THE FUTURE OF INFORMATION SCIENCES

FF press

INFuture 2019

7th International Conference
The Future of Information Sciences
INFuture2019: Knowledge in the Digital Age
Zagreb, 21-22 November 2019

Organizer

Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
Croatia

Publisher

Faculty of Humanities and Social Sciences, University of Zagreb
Department of Information and Communication Sciences
FF press

All papers were reviewed by at least two reviewers. INFuture relies on the double-blind peer review process in which the identity of both reviewers and authors as well as their institutions are respectfully concealed from both parties.

ISSN 2706-3518

DOI: https://doi.org/10.17234/INFUTURE.2019

THE FUTURE OF INFORMATION SCIENCES

INFUTURE2019

KNOWLEDGE IN THE DIGITAL AGE

Edited by

Petra Bago, Ivana Hebrang Grgić, Tomislav Ivanjko,
Vedran Juričić, Željka Miklošević and Helena Stublić

Zagreb, November 2019

CONTENTS

https://doi.org/10.17234/INFUTURE.2019.1

Preface

This is the seventh publication in the series of biennial international conferences, *The Future of Information Sciences (INFuture)* organised by the Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb.

Since its beginnings twelve years ago, the *INFuture* conference has been providing a platform for discussing both theoretical and practical issues in information organization and information integration through the explorations of how developments in information and communication technology influence the future of the field of information sciences. Education and research in information sciences and its interdisciplinary scope and application is of particular interest to this conference which is aimed at researchers and professionals from the broad field of information and communication sciences and related professions.

The title of this year's conference is *INFuture2019: Knowledge in the Digital Age*. The conference explores the influence the information and communication sciences have on the society as a whole.

The *INFuture2019* conference consists of 26 papers from 58 authors from nine countries - Austria, Croatia, Germany, Netherlands, Norway, Slovenia, South Korea, Sweden and United States. This publication follows the five topics of the conference: Digital Heritage (DH), Information Systems and Management (IS&M), Language and Data Technologies (L&DT), Integration of ICT in Education (ICTiE) and Media and Communication (M&C). All papers published in this publication were submitted to a rigorous, double-blind, peer review process by leading experts.

Keynote paper presented an important topic of the impact of digitisation on language and changes it brings. Current trends in the language industry were analysed and the author makes a case for digital linguistics as an interdisciplinary field of study that can tackle many issues in the future of digitised society.

The topic on *Digital Heritage (DH)* elaborates the issues of digitisation and records management, digital repositories and digital libraries, interactive multimedia content, crowdsourcing initiatives and platforms, 3D and GIS platforms and systems, virtual curating, digital storytelling and digital humanities.

The topic on *Information Systems and Management (IS&M)* analyses issues of system design, security and interface, information and communications technologies, open source and freeware, e-government and e-services and blockchain based technologies.

The following topic *Language and Data Technologies (L&DT)* is focused on natural language processing, machine translation, data science and big data analytics, machine learning and data mining, language resources and e-lexicography and e-encyclopaedic sciences.

Integration of ICT in Education (ICTiE) is the conference topic covering e-learning, distant learning and MOOC, virtual and collaborative environments, learning in digital environment, plagiarism detection, service-learning and ICT and simulation and gamification in education.

The last topic, *Media and Communication (M&C)*, explores the social media and Internet technologies, new media and new communications channels, digital democracy, digital literac, media literacy, information literacy and information ethics.

We hope that this year’s issue of conference proceedings will bring new insight and findings regarding the broad and ever-changing information and communication field and broaden our knowledge about the topics covered in the papers. Also, it should shed new light and motivate and inspire new ideas and researches.

See you INFuture!
Editors

https://doi.org/10.17234/INFUTURE.2019.2

Language in the Age of Dataism

Špela Vintar
Faculty of Arts, University of Ljubljana, Slovenia
spela.vintar@ff.uni-lj.si

Summary
The digital age brings dramatic changes to language and communication; its effects can be seen in the ways we use language, the channels we use to communicate and the manners in which ideas are spread. From the other end of the spectrum, our linguistic behaviour, communications and knowledge are transformed into data which can be used or bought to feed intelligent technologies. The article presents a bird's eye view of this dynamics of change, first by focusing on the impact of digitisation on language itself, further by analysing current trends in the language industry where traditional services are being replaced by technology- and data-driven solutions, and finally by exploring the impact of these technologies on man and society at large. We make a case for digital linguistics as an interdisciplinary field of study which adopts a human-centred approach to the sociolinguistic, technological, economic, infrastructural and ethical issues emerging with regard to language in the digital age.

Key words: digitisation, language change, language industry, digital linguistics

Introduction

For some time now, the effects of digitisation on humanity no longer inspire just awe in the face of technological advances but increasingly raise concerns. In less than thirty years of its existence the internet has evolved from a medium charged with tremendous potential for freedom of communication, thought and global cooperation into its shadowy reverse – an environment which has become indispensable but obscured by infringements of privacy, security, dignity, intellectual property rights and competition laws. As Frank Pasquale observes in *The Black Box Society*, the "democratization" promised by Web 2.0 had "a different – even an opposite effect. The very power that brought clarity and cooperation to the chaotic online world also spawns marketing, unfair competition, and kaleidoscopic distortions of reality" (Pasquale, 2015: 98).

With rapid advances of Artificial Intelligence, similar concerns are arising in view of the many scenarios where machine learning algorithms are already replacing human decision-making. The fundamental questions are not whether certain jobs will disappear, which work environments will replace humans by robots and when this is likely to happen for most fields of human endeavour. A more complex set of questions refers to issues like AI bias ("Is the machine fair?") and the moral status of AI ("Is it good or evil?"). In the *Cambridge Handbook of Artificial Intelligence* Bostrom and Yudkowsky lay the foundations for an ethics of AI, acknowledging that "[t]he term 'Artificial Intelligence' refers to a vast design space, presumably much larger than the space of human minds (since all humans share a common brain architecture)" and that certain criteria which apply to humans performing social functions must also be considered in an algorithm intended to replace human judgement: responsibility, transparency, auditability, incorruptibility, predictability (Bostrom and Yudkowsky, 2014).

The beginnings of the age of Big Data celebrated a technological milestone: a point in time when the computational and storage capacity on the one hand and the availability of digital data on the other would no longer present a bottleneck for development. But consider the difference between data collected as sample of human activity in order to build better models, and *collective* data gathered through recording *all* human activity in order to be used, sold and resold by techno-giants and governments alike – this transition marks the beginnings of dataism, which, by Harari's definition, declares that "the universe consists of data flows, and the value of any phenomenon or entity is determined by its contribution to data processing" (Harari, 2016: 351).

It is against this background that we reflect on language in the digital age, whereby our focus shifts from language as a communicative device, language as an economic or business activity to language used as data. It will become clear that from all of these three aspects language has undergone profound changes under the influence of technology, and some of these changes may clearly be regarded as positive. In fact, while popular media will have us believe that the future of *everything* is rather bleak, language in the digital age is, in many respects, thriving.

Digitisation and language change

Languages change over time, and the factors involved in this process range from social, political, technological and economic influences to interventions by normative bodies. It is therefore only to be expected that digitisation and the appearance of numerous new channels of communication would have an impact on language use, and this is often reflected in news articles with titles such as "Is the Innanet RUINING teh English Language??? ¯\(°_o)/¯"[1] or "L3t's t@lk internet"[2]. Linguists have been alert to this topic since the early days of texting (Crystal, 2008), and the expansion and diversification of digital media gave rise to numerous studies exploring their effects on language as a whole or on the use of written language by youth (Baron, 2008; Lenhart, 2008; Thurlow, 2007; Crystal, 2011). In his comprehensive and detailed review of the field of computer-mediated communication (CMC), Androutsopoulos (2011) provides a sociolinguistic set of conditions which shape 'digital networked writing', defining it as "vernacular", "interpersonal and relationship-focused", "unplanned and spontaneous" and "dialogical and interaction-oriented". In a critical synthesis of research studies spanning over three decades, Androutsopoulos demonstrates that much of the language change ascribed to digital media is restricted to lexis, with notorious lists of CMC-typical acronyms and other lexical innovations from the field of technology. The effects of the internet on spoken language seem to be negligible, but the productivity of neologisms derived from social media seems boundless across (written) genres and in languages other than English.

As for netspeak ruining school writing and negatively influencing literacy, evidence is less conclusive, and it is clear that such studies are methodologically difficult to conduct. Lenhart (2008) reports on a large scale study of the attitudes and habits of US teens comparing their out-of-school written communication and school writing, and the prevailing opinion of teens was that texting and communicating via digital media was *not* writing, and that electronic communication had little or no impact on their written production at school. Similarly, Androutsopoulos (2011) mentions an empirical study by Dürscheid and Wagner (2010) carried out in German-speaking Swiss schools, where results suggest that out-of-school digital writing does not visibly influence institutional language production.

This is not to say that the entire landscape of language use has not dramatically changed, mainly through the emergence of new digital genres, and an "unprecedented scale of publicness" that tweets, blogs, posts, news comments and user reviews can achieve. The internet is a mixture of editorial, professionally-crafted content intertwined with vernacular, spontaneous, informal texts; a "manifestation of the intermingling of the private and the public that characterises late modernity" (Androutsopoulos, 2011). We might add that the private/public is only one of the dimensions along which internet discourse is intermingled, other candidate variables being standard/non-standard, true/fake, predominantly textual/predominantly visual, monolingual/multilingual, human-written/machine-written, and many more.

In recent years, a number of language resources, tools and methods have been developed which allow researchers to ask not just whether internet language is different, but *how* different it is. Such studies attempt to quantify the degree to which a certain language variety deviates from standard language, whereby basic corpus pre-processing steps such as lemmatization and PoS-tagging need to be fundamentally adapted or even developed anew to accommodate the transformations and innovations found across genres of the web. In an interesting study of tweets in three closely related languages of former Yugoslavia, Serbian, Croatian and Slovene, Miličević et al. (2017) perform a thorough investigation of spelling transformations and report on a number of similarities and differences. In all three languages, frequent transformations include the omission of diacritics, repetition of certain

[1] https://gizmodo.com/is-the-innanet-ruining-teh-english-language-_o-1680686542 (22. 1. 2015)

[2] https://www.deccanherald.com/sunday-herald/sh-top-stories/l3ts-tlk-internet-668377.html (6. 5. 2018)

vowels for emphasis and omission or transformation of word-final vowels or suffixes. In general, the transformation frequency is highest in Slovene (17%) and lowest in Serbian (10%), with Croatian in the middle (13%), and if the omissions of diacritics are not counted Slovene drops to 15% and Serbian to just over 3% of transformed tokens. This difference is significant – it means that in an average Slovene tweet between 4 and 5 words will be spelled in a non-standard way, while in Serbian only one or none. It would appear, at least for these three languages, that the tendency of a language towards the use of non-standard forms correlates with the level of digital maturity of its country,[3] which is an unexpected finding.

On the other hand, the authors of the study observe that transformations in Serbian, while lower in frequency, occur at more varied positions and indicate a more playful and creative use of language than Slovene or Croatian. On the whole, twitterese and other types of internet discourse mirror layers and layers of social, cultural, political, economic and historical circumstances, and therefore any study of computer-mediated communication limiting itself to just linguistic features necessarily remains incomplete. More importantly, in the same way that virtual communities are communities with their own sociological features, cyber language is a language form in its own right whose properties cannot be described in terms of deviation or transformation from its standard or spoken relatives.

Digitisation affects language beyond the scope of netspeak and genres predominantly residing on the internet. Today, texts are created with the aid of AI technologies and although these are trained on large samples of human language, neural networks may have given rise to a new set of dialects. We are referring mostly to machine translation and the various levels of post-editing applied before such texts are made public. As shown by recent surveys of the language industry which we present in more detail in the next section, the use of MT is growing in all strands of professional translation, but few studies have systematically analysed the properties of post-edited texts. A recent paper by Antonio Toral (2019) fills this gap by addressing the question whether human translation and post-edited machine translation differ significantly in terms of several quantifiable features: lexical variety, lexical density, length ratio and part-of-speech sequences.

The underlying intuition is that translations produced by humans from machine-translated drafts must be somehow different from translations produced by humans from scratch, and Toral performs a number of experiments across six language pairs verifying the existence of *post-editese*. As his results show, post-edited texts have lower lexical variety and density than human translations, and their sentence length and PoS sequences are closer to the source than the target language. This is in line with the so-called "translation universals", the properties of translations which appear across language pairs and include phenomena such as normalization, shining-through and source language interference (Baker, 1993; Maurane, Kujamäki, 2004). Toral's experiment thus proves two important things: firstly, that MT has a lower percentage of content words than HT and is therefore lexically simpler, and secondly that humans striving to improve on MT and create a human-like translation fail to do so, at least as far as lexical variety and PoS sequences are concerned. The author concludes with a cautionary note that "the extensive use of PE rather than HT may have serious implications for the target language in the long term, for example that it becomes impoverished". It remains interesting though that – as Toral himself and several other authors point out (Green, 2013, Bowker and Buitrago Ciro, 2015) – humans do not necessarily perceive HT as better or more acceptable than PE. A recent study by Screen (2019) compares the quality of human and post-edited translations from the end-user perspective. The experiment uses both eye-tracking and end-user assessments of readability and comprehensibility, and the results show no statistically significant difference or inferiority of post-edited texts.

AI and the language industry

In the previous section we briefly discussed some instances where digitisation has impact on language itself, both within and beyond the scope of internet communication. We now turn our attention to the economic sector of language-related services generally referred to as the language industry, which traditionally revolved around translation and interpreting but is increasingly diversified and, as we shall see, datafied. The importance of aggregating translation data became apparent with the emergence of Translation Memory tools, commonly known as Computer-Aided Translation or simply

[3] For Slovenia and Croatia, see the Digital Economy and Society Index 2019, for Serbia the I-DESI 2018.

CAT tools from the early 1990s. With growing needs for fast translation and localization in the globalizing world the idea that past translation projects should be stored in bilingual segments and recycled in order to boost productivity seemed perfectly logical. However, the reactions of translators to CAT tools were reserved at best, with much opposition to the notion that translation work could be conceived as being repetitive and recyclable.

As with most novelties, the technology gradually became mainstream and is considered indispensable today – according to the latest 2018 Language Industry Survey (LIS, 2019: 17) less than 1% of language services companies report that they are not using CAT tools. An interesting historical trivium is that as early as 1997 Trados Translator’s Workbench, the predecessor of today’s market-leading SDL Trados Studio product suite, boasted the use of neural networks for their fuzzy matching algorithms, thus anticipating the AI era in translation technologies.

The development of statistical MT engines and their growing accessibility brought about another shift, namely that of MT becoming a pre-processing step in professional translation, thereby generating the demand for post-editing. Despite the fact that numerous studies have demonstrated significant productivity gains even with early SMTs (O’Brian, 2007; Guerberof, 2008) the sentiments of practicing translators towards PE remain mixed to this day, as a recent survey by the American Translators Association shows (Zetzsche, 2019). The sentiment however is not shared by language service providers. According to the results of the Language Industry Survey for 2016, 2017 and 2018, the use of MT is growing steadily both by companies and individuals. The latest survey, which is considered representative for Europe but not the rest of the world, states that the number of companies and individuals who are not using MT at all has dropped to 31% and 38%, respectively (LIS, 2016, 2017, 2018).

With the arrival of neural Machine Translation (NMT), the language industry was transported into the age of AI. Even if several respondents of the aforementioned ATA survey on “(Why) Do you use MT?” answered “To get a good laugh”, numerous studies have been performed to prove that NMT systems generally outperform SMT models by two or more BLEU points (Bentivogli et al., 2016; Way, 2018), whereby several authors warn that BLEU may be under-reporting the difference in quality. According to error analyses, NMT produces fewer morphological errors (-19%), lexical errors (-17%), and substantially fewer word order errors (-50%) than its closest statistical competitor, and on average requires about a quarter fewer edits compared to the best phrase-based SMT (Way, 2018).

It is thus not surprising that the report issued after the annual TAUS Global Content Conference (TAUS, 2019), an event which attracted 130 world’s largest players in translation and localization, begins with a chapter titled The Quantum Leap and proclaims that “the NMT revolution of the last few years has pretty much wiped out all previous technologies. In addition to this, MT post-editing (PE) has become mainstream, currently the most widely used set-up is MT in conjunction with some degree of human PE.” The size of the Machine Translation market was estimated at 433 million USD in 2016[4] and was expected to grow at an annual rate of 19%. Google Translate’s *daily* throughput exceeds the volume that all translators in the world translate in a year.

According to some estimates, MT is expected to reach the point of *human parity* by 2029, but on the other hand the language industry voices several concerns regarding the use of NMT in business solutions. The first has to do with the robustness of NMT when dealing with different types of content and different domains. This clearly presents a challenge for language service providers, as varying levels of MT quality may have an impact on productivity, return-on-investment and the payment schemes used for PE. A second challenge is the sentence-based mode of processing for most NMT systems which may result in incoherent and inconsistent translations. Research is being put into paragraph- or document-level NMT which would allow systems to translate content, not isolated sentences.

Comparing reports about the language industry from Europe, such as the LIS (2016, 2017 and 2018), and those from more globally oriented organisations such as TAUS (Massardo et al., 2016, Keynotes Summer, 2019) or GALA[5], it appears that the global or US-based view of the language industry

[4] Global Market Insights: Machine Translation Market Size, https://www.gminsights.com/industry-analysis/machine-translation-market-size

[5] http://www.gala-global.org

anticipates more dramatic changes driven by technology and envisages translation as a utility available to everyone, everywhere and on every device. All reviewed studies however agree in forecasting a rapidly growing demand for translations and other language services, in fact these demands even today quite significantly surpass the capacities of human language service providers.

One obvious consequence of this fact is that the majority of translations reach their audiences as raw MT, and that even in professional translation varying levels of quality are required. Both of these facts are hard to digest for a typical professional translator who was trained to strive towards a single and universal highest quality standard, and the position of most translator training institutes regarding quality remains unchanged.

The datafication of translation

There is another important trend we can discern from the reports, and it concerns data. Translation memories and bilingual corpora have been considered important assets for some time now, and issues of ownership, data protection and intellectual property rights have been a hot topic of debate for over a decade (Smith, 2009). The Language Industry Survey for 2017 (LIS, 2018) introduced for the first time a question about the transfer of user rights or ownership to the client, and responses indicated that approximately half of the respondents would never transfer those rights, while the other half would do so sometimes. The results for 2018 reveal a strong trend towards this transfer, and a breakdown of responses by company size shows that for larger companies the transfer of user rights or ownership is now almost mandatory. Large companies work for large clients, and these adhere to the dataism motto that data is the new fuel.

Another TAUS publication titled The Translation Industry in 2022 (Massardo, van der Meer, 2017) identifies Data as one of the six drivers of change and contains a valuable explanation of the difference between language data and translation data. While the former consists of translation memories, corpora, lexicographical and terminological collections, the latter is essentially metadata (Massardo, van der Meer, 2017: 18):

> *Translation data is typically metadata: data about translation that can be harvested downstream the closure of a translation project/job/task, such as content type, language pair(s), domain, subject, number of characters/words/lines, quote/price, scheduled time, time spent, technologies used, translation stats (e.g. source - translation memory match, automatically propagated, machine translated - edited, approved) date and time of last saving, etc.*
>
> *The analysis of translation data can provide a very valuable insight into the translation processes to find the best resource for a job, to decide what to translate and which technology to use for which content.*

Eavesdropping on the debates amongst the tech giants such as Amazon, Apple, Google, Microsoft, Adobe, and the largest LSPs such as Lionbridge, SDL and TransPerfect, the power of data and the central role of AI remain recurring topics. Language data markets have been established, but a lot of data collection goes on backstage using home-grown solutions. Machine translation is but the most obvious application fuelled by data; there is much demand for other intelligent services such as speech processing, user profiling, sentiment analysis, question answering, social network analytics, and there is a pronounced trend towards machine learning for a better management of multilingual workflows.

In view of these developments it would appear that language as a business, not unlike other technology-driven businesses, is under threat of monopolisation by the big players who simultaneously own the bulk of the data, develop the smartest technologies and increasingly own research infrastructures way more powerful than those provided by the academia or public research funding.

A case for digital linguistics

We have examined some of the challenges that language is facing in the digital age; it is now time to reflect on the possible measures to be taken by researchers, academia, practitioners and policy makers in order not to be reduced to mere instruments of change but assume an active role, and possibly direct the course of development into one which is fairer and more inclusive for all members of society.

The advances that Artificial Intelligence is enabling in natural language processing are truly impressive, and scientific progress is accelerated by the enormous amount of private funding flowing into research and by e.g. Google's policy[6] to openly share some of its AI tools with the community, thus enhancing competition. Clearly though, it will be increasingly hard for researchers to keep up with the speed of discoveries produced by the techno-giants.

It is important to remember that the role of science in these – or any other – times is not to blindly compete in the race towards singularity, but to provide critical insights, analyse impact, advocate responsibility, and safeguard the ethical principles fundamental to our society. With regard to the ethics of AI, strong initiatives are underway within leading research institutes, such as the Future of Humanity Institute[7], the IEEE[8] or the Foundation for Responsible Robotics[9], and the European Commission has recently passed a communication titled Building Trust in Human Centric Artificial Intelligence, which defines AI "not as an end in itself, but a tool that has to serve people with the ultimate aim of increasing human well-being" (EC COM 2019: 168).

Returning to language and AI, ethical concerns regarding the use of human translations to train MT systems have been raised by Kenny (2011), especially because "the role of translators in creating vital data has been mostly downplayed or ignored" by MT developers. She also touches upon another interesting ethical question, namely the (im)possibility of computers communicating like humans. According to Melby and Warner (1995) and Kenny (ibid.), "in order to communicate with others, we must have agency, which involves the capacity to make real choices for which we take responsibility, and we must also regard our interlocutors as having agency. […] Without agency, we are reduced to the status of machines and there is no dynamic general language." It is needless to point out that from today's perspective, with chatbots and automated dialogue systems lurking around every corner of the internet, the ethics of communication seems a considerably more complex issue.

A more recent contribution to the discussion about language resources and the ethics of their reuse was made by Moorkens et al. (2016) who systematically describe the practices prevalent in the language industry regarding data ownership, the "disempowered" translator in precarious working positions and the legal situation "in which laws of copyright are effectively bypassed in content collection, curation, and exploitation, [and which] permits resource holders to retain data at a cost to disempowered human writers and translators". The authors' recommendations for translators include collective bargaining, informing themselves about their legal rights and using TM metadata more effectively in order to explicitly assign usage rights to their assets.

Establishing fair practices for data sharing and a transparent regulative system for its collection and processing is just one of the challenges we need to face up to, and the present situation gives little grounds for optimism. As Pasquale writes, "top legal scholars have already analogized the power relationships in virtual worlds and cloud computing to medieval feudalism" (Pasquale 2015: 218).

Considering all the other profound changes that language and communication are undergoing in the digital society, some of which we have discussed above, it becomes clear that to understand and adequately describe these phenomena an interdisciplinary approach is required, and that linguistics alone, even with all its applied subfields, lacks the methodological inventory to approach this task. Analysing large communication networks, proposing new workflows of content creation, developing intelligent knowledge solutions or modelling emotions, to name but a few non-futuristic scenarios, all require a combination and integration of knowledge from different domains.

If solutions for the processing of natural language were traditionally developed by computational linguists, we are now entering an era in which AI technologies are becoming mainstream in many areas of everyday life, and we may well imagine the not-so-distant future when these now separate intelligences begin interacting to solve complex problems, much like intelligent humans do. As we have demonstrated before, any intelligent technology imposed on the human society has a social

[6] https://ai.google/tools/

[7] https://www.fhi.ox.ac.uk/

[8] https://ethicsinaction.ieee.org/

[9] https://responsiblerobotics.org/

dimension in that it modifies the social practices that were in place before, and it may also have ethical, legal, psychological and other dimensions.

We thus propose the term digital linguistics to designate a human-centred approach to digitally-driven language and communication as well as the study thereof, utilizing methodologies and theoretical backgrounds from a range of "feeder" disciplines: linguistics, including computational, corpus, cognitive, socio- and psycholinguistics; computer and information science, including machine learning, data mining, knowledge modelling and AI; social sciences, including law, journalism, communication and media studies; and the relevant humanities, in particular ethics, psychology and philosophy. The list is not exhaustive and serves primarily to emphasize the interdisciplinary nature of digital linguistics.

We further believe it is paramount that universities and other higher education institutions respond not only to the skills gap reported by employers, but more importantly to the expectations and concerns of the civil society which already feels insecure in the "feudalism" of digital communication channels. One attempt to bridge this education gap is the joint master degree in Digital Linguistics in preparation by a consortium of three universities, Ljubljana, Zagreb and Brno, expected to launch in 2021/2022. The model curriculum was developed within the recently concluded DigiLing[10] project and is based on the findings of a trans-European survey of language-related needs amongst employers.[11]

Conclusion

It seems that digitisation affects language in ways different from what the average person or even linguist might expect. The examples selected for discussion above show that the language of internet communications develops under its own rules, not dissimilar to other language varieties known from pre-internet times. Contrary to urban myth, teenagers do know how to draw the boundary between formal and informal writing, while adults or even language professionals have a hard time distinguishing between human and post-edited translations and do not have a clear preference for either. Machine-translated and post-edited texts are found increasingly acceptable by end-users despite the fact that they exhibit pronounced features of the source language.

Word embeddings and neural networks allow us to discern semantic change (Hamilton et al., 2016) or translate between languages for which no parallel data exist (Johnson et al., 2017), but at the same time language professionals feel disempowered as their intellectual property rights are being ignored in the global data collection frenzy. In this article we attempted to present a selection of recent trends involving language and communication in the digital age, and their implications may range from fantastic to catastrophic, depending on one's point of view. A concluding thought might be that as academics and researchers we should strive towards objectivity and realism in the face of the complex challenges, but also towards a responsible stance and a keen interest in the dynamics of change, the only constant of our times.

References

Androutsopoulos, J. (2011). Language change and digital media: a review of conceptions and evidence. // Standard languages and language standards in a changing Europe / Coupland, N.; Kristiansen, T. (eds.), 145-161

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. // Text and technology: In honour of John Sinclair 233, 250

Baron, N. S. (2008). Always on: Language in an Online and Mobile World. Oxford: Oxford University Press

Bentivogli, L., Bisazza, A., Cettolo, M., Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. // Proceedings of EMNLP 2016

Bostrom, N., Yudkowsky, E. (2014). The ethics of artificial intelligence. // The Cambridge handbook of artificial intelligence, 316-334

Bowker, L., Buitrago Ciro, J. (2015). Investigating the usefulness of machine translation for newcomers at the public library. // Translation and Interpreting Studies 10, 2, 165-186

Hamilton, W. L., Leskovec, J., Jurafsky, D. (2015). Diachronic word embeddings reveal statistical laws of semantic change. // Proceedings of the Association for Computational Linguistics (ACL), Berlin

Crystal, D. (2008). Txtng: The gr8 db8. Oxford: Oxford University Press

Crystal, D. (2011). Internet Linguistics. London: Routledge

[10] https://www.digiling.eu

[11] https://www.digiling.eu/deliverables

Dürscheid, C., Wagner, F., Brommer, S. (2010). Wie Jugendliche schreiben: Schreibkompetenz und neue Medien. Berlin/New York: de Gruyter

EC COM. (2019). Building Trust in Human Centric Artificial Intelligence 168. European Commission, 8. 4. 2019

Green, S., J. Heer, C. D. Manning. (2013). The efficacy of human post-editing for language translation. Chi, 439-448

Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., Eisenstein, J. (2016).The social dynamics of language change in online networks. // International Conference on Social Informatics. Springer, Cham, 41-57

Guerberof, A. (2009). Productivity and quality in MT post-editing. // MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT. August 29, Ottawa

Harari, Y. N. (2015). Homo Deus: a Brief History of Tomorrow. London: Harvill Secker

Johnson, M. Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J. (2017). Google’s multilingual neural machine translation system: Enabling zero-shot translation. // Transactions of the Association for Computational Linguistics 5, 339-351

Kenny, D. (2011). The ethics of machine translation. // Proceedings of the New Zealand Society of Translators and Interpreters Annual Conference. Auckland, New Zealand

Lenhart, A. (2008). Writing, technology, and teens. Washington, DC: Pew Internet and American Life Project. http://pewresearch.org/pubs/808/writing-technology-and-teens

LIS (2016). Language Industry Survey – Expectations and Concerns of the European Language Industry. https://www.euatc.org/industry-surveys/item/download/5_57a02b9c45602ea9f7daf4440a7b2979

LIS (2017). Language Industry Survey – Expectations and Concerns of the European Language Industry. https://ec.europa.eu/info/sites/info/files/2017_language_industry_survey_report_en.pdf

LIS (2018). Language Industry Survey – Expectations and Concerns of the European Language Industry. https://ec.europa.eu/info/sites/info/files/ 2018_language_industry_survey_report_en.pdf

Massardo, I., van der Meer, J. (2017). The translation industry in 2022. TAUS BV, De Rijp, The Netherlands

Massardo, I., van der Meer, J., Khalilov, M. (2016). TAUS Translation Technology Landscape Report. September 2016, TAUS BV, De Rijp, The Netherlands

Mauranen, A., Kujamäki, P.,eds. (2004). Translation universals: do they exist?. Vol. 48. Amsterdam: John Benjamins

Melby, A. K., Warner, C. T. (1995).The possibility of language: a discussion of the nature of language, with implications for human and machine translation. Amsterdam/Philadelphia: John Benjamins Publishing

Miličević, M., Ljubešić, N., Fišer, D. (2017). Birds of a feather don’t quite tweet together. // Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World / Fišer, D., Beißwenger, M. (eds.). Ljubljana: Faculty of Arts, 14-43

Moorkens, J., Lewis, D., Reijers, W., Vanmassenhove, E., Way, A. (2019). Translation resources and translator disempowerment. // Tenth International Conference on Language Resources and Evaluation (LREC 2016), 24 - 28 May. Portorož, Slovenia

O’Brien, S. (2007). An Empirical Investigation of Temporal and Technical Post-Editing Effort. // Translation and Interpreting Studies 2, 1

Pasquale, F. (2015). The black box society. Cambridge, Massachussetts: Harvard University Press

Screen, B. (2019). What effect does post-editing have on the translation product from an end-user’s perspective? // Journal of specialised translation 31, 133-157

Smith, R. (2009). Copyright issues in translation memory ownership. // ASLIB Translating and the Computer 31

TAUS (2019). A Review of the TAUS Global Content Conference in Salt Lake City, UT (USA). TAUS Signature Editions, Amsterdam, www.taus.net

Toral, A. (2019). Post-editese: an exacerbated translationese. // Proceedings of the Machine Translation Summit XVII, 1: Research track. Dublin, Ireland, EAMT, 273-281

Thurlow, C. (2007). Fabricating youth: new-media discourse and the technologization of young people. Johnson, S., Ensslin, A. (eds.) Language in the Media. London: Continuum, 213-233

Way, A. (2018). Quality expectations of machine translation. Translation Quality Assessment. Springer, Cham, 159-178

Zetzschke, J. (2019). (How) Do You Use MT? // The Tool Box Journal 19, 11, 306

https://doi.org/10.17234/INFUTURE.2019.3 Preliminary communication

Quantitative Analysis of Adjectives in the Russian Literary Corpus of Realism and Romanticism

Lorena Kasunić
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
lkasunic@ffzg.hr

Petra Bago
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
pbago@ffzg.hr

Summary
Computational analysis of text is an increasingly important approach used by researchers in the field of digital humanities. A much-debated question is whether computational techniques such as text analysis, which is in fact a quantitative approach, is adequate for analysing literary texts, since literature is considered as a type of artistic expression. In the paper we highlight the importance of the application of computational analysis with a study conducted on a corpus of selected Russian literary texts from the periods of Realism and Romanticism. Texts included in the romantic subcorpus are "Eugene Onegin" by Alexander Pushkin and "A Hero of Our Time" by Mikhail Lermontov. Texts that constitute the realist subcorpus are "Anna Karenina" by Leo Tolstoy and "Crime and Punishment" by Fyodor Dostoevsky. The analyzed texts are translations into the Croatian language. The paper presents current methods and approaches used in computational literature analysis. The focus of this research is the analysis of adjective usage in romantic and realist texts, due to the fact that these two literary periods are based on distinctive poetic principles. The texts were analyzed using the programming language "Python". Part-of-speech tagging was accomplished with an online tagger for Croatian language. Considering that all texts are historical (because they originate in the 19th or early 20th century) difficulties with POS tagging are expected. Results of the research show more similarites in the usage of adjectives between the subcorpora then expected. The paper points out how quantitative methods "borrowed" from the field of natural language processing and statistics can be significant in drawing conclusions about literature and that numbers can be meaningful if interpreted competently.

Key words: digital humanities, quantitative methods, stylometry, POS tagging, Croatian, adjective comparison, Russian Romanticism, Russian Realism

Introduction

Computational analysis as one of the methods in digital humanities is applicable to all digital and analog objects that can be studied, meaning that it is not limited to text objects. However, given that the emphasis in this paper is on the analysis of text and language, computational analysis settings will be presented in this context. Computational analysis is a technique that is possible only if there is digitized text. In case the researcher only has analogue text, it must first undergo the digitalization process (Text Analysis Resources, 2016).

What can computational analysis provide in comparison to traditional study of texts without the help of computers? First, it allows us to read a large number of texts in a short amount of time. Here the term "reading" refers to the possibility of passing through the text while orienting to some specific parameters (e.g. the usage of given names as opposed to family names in Jane Austen's novels). Secondly, texts can be automatically classified. Computers are trained to identify whether the analyzed text is a dictionary, a Greek tragedy, a historical epic poem or a letter. It is also possible to determine the authorship of a particular text based on the analysis of the corpus of a presumed author/s. Thirdly, computational analysis makes it easier to see the link between time-distanced texts (Computational Textual Analysis, 2018). Moreover, computer programs can be used to visualize the collected data (using charts, tables, text annotations, etc.). Furthermore, such analysis can empirically and statistically confirm the validity of initial hypotheses, enabling theorists to obtain evidence for their hypotheses (Kerr, 2017).

Related work

Within the field of computational analysis, there are two main approaches: a quantitative and a qualitative approach. However, these approaches are mainly not so distinctive, and they often overlap. Hoover (2008) gives a definition of a quantitative approach claiming it is an approach to literary texts where features or elements of literary texts are numerically represented, applying strong, precise and widely accepted methods of mathematics to measurement, classification and analysis. The increase in the number of available digital texts raised the interest in this approach and stressed the innovation in the ways in which literary texts are treated.

Petrović and Vranešević (2015) defined a quantitative approach as an innovative way of reading literary texts. What is of interest to the quantitative approach to literary texts is the issue of authorship and style, but it is also concerned with some more specific and complex issues such as: genre, theme, tone of the text, periodization (Hammond, 2016).

One of the most popular application of the quantitative approach in literature is stylometry where literary styles are analyzed using the distant reading method (Laramée, 2018). It rests on the premise that authors write in a distinctive, machine-detectable unique way. Problems which stylometry studies are closest to those addressed by the science of literature, with particular interest in patterns and repetitions that are related to issues of interpretation, meaning and aesthetics. The process of stylometric analysis consists of several complex multifactorial stages of preprocessing, feature extraction, statistical analysis and presentation of results, often by visual means (Eder et al., 2016). The linguistic level and grammatical, orthographic, syntactic and morphological research of the text should also not be neglected. The possibilities of applying computational methods in the process of analysis are especially significant when it comes to drawing conclusions from data that even a professional reader (a university professor, literary critic or literary theorist) cannot "detect" by close reading - in this method, the researchers use just distant reading.

The data on the number of transitive verbs, adjectives or the total number of words in Dickens's Great Expectations can be both obtained manually and by quantitative approach. When counting number of parts of speech manually, there is a greater chance of mistakes and it takes a longer time.

In one of her researches on Kafka's literary corpus Berenike Hermann (2017) conducted keyness analysis and compared extracted keywords from the texts of 4 modernistic German authors. She focused on word classes and noticed (based on the given results) the existance of a high frequency of lemmas that may perform "modal" functions in the discourse in the Kafka's corpus. The research shows that quantitative exploration of single words is quite useful (Berenike Hermann, 2017).

Similarly, Algee-Hewitt et al. (2016) analyzed the frequency of combination of any two consecutive words that repeat themselves in observed nineteenth-century novels (canonical and non-canonical texts). This process of lingustic redundancy revealed significat difference between canonical and non-canonical texts: three-fourths of the canonical texts (from the Chadwyck-Healey collection) was less redundant than three-fourths of the non-canonical texts (collected from libraries). This tells us that authors who used language in a redundant way had a bigger chance of being forgotten and remaining unread (Algee-Hewitt et al., 2016).

Qiu and Zhang (2015) in their paper (where they propose new methods of word segmentation for Chinese novels) conclude the following: "For example, based on the analysis of syntax, major events can be extracted from the novel, the relationship between characters can be automatically detected, and sentiment of the author can be analyzed." Just like in our research, Qiu and Zhang use POS tagging as a baseline segmentor.

Kutuzov (2010) on the other hand investigates word types, word tokens and types to token ratio. He compares K. Vonnegut's two novels and their Russian translations, using mostly statistical methods. Many experts who use computer tools in their study of literature are increasingly advocating that computational analysis should join with traditional studies, and that they should complement each other (Hammond, 2016). Quantitative approaches must be aligned with existing ideals and practices in the humanities. The presentation of the mere fact of how many nouns there are in a particular novel does not serve any purpose if it does not consider and clarify the meaning of such data. One appearance of a particular language feature can be much more interesting and important than ten occurrences of another language feature (Petrović, Vranešević, 2015). Quantitative analysis is great for detecting what is rarely or unconventionally used in specific texts. And this can only be detected by counting and comparing, which computing enables (Hoover, 2008).

Dataset

The research about the usage of adjectives in the corpus was conducted on the Croatian translations of works by Russian romanticists and realists. The corpus was divided into two subcorpora – a realist and a romantic corpus. The realist subcorpus consisted of the novels *Anna Karenina* by Leo Tolstoy (translated by Martin Lovrenčević) and *Crime and Punishment* by Fyodor Dostoevsky (translated by Iso Velikanović). The romantic subcorpus consisted of the verse novel, *Eugene Onegin* by Alexander Pushkin (translated by Ivan Trnski), and the novel *A Hero of Our Time* by Mikhail Lermontov (translated by Milan Bogdanović).

Romanticism and Realism as literary periods rest upon contrary principles. These periods do not share the same worldview nor do they perceive the social and cultural reality similarly. Main characteristics of Romanticism are: rejection of the idea of order and racionalism, emphasis on the emotions, irrationality, subjectivity. Authors are occupied with the idea of a genius, a hero, an exceptional individual who is a visionary creator (Croatian Encyclopedia, n.d.). They often connect that individual with the surrounding nature. Lyrical expressions, outburst of emotional states presented in a form of description (when talking about novels) - all these features are considered typical for the literary period of Romanticism.

Realism, on the other hand, tends to be coprehensive as much as possible. Character shaping is of crucial importance for realist authors. Much space is given to showing social, cultural, economical and political circumstances and conditions. Unlike Romanticism, Realism tries to reduce the ramification of the plot and wants to put a light on character's development (Croatian Encyclopedia, n.d.). Prose, especially novel, is the dominant tool of literary expression. Pushkin and Lermontov are canonical names in Russian romantic literature and because of that they can be considered as representatives of a typical Realism poetics. Taking texts of similar artistic and poetic value makes the comparison more accurate and precise. Description as a narrative technique is present both in Romanticism and Realism but in different ways. Romantic authors describe nature, feelings and melancolic atmosphere. Realist authors describe physical appearance, the space of the plot (wretched houses and flats, public houses etc.) on a very "realistic" way, without idealization or emotional enthusiasm. That is the reason why the usage of adjectives shoud differ in the romantic and the realist subcorpora.

These texts were chosen because they represent the culmination of Russian literature and are representatives of the periods from which they originate. All the texts selected, except *Eugene Onegin*, are deliberately prose because belonging to the same literary form requires the application of similar principles in the structure of the text. Poems and dramas have a specific structure and were therefore not considered. It is well-known that the novel as a form prevailed in Realism, while Russian Romanticism remains known for the texts analyzed in our research, even though Pushkin was a great poet. The main hypothesis was that the use of adjectives would differ in the analyzed subcorpora since Romanticism and Realism are based on, we may say, completely opposed poetics and modes of expression, as well as thematic preoccupations. The texts do not have a balanced number of tokens, so below in Table 1 we have provided the data on the size of the corpus, subcopora and the individual literary texts, punctuation included:

Table 1. Number of tokens in each literary text and subcorpus

	Romantic subcorpus		Realist subcorpus	
	Title of text	Number of tokens	Title of text	Number of tokens
	Eugene Onegin	31,516	*Anna Karennina*	370,282
	A Hero of Our Time	55,134	*Crime and Punishment*	229,328
Total	86,650		599,610	

(subcorpora)		
Total (corpora)	686,260	

Methodology

Python (version 3.6.0) and POS tagging (for the Croatian language[1]) were used in the process of computational analysis. All texts were downloaded from the *eLektire* website in the .txt format for further study[2].

The research was based on the application of statistical methods and methods used in natural language processing. The first step was to preprocess the texts - removing data such as footnotes, notes, titles, author's names, dictionaries of lesser-known words. It was necessary to obtain "pure" literary texts without any metadata, as they would affect the results of the research. Subsequently, the texts were processed using the online POS tagger, through which the texts were morphosyntactically tagged and the lemmatization was performed. All subsequent analyses were performed on the processed text.

It was necessary to check the accuracy of the classifier before the data was obtained. This was done by taking one segment of the text and tagging it manually by one annotator. First 500 tokens (including punctuation) from *Eugene Onegin* were chosen for manual tagging. Since the research was focused on adjectives, manual tagging did not go into deep morphosyntactic analysis, but only verified whether the parts of speech were correctly labeled.

Using the *Python* programming language, we obtained data on the frequency distribution of words and punctuation for each text, with reference to the tags used when tagging texts in Croatian. After that, only the adjectives were generated, in order to show which ones appeared most often in a particular text.

The tagger wrongly labeled some words as adjectives, that we excluded from the analysis. Methods used in the analysis of literary texts are mostly "borrowed" from natural language processing and statistics. In this study, the methods used were frequency distribution, tokenization, lemmatization and POS tagging.

Results

The first results that were obtained were related to the accuracy of the classifier itself. The website[3] of the Croatian POS tagger states that the accuracy for the Croatian language is 92.53%. We manually labeled the first 500 tokens (including the punctuation) of *Eugene Onegin* to evaluate the POS tagger and obtained the accuracy of 87.2%. When studying the data, it was apparent that the classifier did not make any mistakes in tagging punctuation, so the accuracy was calculated if the punctuation was excluded. The accuracy then dropped to 84.2%. Out of the 61 errors counted, 10 were related to the wrong tagging of adjectives. In Table 2 we provide the confusion matrix for POS tags.

Table 2. The confusion matrix for POS tags

		Actual class												
		N	V	A	P	R	S	C	M	Q	I	Y	X	Z
Predicted class	N	93	12	3	3	6	0	1	0	0	0	0	0	0
	V	2	64	3	2	4	0	0	0	0	0	0	0	0
	A	2	4	56	3	3	0	0	0	0	0	0	0	0
	P	2	0	2	45	0	0	0	0	0	0	0	0	0
	R	0	0	1	1	18	0	0	0	0	0	0	0	0
	S	0	0	0	0	0	22	0	0	0	0	0	0	0
	C	0	0	0	3	0	0	15	0	1	0	0	0	0
	M	0	0	1	0	2	0	0	6	0	0	0	0	0
	Q	0	0	0	0	0	0	1	0	3	0	0	0	0
	I	0	0	0	0	0	0	0	0	0	0	0	0	0
	Y	0	0	0	0	0	0	0	2	0	0	1	0	0
	X	0	0	0	0	0	0	0	0	0	0	0	0	0

[1] http://www.clarin.si/info/about/

[2] https://lektire.skole.hr

[3] http://www.clarin.si/info/k-centre/web-services-documentation/

	Z	0	0	0	0	0	0	0	0	0	0	0	0	113

If we look at this from another direction (specifically, if we are interested in how many words the classifier tagged as adjectives, when they were in fact some other part of speech), it comes up to 12 mistakes. Of the 61 errors in total, 22 are connected with adjective annotation.

Next, we calculated the frequency distribution of word types. Table 3 shows the frequency distribution of words and punctuation for each text.

Table 3. Frequency distribution of word types and punctuation

	Eugene Onegin	*A Hero of Our Time*	*Crime and Punishment*	*Anna Karenina*	Romantic subcorpus	Realist subcorpus
Verbs	4,422 (14.03%)	**11,838 (21.47%)**	45,541 (19.86%)	**77,475 (20.92%)**	16,260 (18.77%)	**123,016 (20.52%)**
Nouns	7,546 (23.94%)	8,945 (16.22%)	33,123 (14.44%)	62,494 (16.88%)	16,491 (19.03%)	95,617 (15.95%)
Adpositions	1.730 (5.49%)	3,435 (6.23%)	12,690 (5.53%)	23,449 (6.33%)	5,165 (5.96%)	36,139 (6.03%)
Adverbs	1,364 (4.33%)	3,660 (6.64%)	17,096 (7.45%)	24,184 (6.53%)	5,024 (5.80%)	41,280 (6.88%)
Adjectives	3,131 (9.93%)	3,642 (6.61%)	13,898 (6.06%)	25,218 (6.81%)	6,773 (7.82%)	39,116 (6.52%)
Pronouns	3,215 (10.20%)	7,284 (13.21%)	28,471 (12.41%)	49,032 (13.24%)	10,499 (12.12%)	77,503 (12.93%)
Numerals	470 (1.49%)	554 (1.00%)	1,886 (0.82%)	2,635 (0.71%)	1,024 (1.18%)	4,521 (0.75%)
Conjuctions	1,260 (4.00%)	4,617 (8.37%)	21,388 (9.32%)	32,904 (8.89%)	5,877 (6.78%)	54,292 (9.05%)
Interjections	52 (0.16%)	98 (0.18%)	480 (0.21%)	447 (0.12%)	150 (0.17%)	927 (0.15%)
Particles	666 (2.11%)	1,079 (1.96%)	6,217 (2.71%)	7,722 (2.09%)	1,745 (2.01%)	13,939 (2.32%)
Abbrevations	40 (0.13%)	9 (0.02%)	14 (0.01%)	64 (0.02%)	49 (0.06%)	78 (0.01%)
Residuals	25 (0.08%)	27 (0.05%)	28 (0.01%)	187 (0.05%)	52 (0.06%)	215 (0.04%)
Punctuation	**7,595 (24.10%)**	9,946 (18.04%)	**48,496 (21.15%)**	64,471 (17.41%)	**17,541 (20.24%)**	112,967 (18.84%)
Total	31,516	55,134	229,328	370,282	86,650	599,610

The morphosyntactic tagger is trained on the corpus of texts that are part of the Croatian Language Repository, hrWaC (the Croatian Web Corpus), and the Croatian National Corpus. Data on the first two corpora are freely available on the Internet and they were used to compute the relative frequency distribution of parts of speech in order to compare the results with the corpus of texts used in this research. In *Eugene Onegin*, the biggest difference in percentages is visible in the distribution of punctuation marks: in this text their relative frequency distribution is 24.1%, in hrWaC is 11.89%, or 14.29% in the Croatian Language Corpus. In *A Hero of Our Time*, except for punctuation, there are differences in representation of verbs (21.47% in the analyzed text, 16.05% in hrWaC and 14.82% in the Croatian Language Repository), nouns (16.22% in the analyzed text, 26% in hrWaC, 27.91 % in the Croatian Language Repository) and pronouns (13.21% in the analyzed text, 8.2% in hrWaC and 6.88% in the Croatian Language Repository). In the novel *Crime and Punishment* the biggest differences are again found in the punctuation (we may say that this is the leitmotif in all the analyzed texts), then in the nouns (14.44% is the relative frequency in the novel, the percentages for hrWaC and the Croatian Language Repository are identical to the ones previously mentioned) (12.41%). The same goes for *Anna Karenina*: the relative frequency distribution of nouns is 16.88%, punctuation marks 17.41%, and verbs 20.92%. As far as adjectives are concerned, the largest relative frequency distribution is 9.93% and refers to *Eugene Onegin*, while the rest of the texts have a similar relative frequency distribution for adjectives: *A Hero of Our Time* - 6.61%, *Crime and Punishment* - 6.06%, *Anna Karenina* - 6.81%.

Table 4. The number of individual adjectives in each text

	Eugene Onegin	*A Hero of Our Time*	*Crime and Punishment*	*Anna Karenina*
Descriptive adjectives	**2,961 (94.57%)**	**3,389 (93.05%)**	**12,589 (90.58%)**	**23,160 (91.84%)**
Possessive adjectives	53 (1.69%)	32 (0.88%)	379 (2.73%)	587 (2.33%)
Past participle	117 (3.74%)	221 (6.07%)	930 (6.70%)	1,471 (5.83%)
Total number of adjectives	3,131	3,642	13,898	25,218

Within the category of adjectives, an additional frequency distribution of certain adjectives (descriptive, possessive, and past participle) was performed, as can be seen in Table 4. The classifier recognizes these three types of adjectives and this is the reasoning behind this categorization, although in the grammar of the Croatian language there is a basic categorization of adjectives into descriptive, constructive and possessive. Given that the analysis is predominantly oriented on the presence of adjectives in each of these literary texts, we assembled the data on the 20 most common adjectives found in particular texts. From the results, it is evident that adjectives “sam” (Eng.“alone”) and “sav” (Eng. “entire”) appear in all four texts. These are the most commonly used adjectives. In hrWaC, the frequency distribution for “sam” (Eng. “alone”) is 1,261,043 (0.97% of the total number of adjectives in the corpus) and 5,303,295 (4.07%) for “sav” (Eng. “entire”). In the Croatian Language Repository frequency distribution for “sam” (Eng. “alone”) is 66,518 (0.65% of the total number of adjectives in the corpus) and 288,819 (2.84%) for “sav” (Eng. “entire”). All the other 20 most common adjectives for each literary text can be seen in Table 5. As mentioned in the previous chapter, words (and letters) which were excluded from the table of adjectives are: “moj” (Eng. “mine”), “vaš” (Eng. “your”), “njen” (Eng. “hers”), “Svidrigajlov” (Eng. “Svidrigailov”), “Raskoljnikov (Eng. “Raskolnikov”), “njegov” (Eng. “his”), “l” (Eng. “l”).

Table 5. 20 most common adjectives for each literary text

Eugene Onegin	*A Hero of Our Time*	*Crime and Punishment*	*Anna Karenina*	hrWaC	Croatian Language Repository
sav (Eng. entire)	sav (Eng. entire)	sam (Eng. alone)	sav (Eng. entire)	sav (Eng. entire)	sav (Eng. entire)
mlad (Eng. young)	sam (Eng. alone)	sav (Eng. entire)	sam (Eng. alone)	velik (Eng. big)	hrvatski (Eng. Croatian)
sam (Eng. alone)	velik (Eng. big)	isti (Eng. same)	dobar (Eng. good)	nov (Eng. new)	velik (Eng. big)
star (Eng. old)	čitav (Eng. intact)	cijel (Eng. whole)	isti (Eng. same)	dobar (Eng. good)	nov (Eng. new)
lijep (Eng. beautiful)	hladan (Eng. cold)	velik (Eng. big)	nov (Eng. new)	hrvatski (Eng. Croatian)	dobar (Eng. good)
mio (Eng. dear)	dobar (Eng. good)	posljednji (Eng. last)	lijep (Eng. beautiful)	sam (Eng. alone)	europski (Eng. European)
krasan (Eng. splendid)	mlad (Eng. young)	dobar (Eng. good)	velik (Eng. big)	mali (Eng. small)	sam (Eng. alone)
velik (Eng. big)	crn (Eng. black)	nov (Eng. new)	star(Eng. old)	isti (Eng. same)	državni (Eng. national)
nov (Eng. new)	čudan (Eng. strange)	neobičan (Eng. unusual)	mlad (Eng. young)	cijel (Eng. whole)	politički (Eng. political)

ruski (Eng. Russian)	isti (Eng. same)	čudan (Eng. strange)	veseo (Eng. merry)	ostali (Eng. remaining)	mali (Eng. small)
tanak (Eng. thin)	prav (Eng. real)	mali (Eng. small)	moguć (Eng. possible)	star (Eng. old)	isti (Eng. same)
sladak (Eng. sweet)	bijel (Eng. white)	pijan (Eng. drunk)	čitav (Eng. intact)	mlad (Eng. young)	američki (Eng. American)
čudan (Eng. strange)	štaban (Eng. headquartered)	osobit (Eng. special)	potreban (Eng. neccesary)	važan (Eng. important)	posljednji (Eng. last)
živ (Eng. alive)	uvjeren (Eng. convinced)	jasan (Eng. clear)	posljednji (Eng. last)	potreban (Eng. neccessary)	glavni (Eng. main)
drag (Eng. dear)	posljednji (Eng. last)	prav (Eng. real)	sretan (Eng. happy)	prav (Eng. real)	zagrebački (Eng. Zagreb's)
prost (Eng. vulgar)	pun (Eng. full)	mlad (Eng. young)	mali (Eng. small)	poznat (Eng. famous)	svjetski (Eng. worldwide)
bijel (Eng. white)	lijep (Eng. beautiful)	bolestan (Eng. sick)	visok (Eng. tall)	mnogi (Eng. many)	star (Eng. old)
hladan (Eng. cold)	smiješan (Eng. funny)	glup (Eng. stupid)	bijel (Eng. white)	europski (Eng. European)	ostali (Eng. remaining)
tih (Eng. quiet)	blijed (Eng. pale)	star (Eng. old)	miran (Eng. still)	glavni (Eng. main)	međunarodni (Eng. international)
dobar (Eng. good)	star (Eng. old)	strašan (Eng. terrible)	strašan (Eng. terrible)	visok (Eng. tall)	mlad (Eng. young)

Discussion

The application of the classifier on the corpus of literary texts showed that the classifier successfully performed part-of-speech tagging. The accuracy difference is 5.33% (accuracy of the classifier obtained on tested text fragment: 87.2%), compared to the data about the accuracy of the classifier available on its website (92.53%). One ought to keep in mind that here we are dealing with Croatian translations of historical texts created during the 19th and 20th century and therefore their language differs from typical contemporary Croatian language. To test the accuracy of the classifier a text fragment was taken from *Eugene Onegin* because the text (by its structure and language) deviates the most from everyday language, and therefore it is more likely that the classifier will make a mistake whilst tagging. Because of the lack of information on the dates of translations for *Anna Karenina* and *Crime and Punishment*, one cannot decidedly explain the wrong tagging by the fact that the translation of *Eugene Onegin* is the oldest from the analyzed texts, although this possibility should not be dismissed.

From the results of the relative frequency distribution, we can conclude that the literary texts which make up our corpus have a different number of nouns, verbs, pronouns and punctuation marks in relation to hrWaC and the Croatian Language Repository. *Anna Karenina*, *A Hero of Our Time* and *Crime and Punishment* use more verbs and pronouns and fewer nouns in relation to the above-mentioned corpora. With *Eugene Onegin*, it is a different situation. There are less verbs, pronouns and even conjunctions than in other novels. On the other hand, nouns and adjectives are more present than in other texts. This seems to be the case because this a verse novel that inherited a part of the lyrical influence. Although there is a plot, it is not the central aspect of the novel, and therefore there are fewer verbs. The greater quantity of adjectives and nouns can be explained by the presence of the lyrical mode of expression which strives for descriptiveness, expressiveness, and enumeration. A low percentage of conjunctions can be associated with a high percentage of punctuation marks, which are an essential element of writing in verse. Comparing the two subcorpora, the realist and romantic, and then considering the relative frequency distribution, *A Hero of Our Time* could be assigned to the realist rather than a romantic subcorpus. It must be emphasized that this conclusion was reached without considering the content of the text itself, so it is exclusively based on statistical data. Punctuation generally has a much larger share in all literary texts than in the Croatian Web Corpus and the Croatian Language Repository. This phenomenon is understandable as literary texts often use complex sentences, *stringing*, inserts, and these devices increase the use of commas, colons, ellipses, parentheses, etc.

If we observe the frequency distribution of a particular type of adjective (descriptive, possessive, and past participle), we can observe a uniformity of the ratios within all analyzed texts. The highest in frequency are descriptive adjectives, then past participles and finally possessives. From this, one can read the common feature of Realism and Romanticism - the aspiration to describe, whether the fictional world in which the action takes place, the characters that inhabit it, or feelings and emotional states. What is of greatest interest is which specific adjectives appear in a particular text. We produced lists of 20 most frequent adjectives for each text. Through our research, it became apparent that the adjectives on the top of the list are very similar, namely “sav” (Eng. “entire”) and sam (Eng. “alone”). The romantic subcorpus contains a lot of similarities and repetition of adjectives - both in *Eugene Onegin* and *A Hero of Our Time*, the following adjectives are present: “sav” (Eng. “entire”), “sam” (Eng. “alone”), “lijep” (Eng. “beautiful”), “bijel” (Eng. “white”), “velik” (Eng. “big”), “hladan” (Eng. “cold”), “čudan” (Eng. “strange”), “mlad” (Eng. “young”), “star” (Eng. “old”). Both texts have one or two pairs of mutually opposing adjectives: “mlad” (Eng. “young”) and “star” (Eng. “old”) in *Eugene Onegin* and *A Hero of Our Time* and “bijel” (Eng. “white”) and “crn” (Eng. “black”) in *A Hero of Our Time*. There are similarities in the realist subcorpus as well. Adjectives that appear both in *Crime and Punishment* and in *Anna Karenina* are “sam” (Eng. “alone”), “sav” (Eng. “entire”), “isti” (Eng. “same”), “nov” (Eng. “new”), “velik” (Eng. “big”), “posljednji” (Eng. “last”), “dobar” (Eng. “good”), “mali” (Eng. “small”), “mlad” (Eng. “young”), “strašan” (Eng. “terrible”), “star” (Eng. “old”). What is particularly important for these texts is that the names of characters (Raskolnikov, Svidrigailov and Vronski) are referred to as adjectives. The reason for this is in the suffixes -ov and -ski which are typical for adjectives in the Croatian language, and not for proper names. There are also pairs of adjectives: “velik” (Eng. “big”) - “mali” (Eng. “small”) (*Crime and Punishment*) and “star” (Eng. “old”) - “mlad” (Eng. “young”) (*Anna Karenina* and *Crime and Punishment*). It is interesting that the adjective “strange” appears in all the texts, except in *Anna Karenina*. In fact, a large number of adjectives are common to all four texts, although they belong to different literary periods. This could be perceived as one of the theories often emphasized by literary theorists and that is that one cannot draw clear boundaries between literary periods. Influences are always present and for no great literary text can be said to be a typical realistic or romantic text, for example.

What were the problems we encountered while conducting the research? The first problem occurred when using the classifier. Namely, the realist texts are much more extensive than the romantic ones and when they were supposed to be tagged, they were simply too large and the online classifier was blocked. Therefore, these texts had to be divided into several smaller text files that were then tagged and merged into a single text file that was used in further analysis. When launching the *Python* program code for each text, it was apparent that due to the size of *Anna Karenina* file and *Crime and Punishment* file, it took more time for *Python* to produce the data. This was especially noticeable in the part of the research where 20 of the most frequent adjectives were extracted. As for the manual tagging, there were problems with defining parts of speech for some words because they were outdated (e.g. “priljem”, “ponevju”, “stežne”, “zgolje”) and the annotator did not know the meaning of these words.

The results have shown that there are some differences in the use of adjectives between the two subcorpora, but they are not as drastic as expected. A possible answer might lie in the fact that what makes literary texts differ from one another are language phenomena that are not so great in number, i.e., those that are specific to a certain text. In this particular case that would mean that we should study the adjectives that are in the middle of frequency rankings - they are not so rare that they could be considered as exceptions and not so frequent that their occurrence could be attributed to the general method of structuring and using language in literary art. Given adjectives could be further observed in the surrounding context in which they are occur. It would also be interesting to see if are there any patterns in distribution of different types of adjectives throughout the romantic and the realist subcorpora. Perhaps a similarity in the choice of adjectives may confirm a certain intertextuality and literary influences. That could be a possible direction for future work on this or similar corpora of literary texts.

Conclusion

Quantitative approaches have their supporters, as well as their opponents. They contribute to the improvement and further development of practices within digital humanities. Examples of concrete application of methods in computational analysis show that empirical data can be of use in attempts to interpret literary texts.

We attempted to implement computational methods in order to test our hypothesis that the usage of adjectives differs between the two opposite literary periods. The research focused exclusively on the use of quantitative tools, such as frequency distribution. It also used methods from natural language processing (POS tagging, lemmatization, tokenization). The process of conducting the research has confirmed the usefulness of the quantitative approach in the interpretation of literature, but only if there is a human agent who will be able to interpret the obtained data. The paper endeavored to give empirical results and conclusions which can shed a light on the complicated question of boundaries between literary periods.

Although the beginnings of computer usage in studying historical texts (including literary ones) originate in the 1950s, there is still room for progress. More focus needs to be put on the development of new tools and methods, especially for texts that are not written in world languages such as English. There is still a lack of properly digitized machine-readable Croatian texts, especially historical ones from various (literary) periods. Research is mostly carried out on large canonical texts, and the less famous ones are neglected. There is a need for cooperation of experts from different fields - information science, linguistics, literature, computer science etc. However, despite all the obstacles encountered by digital humanists, computer analysis is increasingly being used as a new approach to literary texts, one that can provide a different point of view and encourage us to ask previously untold or overlooked questions.

References

Algee-Hewitt, M., Allison, S., Gemma, M., Heuser, R., Moretti, F., Walser, H. (2016). Canon/Archive. Large-scale Dynamics in the Literary Field. https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf (25.10.2019.)

Berenike Hermann, J. (2017). In a test bed with Kafka. Introducing a mixed-method approach to digital stylistics. // Digital Humanities Quarterly 11, 4. http://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html (14.3.2019.)

Computational Textual Analysis. (2018). https://guides.temple.edu/corpusanalysis (15.3.2019.)

Croatian Encyclopedia. http://www.enciklopedija.hr/ (24.10.2019.)

Eder, M., Rybicki, J., Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. // The R Journal 8, 1. https://journal.r-project.org/archive/2016/RJ-2016-007/RJ-2016-007.pdf (20.7.2019.)

Hammond, A. (2016). Quantitative Approaches to the Literary. // Literature in the Digital Age: An Introduction / Cambridge: Cambridge University Press, 82-130

Hoover, D. L. (2018). Quantitative Analysis and Literary Studies. // A Companion to Digital Literary Studies / Schreibman, S., Siemens, R. (eds.). Oxford: Blackwell. http://digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-9&toc.id=0&brand=9781405148641_brand (26.2.2019.)

Kerr, S. J. (2017). When Computer Science Met Jane Austen and Edgeworth. // NPPSH Reflections 1, 38-41

Kutuzov, A. (2010). Change of word types to word tokens ratio in the course of translation (based on russian translations of k. Vonnegut's novels). https://arxiv.org/ftp/arxiv/papers/1003/1003.0337.pdf (25.10.2019.)

Laramée, F. D. (2018). Introduction to stylometry with Python. https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python (22.7.2019.)

Petrović, B., Vranešević, D. (2015). Kvantitativna raščlamba Čudnovatih zgoda šegrta Hlapića Ivane Brlić-Mažuranić. // Šegrt Hlapić – od čudnovatog do čudesnog / Majhut, B., Narančić Kovač, S.; Lovrić, S. (eds.). Zagreb: Slavonski Brod: Hrvatska udruga istraživača dječje književnosti: Ogranak Maticea hrvatske, 251-267

Qiu, L., Zhang, Y. (2015). Word Segmentation for Chinese Novels. // Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9323/9538 (24.10.2019.).

Technopedia. https://www.techopedia.com/definition/13698/tokenization (19.7.2019.)

Text Analysis Resources. (2016). https://digitalhumanities.berkeley.edu/resources/text-analysis-resources, (15.3.2019.)

The Stanford Natural Language Processing Group. https://nlp.stanford.edu/software/tagger.shtml (20.7.2019.)

https://doi.org/10.17234/INFUTURE.2019.4

Preliminary communication

Croatian Web Dictionary – Mrežnik vs. Croatian Linguistic Terminology – Jena

Lana Hudeček
Institute of Croatian Language and Linguistics, Zagreb, Croatia
lhudecek@ihjj.hr

Milica Mihaljević
Institute of Croatian Language and Linguistics, Zagreb, Croatia
mmihalj@ihjj.hr

Summary

The Croatian Web Dictionary – Mrežnik *is a four-year project which started on the 1st of March 2017 and the duration of the project is four years. The main result of the project will be a free, monolingual, hypertext online dictionary consisting of three modules (the module for adult native speakers – 10,000 entries, the module for children aged 6 to 10 – 3,000 entries, and the module for non-native speakers learning Croatian – 1,000 entries).* Mrežnik *is based on two Croatian web corpora.* Croatian Linguistic Terminology – Jena *is a new terminological project conducted within the* Struna *program. The project started on the 24th of May 2019 and lasts until the 23rd of November 2020. The main result of the project will be a multilingual database consisting of 1,500 entries. As a specialized corpus of Croatian linguistic terminology doesn't exist, it is compiled in parallel with the database. Although* Mrežnik *and* Jena *differ in their basic goals and approach; one is monolingual and general and the other is multilingual and specialized (terminological), one is compiled from the existing corpora and the other is compiled in parallel with the corpus, they have two important meeting points: 1. General linguistic terminology is presented in* Mrežnik *(mostly but not exclusively in the module for adult native speakers) and 2. Within the* Mrežnik *project,* Glossary of E-lexicographic terminology *is compiled. These four parameters will be compared: 1. wordlist/termlist, 2. relation to the corpus, 3. giving normative information, 4. entry structure. The compilation process and the structure of entries for the same headword will be compared and the important similarities, as well as differences, will be shown. The reason for this comparison is that the two projects are conducted at the same time and strongly influence each other in many aspects.*

Key words: Mrežnik, Croatian Linguistic Terminology – Jena, linguisitic terminology, e-lexicography

Introduction

The project *Croatian Web Dictionary – Mrežnik*[1] aims at creating a free, monolingual, easily searchable hypertext online dictionary of standard Croatian. It will be the first web-born dictionary of the Croatian language. Entries, sub-entries, and meanings will be interconnected, as well as linked to entries in databases created within the framework of the project in parallel with the creation of the dictionary (language advice database, conjunction database with description of groups of conjunctions and their modifiers, database of explanations of the origin of idioms, database of ethnics and ktetics), as well as databases being created by project collaborators or other Institute members within the framework of other projects.[2] *Mrežnik* consists of three modules: the module for adult native speakers of Croatian which will have 10,000 entries, the module for school children which will have 3,000 entries, and the module for non-native speakers which will have 1,000 entries). The dictionary is written in the *TLex* program, which has been adapted to the needs of the project. The main goals of the project are: 1. to create the three dictionary modules, 2. to connect the dictionary with the databases created in parallel with the dictionary, 3. to connect the dictionary with other web sources currently being compiled at the Institute of Croatian Language and Linguistics, 4. to compile a reversed dictionary based on the *Mrežnik* wordlist, 5. to write a monograph on *Mrežnik*. The project

[1] More on *Mrežnik* see in Hudeček 2018; Hudeček, Mihaljević 2017a, 2017b; Hudeček, Mihaljević, 2018a, 2018b.

[2] More on this topic see in Hudeček, Mihaljević 2019a.

started on the 1st of March 2017, so at the moment we are in the second half of the project and more than 5,000 entries have been compiled.

The project *Croatian Linguistic Terminology – Jena* is conducted within the *Struna* program. *Struna* is a database of Croatian Special Field Terminology[3] financed by the Croatian Science Foundation. *Jena* is a year-and-a-half project which started on the 24th of May 2019. The main goals of the project are: 1. to compile 1,500 entries with definitions, synonyms, antonyms, hyponyms and equivalents in English, German, French, Russian, and Swedish in the *Struna* database, 2. to collect works on linguistic terminology and present them on the *Jena* website (ihjj.hr/jena/), 3. to write a monograph on Croatian linguistic terminology. At the moment 1035 entries have been entered into the database. However, it is important to note that both *Croatian Web Dictionary – Mrežnik* and *Jena* are conceived as a dynamic dictionary/databases that will be further compiled and edited even after the formal end of the project funding and will not reach their full extent if they do not continue to grow and become an everlasting project of the Institute.

Hypothesis and reason for comparison

The hypothesis of this paper is that although a terminological database obviously differs from a general e-dictionary there are many similarities from which both projects can profit.

The reason for such a comparison is that these two projects are conducted at the same time in the same institution, the head of *Mrežnik* (Lana Hudeček) is the collaborator of *Jena* while the head of *Jena* (Milica Mihaljević) is the collaborator of *Mrežnik*. Thus some results of one project can be applied to the other and vice versa. In the comparison all linguistic terms which appear in *Mrežnik*[4] and the *Glossary of e-lexicographic terminology* compiled within the *Mrežnik* project are taken into account. As both projects are in progress the instructions for the respective team of lexicographers and terminographers can be modified according to new results. The basic points of comparison are: 1. the ways of compiling wordlist/termlist, 2. the approach to the corpus, 3. the approach to normativity, 4. the structure of dictionary entries.

Wordlist vs. termlist

To compile the *Mrežnik* wordlist the frequency lists of *hrWaC* (first 12,000 words) and the *Hrvatska jezična riznica* (first 10,000 words) were overlapped, all words present only in *Hrvatska jezična riznica* and not present in *hrWaC* were extracted, their frequency was multiplied by four, and they were added to the shared list. This wordlist (first 8,000 entries) was juxtaposed with two separate wordlists: the wordlist for the module for children (which was excerpted from textbooks for the first four grades of elementary school with some additions by the collaborators of *Mrežnik*) and the wordlist for the module for non-native speakers which includes 1,000 words taken from a list in textbooks for non-native speakers, to ensure that words found in both these lists (which partially overlap) appear in the list for adult native speakers. This wordlist was supplemented with male/female (in cooperation with the project *Male and Female in the Croatian Language*) and aspectual pairs, possessive and descriptive adjectives, adverbs derived from adjectives from the list, nouns ending in -*ost* derived from adjectives from the list, numerous grammatical and semantic groups, etc. This resulted in a wordlist of 10,000 words with two separate wordlists of 3,000 words (for children) and 1,000 words (for non-native speakers).

The wordlist of the module for children was considered as the basic wordlist and we first began compiling the entries for words from this list in order to make processing for the module for children as compatible as possible with that for adult native speakers (Hudeček, Mihaljević, 2018b).

[3] The Institute of Croatian Language and Linguistics was chosen to serve as the national coordinator. The objective of the program in a broader sense is to lay the foundation for the development of national terminology policy, to establish various forms of more structured education in this field, and to intensify long-term cooperation with national and international academic and other institutions dealing with different aspects of terminology work, with the Croatian Standards Institute and with other interested parties. Within the program, a terminology database has been developed to store and terminographically manage standardized and harmonized Croatian terms from various subject fields and their equivalents in English and other languages. Experts from eighteen domains have so far joined the program with the aim of standardizing the terminology of their respective disciplines. http://struna.ihjj.hr/en/about/.

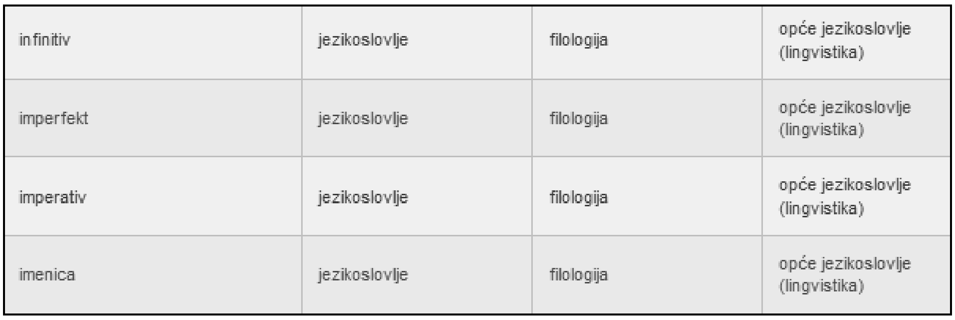
[4] Of course some linguistic terms have a non-linguistic meaning which occurs in *Mrežnik* and doesn't occur in *Jena* but this was not the subject of our analysis.

The *Jena* termlist consisting of 1,500 terms was compiled by project collaborators divided into workgroups by subject fields: basic linguistic terminology, cognitive linguistics, contact linguistics, dialectology, e-lexicography and corpus linguistics, generative linguistics, glottodidactics, language history, lexicography, lexicology, onomastics, orthography, phraseology, pragmatics, sociolinguistics, terminology, translation theory, valency theory. Table 1 shows a small extract from the termlist divided by subject fields.

Table 1. Extraction of *Jena* termlist by specialized subject fields

1.1 Generative linguistics	**1.2 Cognitive linguistics**	**1.3 Phraseology**	**1.4 Translation theory**	**1.5 Language history**
1.6 E-jezik 1.7 generativna gramatika 1.8 I-jezik 1.9 jezična moć 1.10 jezična sposobnost 1.11 jezična uporaba 1.12 logički problem jezičnoga usvajanja 1.13 negativni dokazi 1.14 objasnidbena prikladnost 1.15 opisna prikladnost 1.16 oskudnost poticaja 1.17 pozitivni dokazi	1.18 apsolutni prostorni sustav 1.19 apstrahiranje 1.20 argumentna struktura 1.21 asimetrija izvornoga i ciljnoga okvira 1.22 automatsko prepoznavanje metafora 1.23 autonomistički gramatički pristup 1.24 konceptualne integracije 1.25 ciljna domena 1.26 dinamični razvojni model 1.27 dinamika sile 1.28 diskursna analiza vođena metaforom	1.29 frazeologija u užemu smislu 1.30 frazeologija u širemu smislu 1.31 paremiologija 1.32 krilatologija 1.33 zoonimna frazeologija 1.34 somatska frazeologija 1.35 internacionalna frazeologija 1.36 nacionalna frazeologija 1.37 posuđena frazeologija 1.38 arhaična frazeologija 1.39 dijalektna frazeologija 1.40 regionalna frazeologija 1.41 frazeološki obrat	1.42 automatsko prevođenje 1.43 doslovno prevođenje 1.44 književno prevođenje 1.45 komunikacijski model prevođenja 1.46 ljudsko prevođenje 1.47 pismeno prevođenje 1.48 simultano prevođenje 1.49 slobodno prevođenje 1.50 strojno prevođenje računalno prevođenje 1.51 traduktologija 1.52 univerzalni prevodilac	1.53 starohrvatski jezik 1.54 filološke škole 1.55 zagrebačka filološka škola 1.56 zadarska filološka škola 1.57 riječka filološka škola 1.58 škola hrvatskih vukovaca 1.59 štokavski hrvatski književni jezik 1.60 čakavski hrvatski književni jezik 1.61 kajkavski hrvatski književni jezik 1.62 ozaljski književno-jezični krug

These terms will not appear as headwords of entries or subentries in *Mrežnik*. However, terms belonging to basic linguistic terminology and orthography will appear in *Mrežnik* as well as in *Jena*. Some terms belonging to basic linguistic terminology are shown in the text bellow. Figure 1 shows an extraction of the wordlist in *Jena*.



infinitiv	jezikoslovlje	filologija	opće jezikoslovlje (lingvistika)
imperfekt	jezikoslovlje	filologija	opće jezikoslovlje (lingvistika)
imperativ	jezikoslovlje	filologija	opće jezikoslovlje (lingvistika)
imenica	jezikoslovlje	filologija	opće jezikoslovlje (lingvistika)

Figure 1. General linguistic terminology from *Jena*

These linguistic terms will also be entries in *Mrežnik*.

Corpus-based

Both *Mrežnik* and *Jena* are corpus-based, and not corpus-driven. This means that the corpus and all data extracted from it serve only as guidelines. The *Glossary of E-lexicographic Terminology* on the *Mrežnik* website ihjj.hr/mreznik defines a corpus-based dictionary as follows: *a dictionary for which the lexicographer uses a corpus, but can freely decide what should be included in the dictionary,* *allowing the dictionary to be supplemented with words from other sources if necessary, as well as collocations and meanings not attested in the corpus.* The reason for this approach is that neither of the corpora on which *Mrežnik* is based (*Croatian Web Repository* online corpus (riznica.ihjj.hr/index.hr.html) and *Croatian web corpus – hrWaC* (nlp.ffzg.hr/resources/corpora/hrwac/) are representative of the Croatian language (hrWaC is primarily based on the colloquial and journalist style and *Croatian Web Repository* on the literary style), they are not corpora of the standard language nor are they balanced corpora. It follows that, in composing an entry, lexicographers can add meanings to a particular entry or to the collocation field even if they do not appear in the corpus.

Data extraction from the corpora for *Mrežnik* as well as for *Jena* is performed with the SketchEngine web tool, which allows the display of lexeme context through WordSketches, the most common collocations sorted into syntactic categories and the discovery of good examples of word usage or collocations. After lexicographic processing of *Mrežnik* is completed, the data will be exported from TLex to the web application and the CLARIN European science infrastructure repository (clarin.si repository and the github.com public data management system). This will make *Mrežnik* available for use both via a web application and for machine implementation by downloading data from the CLARIN repository.

Jena is based on the corpus *Jezikoslovlje* composed specially for the needs of the project. It consists of a corpus of linguistic papers and monographs compiled under SketchEngine. The *Jena* corpus which has been compiled by project members and collaborators is the corpus of standard language (in the field of linguistics) but it is as yet not representative enough. Moreover, on many modern linguistics topics there are not many texts in Croatian and many Croatian terms have to be coined by the authors (specialists of the particular linguistic field) themselves. From this corpus a term list has been compiled which contrasted the words appearing in the corpus with the words from *hrWaC*. The basic term list is still the one created by project collaborators but it will be checked against the one created by Sketch Engine from the corpus so *Jena* will also be corpus-based. The *Jena* corpus is also helpful when creating definitions and deciding on the normative status of synonymous words.

Normativity

The Croatian Web Dictionary – Mrežnik is a normative dictionary and *Jena* is a normative terminological database. The normative nature of *Mrežnik* is apparent in the following: 1. the selection of entry-words, 2. giving normative advice in all three modules, 3. the selection of forms acceptable by the standard language norm in the grammatical block, 4. the selection of examples (the dictionary collaborators try to select examples with no language errors while examples with language errors are edited), 5. the accentuation of entry-words and forms in the grammatical block according to the standard language norm.

The most important normative aspect of *Jena* is differentiating between the preferred, allowed, non-preferred, obsolete, and jargon terms (as will be shown in the examples below). If needed normative advice is given in the field note, e.g. why the preferred term is *točka sa zarezom* and not *točka-zarez* as shown in table 2.

Word entries vs. term entries

Two important meeting points of *Mrežnik* and *Jena* are 1. General linguistic and orthographic terminology is presented in *Mrežnik* and 2. Within the *Mrežnik* project, a *Glossary of E-lexicographic Terminology* is compiled (ihjj.hr/mreznik/page/pojmovnik/6/).

General orthographic terminology in *Mrežnik* and *Jena*

Table 2 illustrates the structure of the entries *točka* (period) and *točka sa zarezom* (semi-colon) in *Jena* and compares them to the respective entry or subentry in *Mrežnik*:

Table 2. Entries *točka* (period) and *točka sa zarezom* (semicolon) in *Jena* and *Mrežnik*

1.63 Jena	***1.64 Mrežnik***
1.65 točka 1.66 **unesen:** 01.08.2019, 20:57 **faza obradbe:** urednik pregledao **status naziva:** preporučeni naziv **definicija:** pravopisni znak koji stoji na kraju rečenice te iza kratica i rednih brojeva **vrelo definicije:** Jozić, Željko i dr. 2013. Hrvatski pravopis. Institut za hrvatski jezik i jezikoslovlje. Zagreb. **područje:** jezikoslovlje **potpodručje:** pravopis **jezična odrednica:** imenica **rod:** ženski **broj:** jednina 1.67 **istovrijednica - engleski:** period; full stop 1.68 **njemački**: Punkt 1.69 **francuski**: point 1.70 **ruski**: то́чка 1.71 **švedski:** punkt 1.72 simbol: . **poveznica:** http://pravopis.hr/pravilo/tocka/55/	**1.73 točka** 1.74 *pravop.* Točka je pravopisni znak (.) koji stoji na kraju rečenice te iza kratica i rednih brojeva. 1.75 - Definicija mora počinjati malim slovom i nema točku na kraju. 1.76 - Argument koji govori u prilog tomu da se paratakička rečenica ne razlikuje samo formalno od dviju rečenica, tj. da se ne može promatrati kao dvije rečenice koje su odijeljene točkom, odnosno koje se od dvorečeničnog ustrojstva razlikuju samo formalno. 1.77 **Koordinacija**: točka i crtica, točka i uskličnik, točka i zarez 1.78 **Poveznica** *Hrvatski pravopis*: http://pravopis.hr/pravilo/tocka/55/
1.79 **Comparison:** *Mrežnik* and *Jena* have *točka* as a headword of the entry. *Jena* has only one meaning of *točka*, while *Mrežnik* has many meanings only one of which is the meaning in the orthographic sense. *Mrežnik* also has many subentries of *točka* one of which is *točka sa zarezom* (semicolon). They have similar definitions, but while *Jena* gives the source of the definition *Mrežnik* gives examples and collocations (coordination). These examples are taken from the *Jena* corpus as it was difficult to find adequate examples from two corpora on which *Mrežnik* is primarily based. Both are connected to the same paragraph from *Croatian Orthography Manual*. In *Jena* equivalents in English, German, French, Russian, and Swedish are given. While in *Mrežnik* the sign (.) is a part of the definition given in brackets due to the very strict structure of the terminological database (brackets cannot be included in the definition) in *Jena* it is included in a special symbol field.	
1.80 točka sa zarezom 1.81 **unesen:** 01.08.2019, 20:58 **faza obradbe:** urednik pregledao **status naziva:** preporučeni naziv **definicija:** pravopisni znak koji se piše pri jačemu odvajanju od onoga koje označuje zarez, a slabijemu od onoga koje označuje točka **vrelo definicije:** Jozić, Željko i dr. 2013. *Hrvatski pravopis*. Institut za hrvatski jezik i jezikoslovlje. Zagreb. **područje:** jezikoslovlje **potpodručje:** pravopis **jezična odrednica:** višerječni naziv 1.82 **istovrijednica - engleski:** semicolon 1.83 **njemački**: Semikolon 1.84 **francuski**: point-virgule 1.85 **ruski**: то́чка с запято́й 1.86 **švedski**: semikolon 1.87 **simbol**: ; 1.88 **napomena:** U hrvatskome pravopisnom nazivlju u istome se značenju upotrebljavaju nazivi *točka-zarez i točka sa zarezom*. Budući da u nazivlju istoznačenice nisu poželjne, a polusloženice se ne uklapaju u strukturu hrvatskoga jezika te ih je, kad je to moguće, bolje zamijeniti istoznačnim nazivom drukčije strukture, prednost se daje nazivu *točka sa zarezom*. 1.89 **nepreporučeni naziv:** točka-zarez **poveznica:** http://pravopis.hr/pravilo/tocka-sa-zarezom/62/	1.90 **točka sa zarezom** pravop. 1.91 Točka sa zarezom pravopisni je znak (;) koji se piše pri jačemu odvajanju od onoga koje označuje zarez, a slabijemu od onoga koje označuje točka 1.92 - Definicije su u kurzivu i međusobno su odvojene zarezom, a sinonim koji nije u kurzivu odvojen je točkom sa zarezom. 1.93 - Veoma je često u engleskome tekstu uz veliko slovo u okomitome nabrajanju i točka sa zarezom. 1.94 •**normativna napomena**: U hrvatskome pravopisnom nazivlju u istome se značenju upotrebljavaju nazivi *točka-zarez* i *točka sa zarezom.* Budući da u nazivlju istoznačenice nisu poželjne, a polusloženice se ne uklapaju u strukturu hrvatskoga jezika te ih je, kad je to moguće, bolje zamijeniti istoznačnim nazivom drukčije strukture, prednost se daje nazivu *točka sa zarezom.* 1.95 **Mrtvi sinonim**: točka-zarez 1.96 **Poveznica**: *Hrvatski pravopis*: http://pravopis.hr/pravilo/tocka-sa-zarezom/62/ 1.97

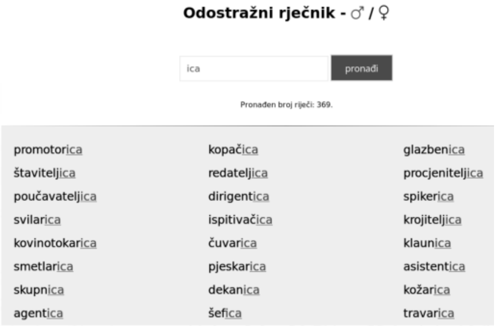
1.98 **Comparison:** In *Jena točka sa zarezom* (semicolon) is an entry while in *Mrežnik* it is a subentry of the entry *točka*. The reason for this is that a multiword term has the same terminological status as a single word term. They have similar definitions but only *Jena* states that the source of the definition is the *Croatian Orthographic Manual*. *Mrežnik* gives examples from the corpus while *Jena* has no examples. *Mrežnik* gives another synonymous term *točka-zarez* as a „dead synonym" which means it is not an entry in *Mrežnik*. In *Jena* there are no synonymous entries and *točka-zarez* is given as a non-preferred term. Both sources give the same explanation why *točka sa zarezom* is preferred to *točka-zarez* but this explanation occurs in the note field in *Jena* and in the field normative advice in *Mrežnik*. However, both are connected to the paragraph on semicolon from the *Croatian Orthographic Manual.* Both *Mrežnik* and *Jena* state that this term belongs to the field of orthography. In *Jena* equivalents in English, German, French, Russian, and Swedish are given. Examples in *Mrežnik* are taken from the *Jena* korpus, as it was difficult to find an adequate example in the two corpora on which *Mrežnik* is primarily based.

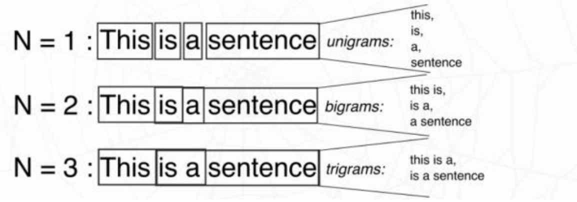
Similar results could be shown when comparing some other entries of general linguistic terms, e.g. *imenica* (noun), *padež* (case), *sklonidba* (declension), *sintaksa* (syntax).

Glossary of e-lexicography and *Jena*

Glossary of E-lexicography compiled within the *Mrežnik* project and in collaboration with the *Jena* project consists of names and terms relevant for e-lexicography. This *Glossary* is an important source for *Jena* as from it most of the terms (not names) are taken over into the *Jena* database. In table 3 the comparisons of the entries *odostražni rječnik* (reversed dictionary) and *n-gram* is shown in *Jena* and the *Glossary of e-lexicography*.

Table 3. Entries *odostražni rječnik* (reverse dictionary) and *n-gram* (n-gram) in *Jena* and *Glossary of e-lexicography*

1.99 Jena	***1.100 Glossary***
1.101 odostražni rječnik 1.102 **unesen:** 04.08.2019, 18:12 **faza obradbe:** urednik uređuje **status naziva:** preporučeni naziv **definicija:** rječnik u kojemu su riječi abecedirane od kraja riječi **vrelo definicije:** Pojmovnik, Mrežnik. http://ihjj.hr/mreznik/page/pojmovnik/6/. **područje:** jezikoslovlje **potpodručje:** e-leksikografija i korpusno jezikoslovlje **dopušteni naziv:** odostražnik **jezična odrednica:** višerječni naziv **istovrijednica - engleski:** reverse dictionary 1.103 **njemački**: rüchläufiges Wörterbuch 1.104 **francuski**: dictionnaire inverse 1.105 **ruski**: Обратный словарь 1.106 **švedski**: baklängesordbok; finalalfabetisk ordbok 1.107 **poveznica**: https://borna12.gitlab.io/odostraznji/ 1.108 **napomena:** Nazivu *odostražni rječnik* daje se prednost pred nazivom *odostražnik* zbog sustavnoga odnosa s nazivljem ostalih vrsta *rječnika (opći rječnik, posebni rječnik, abecedni rječnik, normativni rječnik, deskriptivni rječnik, terminološki rječnik* itd.)	1.109 **odostražni rječnik** (engl. reverse dictionary) rječnik u kojemu su riječi abecedirane od kraja riječi Rückläufiges Wörterbuch des Serbokroatischen (1965. – 1967.) mrežno je dostupan na https://www.uibk.ac.at/slawistik/institut/matesic.html. Demoinačica odstražnoga rječnika naziva za vršitelje/vršiteljice radnje (https://borna12.gitlab.io/odostraznji-mz/, izradio Josip Mihaljević): 1.110  **Odostražni rječnik - ♂ / ♀** ica Pronađen broj riječi: 369. promotorica kopačica glazbenica štaviteljica redateljica procjeniteljica poučavateljica dirigentica spikerica svilarica ispitivačica krojiteljica kovinotokarica čuvarica klaunica smetlarica pjeskarica asistentica skupnica dekanica kožarica agentica šefica travarica 1.111 1.112 1.113 **odostražnik** v. odostražni rječnik
1.114 **Comparison:** *Jena* has only a definition and additional information is given in the note section. In the note the reasons for selecting *odostražni rječnik* as the preferred term are explained. In *Jena* equivalents in English, German, French, and Russian are given.	
1.115 n-gram 1.116 **unesen:** 04.08.2019, 16:41 **faza obradbe:** urednik pregledao **status naziva:** preporučeni naziv **definicija:** sekvencija određene duljine koju sačinjavaju znakovi ili riječi koje se	1.118 **n-gram** sekvencija određene duljine koju sačinjavaju znakovi ili riječi koje se pojavljuju unutar teksta; pri radu s korpusima n-grami se odnose na sekvencije riječi; unigram je jedna riječ, bigram je sekvencija od dvije riječi, trigram je sekvencija od tri riječi itd.

pojavljuju unutar teksta korpusa **vrelo definicije:** Pojmovnik, Mrežnik. http://ihjj.hr/mreznik/page/pojmovnik/6/. **područje:** jezikoslovlje **potpodručje:** e-leksikografija i korpusno jezikoslovlje **podređeni pojam:** bigram; trigram; unigram **jezična odrednica:** imenica **rod:** muški **broj:** jednina **istovrijednica - engleski:** n-gram **istovrijednica - njemački:** N-Gramme **istovrijednica - francuski:** n-gramme **istovrijednica - ruski:** N-грамма 1.117 **švedski**: n.gram	1.119  N = 1 : This is a sentence unigrams: this, is, a, sentence N = 2 : This is a sentence bigrams: this is, is a, a sentence N = 3 : This is a sentence trigrams: this is a, is a sentence
1.120 Comparison: In *Jena unigram, bigram,* and *trigram* are added as subordinate terms and they have a separate definition. In the *Glossary* they are explained under *n-gram* but also have separate definitions in the glossary. The *Glossary* is added as a source in *Jena*. In *Jena* equivalents in English, German, French, Russian, and Swedish are given.	

Jena vs. *Mrežnik*: entry structure

From the general structure of *Struna* these fields have been activated for *Jena* (Table 4).

Table 4. Fields in *Jena*

1.121 Field	**1.122 Explanation**
1.123 entry word	1.124 can be a multiword entry
1.125 grammatical data	1.126 only word class and gender and number for nouns
1.127 definition	1.128 with *genus proximum* and *differentia specifica*, not a whole sentence, starts with the small letter and does not end with a period; one term can have only one definition.
1.129 field, discipline	1.130 subfields of linguistics, e.g. generative linguistics, cognitive linguistics, pragmatics, etc.
1.131 synonyms	1.132 divided into preferred terms, allowed terms, depicted terms, obsolete terms, and jargon terms
1.133 antonyms (added for the purpose of this project)	1.134 defined in *Jena* by a similar definition
1.135 subordinate terms	1.136 defined in *Jena*
1.137 source of the definition	1.138 if the definition was taken over from a source and not formed by the compiler the source should be stated
1.139 equivalents in English, Russian, French, German	1.140 written (or checked) by experts for each of the languages
1.141 abbreviation or acronym	1.142 given if any
1.143 connected to other sources	1.144 the entries are often connected to the *Croatian Orthographic Manual* or *Croatian School Grammar*, sometimes they are connected to other sources, e.g. articles in the journal *Hrvatski jezik*
1.145 note	1.146 in the note relevant additional information is given
1.147 phase in the compilation process are recorded	1.148 different phases are: written by the author, checked by the editor, checked by the terminologist, checked by the language editor, finished

The diagram of the structure of *Mrežnik* is described in detail in Hudeček, Mihaljević, 2019a. In table 5 the main differences between the structure of *Jena* and *Mrežnik* are shown:

Table 5. Differences between the structure of *Jena* and *Mrežnik*

1.149 Jena	***1.150 Mrežnik***
1.151 headword – can be multiword, not accentuated	1.152 no multiword headwords, headwords and forms are accentuated
1.153 grammatical data, word class or multiword – for nouns gender and number	1.154 gives much more grammatical data as well as accentuated forms
1.155 definition – one headword has only one definition, if needed the same headword has multiple entries	1.156 a headword can have multiple senses and definitions. very often the same entry has many meanings and only one belongs to the field of linguistics (e.g. *crtica*, *točka*, *atribut*) and sometimes the same word has more than one meaning in the field of linguistics (e.g. *pravopis*, *rječnik*, *fonologija*)

1.157 field, discipline, has a list of disciplines and sub-disciplines	1.158 differentiates between linguistics, grammar and orthography, doesn't differentiate between sub-disciplines
1.159 synonyms – differentiates between the preferred term, allowed term, non-preferred term, obsolete term, and jargon term	1.160 gives synonyms, differentiates between synonyms that are dictionary entries and that are not (synonyms and dead synonims), doesn't differentiate between the status of synonyms
1.161 antonyms – all given antonyms are dictionary entries	1.162 differentiates between antonyms which are dictionary entries and which are not (dead antonyms)
1.163 subordinate terms – a very important field for building the terminological system	1.164 sometimes gives subordinate terms
1.165 source of the definition	1.166 doesn't give data on the source of definition
1.167 equivalents in English, Russian, French, German, and Swedish	1.168 doesn't have equivalents in foreign languages
1.169 abbreviation or acronym given in a separate field	1.170 abbreviations and acronyms are given as synonyms and not in a separate field
1.171 symbols are given in a separate field	1.172 symbols, if needed, are included in the definition
1.173 connected to other sources – mostly connected to *Croatian School Grammar*, articles from the journal *Hrvatski jezik* and *Croatian Orthography Manual*[5]	1.174 connected to a number of sources[6]
1.175 additional information given in the note	1.176 differentiates between the pragmatic note and the normative note (language advice)
1.177 records the phase in the compilation process	1.178 doesn't state explicitly the phase in the compilation process
1.179 context	1.180 examples and collocations

An important difference between *Mrežnik* and *Jena* is the approach to collocations. In *Mrežnik* they have a separate field where they are introduces by questions, e.g. *What is xxx like?*, *What does xx do?*, *What can we do with xxx?*, *Coordination*, *What is mentioned in connection with xxx?* In *Jena* there is no special collocation field and they can be either introduced as subordinate terms which than have separate entries, explained in the note or ignored.

Results of the comparison

The results of the comparison prove the hypothesis that two such projects as *Jena* and *Mrežnik* can be compared and that they can mutually profit from each other and such a comparison. The results of the comparison are shown in table 6.

Table 6. Comparison of *Jena* and *Mrežnik*

Points of comparison	***Mrežnik***	***Jena***	**Benefits**
wordlist/termlist	terms extracted from the two corpora and supplemented by the lexicographers	termlist compiled by field specialists and supplemented by corpus data	*Mrežnik* wordlist was checked against the *Jena* termlist and all terms belonging to general vocabulary are included into *Mrežnik* *Jena* termlist was supplemented by the terms from the *Glossary of e-lexicography*
corpus	corpus-based	corpus-based (but the role of the corpus is not as important as in *Mrežnik*)	for linguistic terms in *Mrežnik* examples can be taken form the *Jena* corpus
normativity	normative dictionary, normativity expressed in the normative advice note	differentiates between preferred, allowed, non-preferred, obsolete, and jargon terms	normative advice given in *Mrežnik* and in *Jena* is always the same (it can be explained differently), i.e. once the normative status of a term is determined within one project the same normative status is given to the term in the other

[5] See in Hudeček, Mihaljević 2017c; Jozić et al., 2013.

[6] See Hudeček and Mihaljević in print.

			project
the structure of dictionary entries	The structure of dictionary entries is determined by the project head in collaboration with project members	the structure is only partially flexible as it is limited by the *Struna* database	projects have different structure but all relevant data determined by the research on either of the projects could be included into the other project; the experience on *Mrežnik* resulted in the inclusion of the field antonyms in *Jena*

Conclusion

The aim of this paper is to show how two projects conducted at the same time in the same institution can influence each other and how experience with one project as well as the corpus and data from one project can help the other project. As we started working on this paper at the very beginning of *Jena*, this gave us the perfect opportunity to test the hypotheses that the work on a terminological project can help the work on a general e-lexicographic project as well as benefit from it. Despite many differences, *Mrežnik* and *Jena* are closely connected as the *Glossary of E-lexicography* serves as one of the sources for *Jena* and as some definitions of general linguistic and orthographic terms from *Mrežnik* can serve as a starting point for composing the entry of the same headword in *Jena* (as shown above). We also found the corpus compiled for *Jena* very useful for collocations and examples in *Mrežnik* as with some frequently used words which have many meanings in the general language (as for example *točka* shown above as well as *čestica*, *prilog*, *prijedlog*) it was difficult to find the adequate example in the general corpus.

At the end of Jena, entries from *Mrežnik* will be connected with external links to entries in *Jena* as is *Mrežnik* already connected to some other finished *Struna* projects. Some other projects conducted at the Institute are connected to one or both of the analysed projects, e.g. the project *Orthographic Manual of Religious Terminology* is connected with *Mrežnik* as religious terms also form an important part of *Mrežnik*. The project *Male and Female in the Croatian Language* is strongly connected to both of the analysed projects as in *Mrežnik* each noun denoting a male person is connected to the noun denoting a female person[7] and in *Jena* special attention is paid to professional nouns in the field of linguistics (male and female).

Acknowledgments

This paper is written within the research project *Croatian Web Dictionary – MREŽNIK* (IP-2016-06-2141) and the project *Croatian Linguistic Terminology – Jena* (Struna-2017-09-05), both financed by the Croatian Science Foundation.

References

Hudeček, L. (2018). Izazovi leksikografske obrade u jednojezičnome mrežnom rječniku (na primjeru Hrvatskoga mrežnog rječnika – Mrežnika). // Visnyk of Lviv University: Series Philology 69, 29-38

Hudeček, L., Mihaljević, M. (2017a). A New Project – Croatian Web Dictionary MREŽNIK. // The Future of Information Sciences. INFuture2017, Integrating ICT in Society / Atanassova, I. et al. (eds.). Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, 205-213

Hudeček, L., Mihaljević, M. (2017b). Hrvatski mrežni rječnik – Mrežnik. // Hrvatski jezik 4, 4, 1-7

Hudeček, L., Mihaljević, M. (2017c). Školska gramatika hrvatskoga jezika. Zagreb: Institut za hrvatski jezik i jezikoslovlje

Hudeček, L., Mihaljević, M. (2018a). Croatian Web Dictionary Mrežnik: One year later – What is different? // Proceedings of the Conference on Language Technologies & Digital Humanities / Fišer, D., Pančur, A. (eds.). Ljubljana: Oddelek za prevajalstvo, Inštitut za novejšo zgodovino, 106-113

Hudeček, L., Mihaljević, M. (2018b). Hrvatski mrežni rječnik – Mrežnik: Upute za obrađivače. http://ihjj.hr/mreznik/uploads/upute.pdf (27.10.2019)

Hudeček, L., Mihaljević, M. (2019a). Croatian Web Dictionary – Mrežnik – Linking with Other Language Resources. // Electronic lexicography in the 21st century – Proceedings of the eLex 2019 conference / Kosem, I. et al. (eds.). Leiden: Lexical Computing CZ s.r.o, 72-98

Hudeček, L., Mihaljević, M. (2019b). Profesijski nazivi u hrvatskoj e-terminografiji i e-leksikografiji. // Studia lexicographica 13, 24, 75-95

Institut za hrvatski jezik i jezikoslovlje. Glossary of E-lexicographic Terminology. 14.10.2019. http://ihjj.hr/mreznik/page/pojmovnik/6/ (30.08.2019)

Institut za hrvatski jezik i jezikoslovlje. (2019). Hrvatsko jezikoslovno nazivlje – Jena. 23.10.2019. http://ihjj.hr/jena/ (30.08.2019)

[7] More on the topic see in Hudeček, Mihaljević, 2019b.

Institut za hrvatski jezik i jezikoslovlje. (2010). Hrvatsko strukovno nazivlje – Struna. 16.09.2010. http://struna.ihjj.hr/ (30.08.2019)

Institut za hrvatski jezik i jezikoslovlje (2017). Muško i žensko u hrvatskome jeziku. 1.11.2017. http://ihjj.hr/projekt/musko-i-zensko-u-hrvatskome-jeziku/72/ (20.08.2019)

Institut za hrvatski jezik i jezikoslovlje. Religijski pravopis. 5.11.2015. http://ihjj.hr/projekt/religijski-pravopis/23/ (20.08.2019)

Jozić, Ž. et al. (2013). Hrvatski pravopis. Zagreb: Institut za hrvatski jezik i jezikoslovlje

https://doi.org/10.17234/INFUTURE.2019.5

Review paper

Enhancing Encyclopedic Characteristics Using Geotagging
Why It Matters?

Jasmina Tolj
The Miroslav Krleža Institute of Lexicography, Zagreb, Croatia
jasmina.tolj@lzmk.hr

Ivan Smolčić
The Miroslav Krleža Institute of Lexicography, Zagreb, Croatia
ivan.smolcic@lzmk.hr

Zdenko Jecić
The Miroslav Krleža Institute of Lexicography, Zagreb, Croatia
zdenko.jecic@lzmk.hr

Summary

Through the last couple of decades, encyclopaedias have transformed significantly, becoming digital and born-digital, full of multimedia, and today among other characteristics being well connected via hypertext and metadata. At the same time, in digital humanities an increasing emphasis is being put on mapping or geotagging archival data, with special emphasis on networking projects and international cooperation. Digital platforms are being designed for publishing, describing, presenting, searching or browsing through historical sources. Some projects take the temporal aspect further, enabling not just retrieving historical data but also offering navigation through space and time via interactive maps. Encyclopaedias joining such initiatives would allow for new ways to explore its content and connecting encyclopaedic knowledge with specific artefacts and locations so users could visit them. Some encyclopaedic projects include a similar approach to their content, but further development is needed. In this paper, authors analyse existing conditions among web-based encyclopaedic projects (such as the Brockhaus Encyclopaedia and the Slovenian biographical lexicon), the data types appropriate for geotagging, opportunities and perspectives. This paper will also explain how using geotagging enhances encyclopaedias' characteristics. This would contribute to further encyclopaedia's interactivity and allow for new ways for users to explore content and learn, in turn contributing to its greater usage.

Key words: web-based encyclopaedia, geotagging, geospatial anchoring

Introduction

Starting from their very beginnings, encyclopaedias have introduced a new method of organizing and structuring knowledge and were thusly designed as a high-quality, simple, and quickly available source of information. As such, encyclopaedic works are characterized by accuracy, relevance, objectivity, comprehensiveness, credibility, timeliness, consolidation, complexity and structure (Jecić, 2013). With technological advancements, encyclopaedias first became digital, mostly meaning that previous works were digitized and supplemented by multimedia, and later have developed further to become comprehensive web-based projects. Wikipedia, initiated in 2001 as the global multilingual web-based edition based on mass collaboration, was among the first to explore a new way of thinking about encyclopaedic works. It was the first to offer open access to its knowledge, making it available for anyone with an internet connection. Encyclopaedic characteristics have since developed as well and now hold an even greater epistemological value regarding its update possibilities (timeliness), collaboration potential, being unlimited in scope, having far more possibilities for information retrieval and options for content interconnectivity (Smolčić et al., 2017).

Further advancement, especially in digital humanities and archival communities, allow for continued development. Increasing emphasis is being put on mapping or geotagging data, with special emphasis on networking projects and international cooperation, such as the Topotheque[1] and the Time Machine[2] project. Such digital platforms are designed for publishing historical sources, such as photographs, documents or audio-visual records and then using interactive IT tools to describe, present, search or browse through them. As a data base of relevant information and a source of condensed knowledge, encyclopaedia must also keep up. Regarding encyclopaedias' future, Prelog (2010) writes about placing the developmental emphasis on more complex linking of articles, where expanding the original context in which some concept is introduced is enabling a dynamic approach. This would allow new creative ways of searching (or browsing) and using relevant information. Since much of the information contains a spatial dimension, it would be worth exploring what information can and should be linked to a geographical location, thereby enhancing search options, interactivity and comprehensiveness, thusly furthering the encyclopaedic concept.

The idea of enriching information on the internet with a spatial component was first introduced in the 1990s (Herring, 1994) which later developed into the term geoweb, relating to virtual maps on the internet and the ability to link abstract information (textual, multimedial or entire web pages) with geographic locations, allowing for search options based on geographic locations (Voloder, 2010). Encyclopaedic content, with its comprehensiveness and knowledge synthesis, surely offers plenty opportunity to be tagged to geographic locations (geotagged). The authors believe that encyclopaedias could play an even greater role in the so-called *knowledge ecosystem* and should therefore be further advanced by geotagging its content. This paper will present a brief overview of a few encyclopaedic and non-encyclopaedic projects that have in a way geotagged their content, and propose guidelines for further development in encyclopaedics. For encyclopaedias to reach their full potential in geotagging their content, however, it will be necessary to determine typology of data appropriate for geotagging and develop a more complete methodology for geotagging encyclopaedic content.

Scope and methodology of research

Three encyclopaedic and three non-encyclopaedic projects will be analysed in order to present the potential for use of spatial data and to what extent it is implemented in encyclopaedics. The non-encyclopaedic projects included are the Topotheque project (platform) which is an online archive operated in local entities, the eCultureMap[3] which is an online geographical knowledge map connecting and visualizing digital cultural objects, and the Time Machine project, a platform to build a map of European history that spans thousands of years. These projects all relate to archival science, part of social sciences just as encyclopaedics, and were chosen to give example of what can be done using, among other, geospatial data.

The encyclopaedic projects presented include the Slovenian Biography portal[4], which is an ongoing web-based project, the German Brockhaus encyclopaedia[5], which is the most significant German encyclopaedia and one of the largest in general among encyclopaedic works, and the Encyclopedia Virginia[6], which focuses on the history and culture of Virginia, USA. These three, biographical, general and national encyclopaedic projects, have in different ways put the geospatial component in use and therefore represent the practice of geotagging in encyclopaedic projects. Analysis of these six projects will allow insight into how to further improve web-based encyclopaedic projects using the spatial component of information and how this in turn enhances encyclopaedic characteristics.

The use of spatial features in digital humanities and archival science

Different archival and cultural communities have recognised a need to become more approachable to the general public and some to seek cooperation with gathering more materials to catalogue and save. One project that started as a private toolkit to manage (index, date and localize) a private footage has evolved into a collaborative online archiving platform for the public to save local and historically

[1] Topotheque https://www.topothek.at/en/ (26.7. 2019)

[2] Time Machine https://www.timemachine.eu/ (26.7. 2019)

[3] eCultureMap http://eculturemap.eculturelab.eu/eCulture14m/Map.html (30.7. 2019)

[4] Slovenska biografija https://www.slovenska-biografija.si/ (30.7. 2019)

[5] Brockhaus https://brockhaus.de (30.7. 2019)

[6] Encyclopedia Virginia https://www.encyclopediavirginia.org/ (30.7. 2019)

relevant material or knowledge with a European perspective – the Topotheque[7]. The platform allows users to create archives where image and file content is sorted, and therefore can be searched, by key words, date (timeline slider) and perspective[8] on the map[9]. There is usually additional information, as a description or information on the owner of materials (often time museums and other institutions, not just private persons). The Topotheque was designed to not only act as means to preserve culture and history, but to also serve as a research tool. V. Lemić (2019) points out that all local collections are aggregated on the Europeana, European Union's digital platform for cultural heritage, facilitating swift exchange, gathering and presenting of information and allowing for new connections to be made in presenting cultural heritage, and to be used in other cultural, scientific or educational projects.

The eCultureMap is also related to Europeana[10]. It is an interactive online map connecting and visualizing digital cultural objects[11]. It allows geographical overview of Europeana's content, the use of spatial tools for searching and browsing, and it consists of four components, the mapping (the interactive map, with automatic translation from English to native languages of metadata), route planning (for browsing objects along a selected path, with distances, images and links to detailed information), search (of Europeana content free text), and mobile component (using locational services with previous three components). The quality of added geographical metadata is crucial for efficacy. This map provides a valuable resource for creative industries, promotion of culture, education, cultural tourism and adds even more value to Europeana (Zakrajsek, Vodeb, 2014).

The most ambitious project among its peers is the Time Machine project that would map millennia of European historical and geographical evolution. It proposes to build a large-scale digitisation and computing infrastructure (through simulation, multi-scale modelling and artificial intelligence technologies) that would digitise great volumes of information from Europe's historical archives, museum, libraries and geo-historical datasets. The project is one of six that won The European Commission's pan-European competition for researchers addressing grand scientific and technological challenges of strategic importance for Europe and with potential to change the future (Aigner, 2019). One of many proposed project's outcomes is to identify larger patterns, correlations and connections that open new perspectives for in-depth analysis of the culture and history[12]. It is evident that mapping information, providing a new dynamic approach to research, and connecting institutions that safeguard history and culture has been put high on Europe's priority list.

The use of spatial features in encyclopaedic projects

Slovenian biography

The Slovenian biography portal gathers biographies of notable figures from Slovenia's history and was created by the Slovenian Academy of Sciences and Arts, its Scientific Research Centre, the Jožef Stefan Institute and Seven Past Nine Ltd. It is comprised of biographies from three biographical lexicons, two of which were first published in print: Slovenian Biographical Lexicon that was published 1925–91 (around 5,000 entries entirely digitized and available on the portal, edited by Cankar et al.), Primorska Slovenian Biographical Lexicon that was published 1974–94 (containing 4,400 entries, out of which about 40% digitized and published on the portal so far, edited by Jevnikar et al.) and New Slovenian Biographical Lexicon that started as a web-based lexicon in 2013 (currently containing around 500 entries, edited by Svetina et al.). The portal has over 9,000 entries and is an important reference work for research in Slovenian humanities, social sciences and the history of the natural sciences.

Slovenian Biography uses open software and standards (Erjavec, Dokler, Vide Ogrin, 2017). To encode and structure information XML is used as a markup language, as it is compliant with the Text

[7] It was initiated by Alexander Schatek, an entrepreneur and industrial designer, in cooperation with the International Centre for Archival Research (ICARUS) and Europ's co:op project – Community as Opportunity: the Creative Users' and Archives' Network.

[8] Meaning that a photograph's position is not only shown as a point on a map, but the perspective in which the tagged image or video was taken can also be presented.

[9] The portal uses an embedded Google Map.

[10] In fact, it re-uses its content.

[11] It was developed in the EU project Carare (Connecting Archaeology and Architecture in Europeana).

[12] The Time Machine factsheet (updated June 2019). https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60088 (6.8.2019)

Encoding Initiative guidelines for electronic text encoding and interchange (TEI, an extensive and flexible schema used to digitally represent texts)[13]. Among other (name, sex, occupation, etc.), data is encoded using the TEI module for places, which allows for detailed annotation. Information that is detailed includes settlement name, region type and name, country and geographical coordinates. Gazetteers are used to ensure the standard form of settlement names and additional information is encoded for settlements that no longer exist or have changed name.

This kind of detailed markup is then used to present information and allow for various browsing and searching options. Slovenian biography portal offers a search bar for simple and complex search queries, and many browsing options: alphabetical index, timeline (to select the period in which persons lived), by profession and occupation, status (nobility, academy member, etc.), born or died on this day, by family, and using the map[14]. The map presents the places of birth and death of all persons included in the lexicon, allowing users to browse by place of birth and death. When a user chooses a place on the map to explore, initial information given are the names of person (in *Surname, Name* form, that act as hyperlinks), divided into categories of *born here* and *died here,* along with years of birth and death. The hyperlinks link to the person's biography article in the lexicon. Other than that, the complex search option allows to search for persons by place of birth or death by name of settlement, region name or country. This allows for a new perspective on browsing through or researching Slovenia's remarkable persons.

Brockhaus Encyclopedia

In a paper discussing geospatial anchoring (geotagging) of encyclopaedia articles, Kienreich, Granitzer and Lux (2006) present how this was done for the German encyclopaedia Brockhaus. The work references the online version of the encyclopaedia at that time, which today does not offer the discussed features, but the work gives valuable insight into problems and possibilities of geotagging encyclopaedic content. This general encyclopaedia was published by Bibliographisches Institut and F. A. Brockhaus AG as the leading work in European German-speaking countries, with about 240,000 articles and the world atlas with 2,000,000 named geospatial references as well as detailed surface maps[15].

Encyclopaedia's articles were geotagged[16] to named geographical entities in the atlas, and Kienreich, Granitzer and Lux (2006) divided geospatial references chosen for an article in four categories: direct, indirect and symbolic geographical references, and temporal spatial references. A direct reference is formed when an articles topic is explicitly linked to a geographical entity, e.g. an article on a specific town is referenced to the same town on the atlas. An indirect geographical reference is formed when a geographical entity to be tagged (anchored) does not correlate to an article in the encyclopaedia's knowledge space. In a biographical article, for an example, a place of birth will be a geographical reference, but it might not have its own article (a non-German tow, for an example). Such geographical references create relations between encyclopaedia articles that would otherwise not be connected because they don't overlap in the topical but do in the geographical domain. In cases where an article does not name a specific geographical entity, a symbolic reference is created, for an exemplary instance of the object class in question, for the largest of such objects or for the most important such object (as decided by editorial staff ranking). Temporal spatial references are created for articles on historical events, which require making references in a time-dependent manner (e.g. reference to a country, but only in a specific time period).

The map representing this geotagged content[17] can be browsed and searched, as there is a search bar to search for geographical entities that then displays search results in a relevance-ranked list. Further options include focusing of a selected rectangle in the view, taking measurements, and overlapping satellite imagery, topological and geopolitical information. The authors point to challenges in encyclopaedical and spatial ambiguities, as well as name changes in geographical entities over time. They found that users benefit from geospatial anchoring by having one more option to navigate

[13] TEI: Text Encoding Initiative http://www.tei-c.org/ (30.7. 2019)

[14] The portal uses an embedded Google Map.

[15] Apart from detailed surface maps, the atlas also contains high-precision altitude information, not just on Earth, but on Moon and Mars as well.

[16] The authors refer to it as geospatial anchoring.

[17] Which utilises a web-based atlas client application (not Google Maps).

between encyclopaedia articles (through geospatial references), making the content better connected and the encyclopaedia more dynamic.

Encyclopedia Virginia

Encyclopedia Virginia is a project of Virginia Humanities and the Library of Virginia[18] focusing on the history and culture of Virginia. It is an ongoing project, first published in 2008. Encyclopaedia's entries are topical and biographical, accompanied by primary documents and media objects, including images, audio and visual clips, links to Google Street View tours of historic sites and museum objects (in partnerships with museums and cultural institutions in Virginia)[19]. Each entry contains rich metadata, allowing users to search and browse content in various ways, via map, static or interactive time lines and with *This Day in Virginia* feature. Future plans are to expand options with, among other, an option for mobile users to read encyclopaedia's entries appropriate to the geographical spot on which they stand.

The map[20] allows searching by search term, time span and over sixty categories (such as Slavery, Geography, Architecture etc.), or it can be visually browsed. Each marker gives a brief description and a link to a correlating entry, and sometimes a date, event type (death, birth etc.) or place (e.g. highway marker, a park, a stream). The encyclopaedia offers over sixty virtual tours of museums, historic mansions, plantations, former slave dwellings and other. The tours offer 360° views of interior, exterior and sometimes the surrounding area. This provides a more vivid approach to researching Virginia's past, especially the reminder of unjust treatment of people of colour, to general public.

Discussion

There are some similarities in how the Brockhaus Encyclopedia and the Slovenian biography use the geotagged content. Both display the information on a map and allow for new ways to search and browse content, along with making new connections were there were previously none, and allowing for a more dynamic browsing session. The Slovenia's biographical portal holds one type of encyclopaedic article, the biographical article, so naturally there are fewer opportunities to advance the content by mapping than there are in a general encyclopaedia such as the Brockhaus. This is where it becomes clear that anchoring information to geographical locations can provide new paths (connections) for users to explore. Where, for an example, a search for a certain writer can lead to a town of his birth, then the town is connected to some historical event, which in turn leads to a new historical figure and so on.

What this type of encyclopaedia could further include as a feature is to point to not only its own articles, but to geographical locations, such as artefacts in museums or on display in public (statues and such), or perhaps provide images of such objects and locations. The Encyclopedia Virginia offers a similar feature, if not more advanced, providing virtual tours. It, however, lacks other features, such as the map that would visually represent locations of all virtual tours. These described encyclopaedic projects use geographical component of information to further their users' experience and also to enhance encyclopaedia's characteristics.

Geotagging advances encyclopaedia's connectivity, search options, interactivity and adaptability. Connectivity is made better throughout encyclopaedia's own content, but it can also be enhanced by connecting to outside content, such as objects in open space or museums. Spatial metadata used in geotagging allows search options by place, and when presented on a map it also allows better interactivity and adaptability as it facilitates new ways of finding (or browsing) content. While geotagging enhances encyclopaedic characteristics, it also presents new research options in other branches of social science, such as history, political science or sociology, just as it does in non-encyclopaedic projects.

Regarding the described encyclopaedic projects, it can be seen that there is no significant attempt in making many connections to other institutions, while described non-encyclopaedic projects try hard to do so. Perhaps encyclopaedias should also try to be included in Europeana or similar aggregating

[18] The Library of Virginia became a partner of the project in 2012.

[19] Encyclopedia Virginia. https://www.encyclopediavirginia.org/about (30.7.2019)

[20] The website uses an embedded Google Map.

projects, such as the Topotheque and the eCultureMap are, because they too are a sort of *storage* of history and knowledge.

While searching for information on the internet a user is flooded with enough to never read through it all, what is available is unstructured and often unverifiable. Unlike an internet search, encyclopaedias provide a compendium of trusted and concise information and knowledge, where articles are prepared by experts in a given domain and are guaranteed to describe a topic accurately and exhaustively. Due to their basic characteristics (accuracy, objectivity, relevance, etc.), encyclopaedias have the potential to play an important role in informing and educating the public, which is of great importance in this so-called *post-fact era*, where evidence is often overpowered by emotion and confidence in institutions, expertise and the media is on the decline. The European Union has recognized the importance of encyclopaedias and had conducted an analysis of European countries' online encyclopaedias (Bentzen, 2018). It found disproportion between the sheer amount and verifiability of information and how this impacts the European democracy system and values. The analysis emphasized the importance of information availability regarding at least basic history and culture in the so-called *knowledge ecosystem* and the importance of encyclopaedias as crucial factors in avoiding manipulation as they hold great potential for informing users seeking information and knowledge. For these reasons it is crucial to keep advancing encyclopaedic science to maximize encyclopaedias' usability.

Geotagging in encyclopaedic projects is underutilized and many encyclopaedias do not use spatial tags at all[21], giving a lot of room for further research on this topic. For these reasons it is crucial to advance encyclopaedic science further and to create a more complete methodology for geotagging encyclopaedic content. To do that, it is necessary to investigate solutions to many challenges that arise, such as to determine which information is relevant for creating geospatial tags, developing automatic or semi-automatic tools to geotag existing projects, and most importantly determine typology of data appropriate for geotagging.

References

Aigner, T. (2019). Europe builds a Time Machine. Big Data of the Past is becoming a reality. Vienna: Insights, ICARUS – International Centre for Archival Research

Bentzen, N. (2018). Europe's online encyclopaedias. Equal access to knowledge of general interest? Bruxelles: European Parliamentary Research Service

Erjavec, T., Dokler, J., Vide Ogrin, P. (2017). Slovenian Biography. // Proceedings of the Second Conference on Biographical Data in a Digital World, (Vol-2119) / Fokkens, A., ter Braake, S., Sluijter, R., Arthur, P., Wandl-Vogt, E. (eds). Linz: CEUR-Workshop Proceedings, 16-21

Herring, C. (1994). An Architecture for Cyberspace: Spatialization of the Internet. Champaign: The United States Department of Defense, U.S. Army Construction Engineering Research Laboratory

Jecić, Z. (2013). Enciklopedički koncept u mrežnom okruženju. // Studia lexicographica 7, 2(13), 99-115

Kienreich, W., Granitzer, M., Lux, M. (2006). Geospatial anchoring of encyclopedia articles. // Proceedings of the International Conference on Information Visualisation. / Bilof, R. S. (ed). London: Institute of Electrical and Electrics Engineers, 211-215

Lemić, V. (2019). Topoteka – naša povijest, naš arhiv. http://www.skole.hr/upload/portalzaskole/newsattach/16617/Topoteka.pdf (6.8.2019)

Prelog, N. (2010). Od tko i što do kako i zašto – budućnost online enciklopedija. // Studia lexicographica 4, 2 (7), 164-176

Smolčić, I.; Tolj, J., Jecić, Z. (2017). Epistemological Value of Contemporary Encyclopedic Projects. // INFuture 2017: Integrating ICT in Society / Atanassova, I., Zaghouani, W., Kragić, B., Aas, K., Stančić, H., Seljan, S. (eds.). Zagreb: Faculty of Humanities and Social Sciences, University of Zagreb, 141-149

The Time Machine factsheet. (2019). https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60088 (6.8. 2019)

Voloder, I. (2010). Strategija razvoja geoweba s informacijskog, tehnološkog, kulturološkog i poslovnog stanovišta. Doctoral dissertation. Zagreb: Filozofski fakultet

Zakrajsek, J. F., Vodeb, V. (2014). eCultureMap – Link to Europeana Knowledge. // Communications in Computer and Information Science (Vol. 416 CCIS) / Bolikowski, L., Casarosa, V., Manghi, P., Goodale, P., Houssos, N., Schirrwagen, J. (eds.). Springer International Publishing Switzerland, 184-189

[21] Encyclopaedia Britannica, for instance, does not geotag its content.

https://doi.org/10.17234/INFUTURE.2019.6

Preliminary communication

A Corpus-Based Approach to Reevaluation of Croatian Verb Classification

Danijel Blazsetin
Faculty of Humanities and Social Sciences. University of Zagreb, Croatia
dblazset@ffzg.hr

Petra Bago
Faculty of Humanities and Social Sciences. University of Zagreb, Croatia
pbago@ffzg.hr

Summary
Croatian grammar textbooks have a long tradition of classifying verbs based on their morphosyntactic characteristics. Conclusions, such as the frequency or productiveness of a class, were drawn without having the insight into a big corpus. Corpora used in such descriptions were not described and were presumably made of literary works which is, in our opinion, describing a form of the Croatian language distant from its everyday use. The corpus used for analyzing verbs in this paper is hrWaC which contains 1.9 billion tokens and about 90,000 verbs. This corpus was selected with the intention of describing and analyzing a less formal and less standardized language This paper offers a corpus-based approach to the problem of verb classification and emphasizes the importance of NLP methods in the process of classification as they fasten and simplify it. The paper gives a brief introduction to verbs, their morphological characteristics and their classification. By extracting verbs from the Croatian web corpus hrWaC and processing them computationally, the paper gives an insight into the verb distribution in the Croatian language and points out some difficulties that were encountered during this study. Even though this paper aimed to reevaluate the existing data data, the present findings mostly confirm the claims of previous researches. A number of recommendations for future research are given, foremost, the need of the extension of the language material.

Key words: corpus linguistics, natural language processing, verb classification, grammar textbooks, Croatian language

Introduction

Reviewing Croatian grammar textbooks, one can find information about the frequency of Croatian verb classes. This paper uses modern technology tools and methods to reevaluate the existing statistics regarding Croatian verb classes. Applying natural language processing methods to this field, we can speed up the process of verb classification and get exact information i.e. the frequency of Croatian verb classes. It is important to mention that existing descriptions are mostly based on corpora of standardised Croatian language, whilst the frequency information given in this paper analyzes verbs from a web corpus which is comprised of diverse discourses, including texts written in the standardised version of the Croatian language as well as texts written in informal, colloquial version of the Croatian language. The presented model offers an automatic verb classification system that was applied and tested on the hrWaC web corpus. For the purposes of this paper, it was necessary to build three different groups of verbs that would simulate the corpus as a whole: *common verbs*, *occasional verbs*, *rare verbs*, as will be later explained in more detail. This paper also offers a comparative analysis of the existing works on frequency distributions of Croatian verb classes and the presented research.

Verb classification

Verb classifications are based on the verbs' morphological attributes. In Slavic languages there are two approaches to verb classifications: classifications founded on the verb's present tense base and classifications founded on the verb's infinitive base (Marković, 2012: 217-219). To understand the classifications, we must first define the morphological characteristics that determine a verb's class. If we compare a verb's infinitive form with its present tense form, we can see three elements: the verbs stem, the suffix that denotes its conjugational class and the derivational morph that indicates the verb's tense (Table 1).

Table 1. Comparison of the morphology of the infinitive and the present form.

infinitive form	*gled*-***a***-*ti* (Eng. *to watch)*)	*rad*-***i***-*ti* (Eng. to work)
present form	*gled*-***a***-*m* (Eng. I watch)	*rad*-***i***-*m* (Eng. I work)

Because of linguistic economy, words (including verbs) tend to organise themselves according to their similarities. Verb classifications aim to discover the patterns in the verbs' groupings and describe them (ibid: 197).

As mentioned, there are several ways of approaching the problem of verb classification. It seems like nowadays Croatian linguistics prefers the classifications based on the verbs' present tense form. Hence the presented model uses aforesaid classification as well (Bošnjak Botica, 2013: 65). We can find examples of this classification in the highly influential grammar textbook written by Josip Silić and Ivo Pranjković. Because of its prevalence, it is chosen to be the foundation of the verb class frequency analysis in this paper. The classification defines six verb types that can be divided into several classes (Silić, Pranjković, 2005).[1]

Related work

Information about the frequency and prototypness of verb classes can be found in Croatian grammar descriptions. However, it seems like the majority does not provide information about the corpus on which the classifications were applied to and as such prevents future researchers from comparing the diverging results of grammar textbooks and other related studies (cf. Babić et al., 1991; Raguž, 1997).

The corpus that Jelaska (2003) based her research on was the so called *Moguš's corpus*. Compared to today's corpora, it is small, and compiled from texts written in the standardised variant of the Croatian language. As such, it represents an artificial form of the language (cf. Tadić, 1997: 389-390). Hence we believe that information gathered from that corpus cannot describe the language as it is in its everyday use. Jelaska and Bošnjak Botica (2019) made an extensive research on verb class frequency which contains 24,538 verbs, but they do not provide any information about the corpus the research was based on. The main drawback of these studies is that they do not give any information about the corpus they are based on.

We think that dealing with language should always include modern technologies ie. natural language processing methods. Our research combines corpus linguistics as well as NLP methods, and presents an automated verb classifier that could be used on any given corpus in order to define the frequency of verb classes. Some of the biggest advantages of our approach to verb classification are its automatization, upgradeability and feasibility. The corpus used for the purpose of this paper is the hrWaC corpus.

Methodology

Corpus hrWaC contains 1.9 billion tokens, is made of HTML documents found on the *.hr* top-level domain and is the first of its kind of Croatian language (Ljubešić, Klubička, 2014). It is an annotated and a searchable web corpus that can be accessed via Sketch Engine.[2] We find it crucial that the corpus is based on documents found on the web as it means that it does not only represent the standardised version of Croatian, but also includes the language used in fashion magazines, newspapers, blogs, advertisements, user responses, forum discussions etc. Thus it reflects written language in its everyday use. A corpus of this kind can attest the jargon of different groups; speakers' doubt in using the correct word forms; the usage and frequency of loanwords; problems with orthography and trends in a language. Unfortunately, such corpora have their disadvantages as well.

[1] The verbs of the first type can be sorted into eighteen classes, but in this paper, we ignored those classes as they would not be useful for the comparative analysis.

[2] https://www.sketchengine.eu/

One of the biggest issue is that they are not representative as they do not include all types of texts (for example literary works are barely found in web corpora because they are copyright protected) and the documents that make up the corpus might not be reliable (cf. Fletcher, 2011).[3] In the process of making this model we ran into some *web corpora* specific issues that include incorrectly lemmatised or tagged words and improperly written words. This, on the one hand, means that the documents in the corpus did not overgo a process of selection, meaning that nothing is censored, but on the other hand it raises the problems of reliability of the statistics calculated from the corpus. Besides the benefits of such corpora, we would like to emphasize that one must always bear in mind the nature of alike corpora when he/she relies on its statistics.

Firstly, to classify verbs we had to gather information from the corpus hrWaC. As mentioned, to classify verbs one must *know* the verb's infinitive form and its present tense form. Using Sketch Engine we created a verb frequency list ie. a lemma frequency list of the verbs. Since Sketch Engine only allows the export of 1,000 verbs long lists, we had to define categories that could represent the corpus as a whole. We outlined three categories: *common verbs*, *occasional verbs*, *rare verbs*. Naturally, the *common verbs* category is constituted by the top of the frequency list and contains the 1,000 most frequent verbs. Then we defined the *rare verbs* category, which includes 1,000 verbs that occur from 18 to 25 times in the corpus. We wanted to exclude *hapax legomenon* as well as verbs that are occurring more than once, but are still very rare.[4] Then we had to define the *occasional verbs* category. As the lower bound of the frequency of the *common verbs* is 825 instances and the most frequent verb in *rare verbs* occur 25 times, the *occasional verbs* category had to be somewhere in between. In order to have a consistent methodology, it was mandatory to have a thousand verb category for the *occasional verbs* as well. If we try do define this category on the arithmetic middle of the two numbers we will have a category with only a few hundred members so we decided that this category would include verbs that occur from 140 to 375 times. Therefore the category *occasional verbs* contains 1000 verbs as well.

The three derived lists were merged and, as we only had their infinitive form, their present tense form was defined. This step was done manually.[5] Throughout this process verbs that were not suitable for the analysis were removed, resulting in a corpus that counts 2,588 verbs.[6]

The program compares a string to several regular expressions ie. words to patterns. The present tense form and the infinitive form of a verb are paired and are part of a list of all the verbs. The program compares the infinitive form and the present tense form of a verb to class-specific patterns. For example, the pattern for the infinitive form of the verbs in the fourth class of the third type (e. g. *držati* (Eng. *to hold*)) is defined as follows: _r'.*(š|č|ž|j|št|žd)ati\b'_. The value of this expression is *True* if the string ends with *šati, čati, žati, jati*, *štati* or *ždati*. The regular expression for the present tense form of the same class (e. g. *držim* (Eng. *I hold*)) is _r'.*im\b'_. This expression returns *True* if the string ends with *im*. If both of the expressions' values are *True*, the program classifies the verb (its infinitive and present form) into the matching class and removes it from the list which the program iterates through. Because some verbs would pass several regular expressions, there is a defined order of comparing the verbs with the patterns. For example, the verb *razgledavati* (Eng. *to sightsee*) both in its infinitive and present form *razgledavam* (Eng. *I sightsee*) matches the pattern of the first class of the fifth type ie. _r'.*ati_\b'-r'_.*am\b'_, but it belongs to the second class of the fifth type ie. _r'.*avati_\b'-r'.*am\b'. As it is seen, defining the order is mandatory and is a crucial step in the classification process. When the program is finished, the user gets a document with a *.txt* extension that contains the classified verbs.

[3] Even though these corpora are not representative in its narrower sense, they can, for example in our situation, be a representative corpus of the language on the web.

[4] We must not forget that we are working with a huge web corpus and that same incorrectly written words can occur many times in the corpus. Even the category *common verbs* contains incorrectly written verbs that had to be excluded from the statistics.

[5] We think that in the future, when dealing with huge ammount of data (verbs), we will automate the process of defining the present tense form by utilizing hrLeX (http://nlp.ffzg.hr/resources/lexicons/hrlex/).

[6] We excluded verbs which are not correctly written, which are wrongly lemmatised or which contain typographical errors.

Results and discussion

Before the comparison, we would like to give a brief overview of the existing frequency data of the verb classes. There is only a handful of grammar textbooks that contain information about verb frequency and a few research papers that give insight into verb class frequency. It seems like exact data regarding the number of verbs in a class does not preoccupy Croatian linguists (Marković 2012: 220). Instead of listing all the conclusions about verb class frequency found in grammar textbooks, we will only be illustrating how grammar textbooks inform the reader about verb class frequencies. In Babić et al. (1991) one can find such statements: ”There are approximately sixty verbs like *vidjeti-vidim* (Eng. *to see-I see*)[7] and two hundred more derivatives.” or ”There is a lot of verbs like *misliti-mislim* (Eng. *to think-I think*)”. Raguž (1997) states the following: ”There are a few hundred verbs of the type *vidjeti-vidim* (Eng. *to see-I see*)” and ”There are approximately 6,000 verbs like *misliti-mislim* (Eng. *to think-I think*)”. As we can see, the given information are not precise. There is more accurate and exact information in the works of Jelaska (2003) and Jelaska and Bošnjak Botica (2019). Jelaska (2003) categorized 16,000 of the most frequent verbs that were extracted from the *Moguš's corpus*, while Jelaska and Bošnjak Botica (2019) categorized 24,538 verbs, however the used corpus is unknown. In the following tables (Tables 2, 3 and 4) we can see the results of their research and their comparison with our research (Table 5).

Table 2. Percentage of the classes' representation in regards to all the verbs

Type	Class	100	100 (by type)	16,000 (by type)
a	*gledati-gledam* (Eng. *to watch*))	22%	22%	36%
i	*moliti-molim* (Eng. *to pray*)	26%	37%	30%
	voljeti-volim (Eng. *to love*)	6%		
	bježati-bježim (Eng. *to run away*)	5%		
e	*dignuti-dignem* (Eng *to lift*)	0%	12%	29%
	vjerovati-vjerujem (Eng. *to believe*)	4%		
	davati-dajem (Eng. *to give*)	1%		
	smijati se-smijem se (Eng. *to laugh*)	2%		
	plesati-plešem (Eng. *to dance*)	5%		
ø	*naći-nađem* (Eng. *to find*)		29%	5%

Source: Jelaska (2003: 56)

[7] The translations of the Croatian verbs were added by the authors.

Table 3. Number of verbs in the classes

Representative verbs	Class frequency	Verb type	Verb type frequency
gledati-gledam (Eng. *to watch*)	9,590		
		a	9,590
moliti-molim (Eng. *to pray*)	7,011		
vidjeti-vidim (Eng. *to see*)	509		
trčati-trčim (Eng. *to run*)	225		
		i	7,745
pisati-pišem (Eng. *to write*)	1,325		
smijati se-smijem se (Eng. *to laugh*)	337		
putovati-putujem (Eng. *to travel*)	2,621		
davati-dajem (Eng. *to give*)	67		
viknuti-viknem (Eng. *to yell*)	1,463		
		e1	5,813
naći-nađem (Eng. *to find*)	1,390		1,390
		e1+e2	7,203
Total number	24,538		24,538

Source: Jelaska, Bošnjak Botica (2019: 64)

Table 4. Number of verbs in the classes based on the research presented in this paper

		Representative verb	Number	Number by type	% by class	% by type
I. type[8]		*ići-idem* (Eng. *to go*)	278	278	10.7	10.7
II. type		*viknuti-viknem* (Eng. *to yell*)	96	96	3.7	3.7
III. type	1. class	*pisati-pišem* (Eng. *to write*)	143	172	5	6
	2. class	*pljuvati-pljujem* (Eng. *to spit*)	1			
	3.class	*grijati-grijem* (Eng. *to heat*)	28		1	
IV. type	1. class	*raditi-radim* (Eng. *to work*)	821	897	31.7	34.5
	2. class	*vidjeti-vidim* (Eng. *to see*)	49		1.8	
	3. class	*trčati-trčim* (Eng. *to run*)	27		1	
V. type	1. class	*kopati-kopam* (Eng. *to dig*)	810	952	31.3	36.2
	2. class	*proučavati-proučavam* (Eng. *to study*)	142		4.9	
VI. type	1. class	*kupovati-kupujem* (Eng. *to buy*)	49	187	1.8	6.6
	2. class	*smanjivati-smanjujem* (Eng. *to reduce*)	138		4.8	
∑			2,582+5	2,582+5	100	100

[8] The first verb type was not separated into classes. The Silić and Pranjković grammar textbook (2005) defines 18 classes in the first type. The criteria for the classes are really specific, hence we believe that it would be redundant to separate the verbs in the first type to classes as we will not use those numbers in the comparative analysis.

Table 5. The comparison of the works of Jelaska (2003), Jelaska and Bošnjak Botica (2019) and our research

	Jelaska (2003)	Jelaska and Bošnjak Botica (2019)	Our research
I. type	5%	5.66%	10.7%
IV. type	30%	32.02%	34.5%
V. type	36%	39.08%	36,2%
II. type	29%	23.23%	16.3%
III. type			
VI. type			

Firstly, we would like to emphasize that the research seen in Jelaska (2003) and Jelaska and Bošnjak Botica (2019) analyzes significantly more verbs than our research. However, it is our belief that, even though our research analyzes fewer verbs, due to it being based on three different frequency categories, it can serve as a valid element in the comparison. The statistics conducted in different studies do not differ much. This means that the verbs in hrWaC are similar to those in the *Moguš's corpus*. Thus, putting emphasis on rare verbs might give us a more exciting insight into verb classification. We shall highlight the differences and similarities between the existing statistics here. Verbs like *gledati-gledam* (Eng. *to see-I see*)[9] are the most frequent in every statistic and are followed by the verb type *misliti-mislim* (Eng. *to think-I think*). Classes inside the verb type *misliti-mislim* (Eng. *to think-I think*) differ though. According to Jelaska (2003), 6% of the verbs belong to the class *voljeti-volim* (Eng. *to love-I love*), while 5% to the class *trčati-trčim* (Eng. *to run-I run*). In Jelaska and Bošnjak Botica (2019) and our research, these percentages are 2% and 1%, respectively. It is interesting how the verbs *bosti-bodem* (Eng. *to stab-I stab*)[10] in Jelaska (2003) and Jelaska and Bošnjak Botica (2019) make only 5% of all the verbs, while in our research it is as high as 10.7%. It is our opinion that this percentage would gradually decrease if we added more verbs to our analysis.[11] It is usually said that the verb type *dignuti-dignem* (Eng. *to lift-I lift*) is frequent and productive (cf. Babić et al. (1991); Raguž (1997). However, Jelaska and Bošnjak Botica (2019) and this research found that only 5.96% and 3.7% of all verbs, respectively, belong to this type.

As shown by our analysis, the differences are modest, therefore we believe that comparing different frequency categories of our research could be useful and could give insight to the system of verb classification.

In this research, verbs like *misliti-mislim* (Eng. *to think-I think*) and *gledati-gledam* (Eng. *to see-I see*) are the most frequent in every frequency category. However, it has to be emphasized that while in the common verbs category *misliti-mislim* (Eng. *to think-I think*) makes 40.3% of all the verbs and *gledati-gledam* (Eng. *to see-I see*) only 28.8%, in the rare verbs category *misliti-mislim* (Eng. *to think-I think*) decreases to 30.8% and *gledati-gledam* (Eng. *to see-I see*) raises to 46.8%. The fact that there are verbs in the rare verbs category such as **odblokirati* (Eng. *to unblock someone on social media platforms?), *štrumpfetati* (Eng. *to act like Smurfette; to be an easy girl?*) indicates the prototypness of the class *gledati-gledam* (Eng. *to see-I see*). It is more likely that Croatian speakers will make up the word *štrumpfetati* and not *štrumpfetjeti*.[12]

As expected, the percentage of the verb type *bosti-bodem* (Eng. *to stab-I stab*) decreases with the growth of the number of the analysed verbs. However, atypicality does not always correlate with high frequency. This is supported by the fact that 5.7% of the verbs in the rare verbs category belong to the class *bosti-bodem* (Eng. *to stab-I stab*). Such verbs in the rare verbs category are: *rastresti* (Eng. *to shake up*) *prigristi* (Eng. *to have a bite*), **štići* (Eng. *to arrive*), *crpsti* (Eng. *to draw out*).

The *viknuti-viknem* (Eng. *to yell-I yell*) class is fairly low in all the categories: 2.3%, 3.7% and 5.4%.

[9] To avoid confusion, in this section we will not name a verbs' class to determine it, but a prototype of its class (eg. *gledati-gledam* (Eng. *to see-I see*) instead of type V. class 1.). This is necessary as Jelaska & Bošnjak Botica use a slightly different classification in their works.

[10] These verbs are traditionally classified into the first verb type. They are unique and unusual because their suffix that denotes its conjugational class is a zero morph (ø) and the stem cannot be seen from the infinitive verb form (eg. *jes-ø-ti, jed-e-m* (Eng. *to eat-I eat*)).

[11] We will discuss this statement below.

[12] However, this paper does not aim to define the prototypness of the Croatian verb classes.

Babić et al. (1991) mention that the class *kupovati-kupujem* (Eng. *to buy-I buy*) is big, however in our research there are only 49 verbs out of 2,587 which are classified in this category. Because of the verbal aspect pairs in Croatian language, we can indeed produce a lot of verbs that will belong to this category. However, it seems they are barely present in the written language of the Internet. It would be interesting to compare this frequency distribution to verbs extracted from a spoken corpus.

We analyzed 2,587 verbs, presenting the data as 2,582+5. The five isolated verbs belong to the irregular class and they do not fit into any of the classes. These verbs are: *biti-jesam* (Eng. *to be-I am*), *moći-mogu* (Eng. *to can-I can*), *spati-spim* (Eng. *to sleep-I sleep*), *zaspati-zaspim* (Eng. *to fall asleep-I fall asleep*) and *htjeti-hoću* (Eng. *to want-I want*).

Conclusion

In this paper we offer a corpus-based approach to the problem of verb classification in Croatian language extracting verbs from the hrWaC corpus. As we have seen there are studies that are based on a bigger corpus of verbs. However we believe that the uniqueness of our research lays in the fact that it is based on a web corpus, hence mirrors a variation of Croatian language that is similar to its everyday use. It has to be stated that a research which analyzes 3,000 verbs cannot reflect a true and comprehensive picture of the language, even if frequency categories were assembled. Thus the results of this research have to be dealt with reservations. However, it seems that a similar approach to the problems of verb classifications in Croatian language could shed light on some tendencies in the language and present new numerical data. Although this paper aimed to reevaluate the existing statistics regarding the number of verbs in various verb classes, it, first of all, points out the significance of the usage of NLP methods in language research and linguistics. The strength of the presented tool lays in its reusability and easy application. Future studies could fruitfully explore the issue of verb classification by analysing the corpus as a whole. We believe that a throughout analysis could either decisively confirm the existing data regarding verb classes or verify our initial hypothesis ie. the authors of Croatian grammar textbooks did not have access to this big amount of data so the information about the frequency of verb classes should be reevaluated. This study tried to offer an approach to the process of reevaluation. A verb analysis as shown in this paper could also be useful in the making of a language learning program for those learning Croatian as a second language as it is known that verbs of the same class have the same inflection, and derivational phenomena can also be generalised. Such approaches (ie. that take into account the prototypness and frequency of the verbs and their classes) in language teaching are already a trend and this type of clearly *digitally born data* could expand the previously proposed programs database. If, in the future, a corpus of contemporary and standard Croatian language is made, with the application of this method anyone can come to conclusions regarding the verb class frequencies. On the other hand, the paper highlights that statistics and data made by programs always have to be supervised by humans. The future of linguistics is based on the interdisciplinary approach to language investigation thus researchers have to accept the challenges and incorporate computational methods and tools into their field of interest. However, we should percieve computational methods as tools which help us analyze language and not as an approach that substitutes linguists.

References

Babić, S., Brozović, D., Moguš, M., Pavešić, S., Škarić, I., Težak, S. (1991). Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika: Nacrti za gramatiku. Zagreb: HAZU: Globus

Bošnjak Botica, T. (2013). Opća načela podjela na glagolske vrste u hrvatskome u perspektivi drugih bliskih jezika. // Lahor 1, 15, 63-90

Fletcher, W. H. (2012). Corpus analysis of the world wide web. // The encyclopedia of applied linguistics / Chapelle, C. A. (ed.). Hoboken, NJ: John Wiley & Sons

Jelaska, Z. (2003). Proizvodnja glagolskih oblika hrvatskoga jezika kao stranoga jezika: od infinitiva prema prezentu. // Zbornik Zagrebačke slavističke škole 2002. / Botica, S. (ed.). Zagreb: FFpress Filozofski fakultet, 48-63

Jelaska, Z., Bošnjak Botica, T. (2019). Conjugational Types in Croatian. // Rasprave: časopis instituta za hrvatski jezik i jezikoslovlje 45, 1, 47-74

Ljubešić, N., Klubička, F. (2014). {bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian. // Proceedings of the 9th Web as Corpus Workshop (WaC-9). Bildhauer, F., Schäfer , R. (eds.). Gothenburg: Association for Computational Linguistics, 29-35

Marković, I. (2012). Uvod u jezičnu morfologiju. Zagreb: Disput

Raguž, D. (1997). Praktična hrvatska gramatika. Zagreb: Medicinska naklada

Silić, J., Pranjković, I. (2005). Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta. Zagreb: Školska knjiga

Tadić, M. (1997). Računalna obrada hrvatskih korpusa: povijest, stanje i perspektive. // Suvremena lingvistika 23, 43-44, 387-394

https://doi.org/10.17234/INFUTURE.2019.7

Original scientific paper

Can Societal Impact of Scientific Work Be Measured in the Process of Re-Accreditation of Higher Education Institutions and Public Scientific Institutes in Croatia?

Marina Grubišić
Agency for Science and Higher Education, Zagreb, Croatia
marina.grubisic@azvo.hr

Sonja Špiranec
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
sspiran@ffzg.hr

Summary
In this paper we present an approach to measuring the societal impact of scientific research. Our proposed methodology is based on the theory of productive interaction (Spaapen and van Drooge, 2011). The methodology was tested on the reports produced in the process of the expert evaluation of public higher education institutions (HEIs) at universities and public research institutes in Croatia. For the purpose of quantitative assessment, we have developed a conceptual framework and analyzed the narrative texts of reports based on recognising codified interactions. Finally, we discuss our results in the context of the research area of the Republic of Croatia in the fields of social sciences and biomedicine.

Key words: societal impact of scientific work, Croatian system of higher education and science

Introduction

In this paper, we review several approaches to measuring a societal impact of scientific work and propose a possible methodology for its measurement. Our proposed methodology will be assessed on a case study of the evaluation process of public higher education institutions (HEIs) at universities and public research institutes in Croatia. We will develop a conceptual framework following (Spaapen and van Drooge, 2011) which proposes evaluation based on productive interaction theory.

Measuring the societal impact

Impact assessment in the context of science is complex and challenging. Common approaches, such as citation counts, are often critiqued, while it is emphasized that traditional bibliometric indicators (Holmberg et al., 2015) do not measure impact of science on the wider community. Along with the change in funding allocation (Hicks, 2012) based on indicators, we have a strong public demand that the scientific activity should not be closed within the scientific community. Today's scientific activity is in the transition from a relatively closed system within scientific areas and fields to an open and interdisciplinary structure where knowledge creation is increasingly available to stakeholders throughout the society (Wilsdon et al., 2016). A new way of financing and a new approach to scientific, activity-emphasizing accountability to the wider community has motivated an increasing number of researchers to explore the impact of scientific work on the society at large (Benneworth, Pinheiro and Sánchez-Barrioluengo, 2016). Social impact, obviously, is an elusive concept and hard to grasp. One possible definition is that scientific work has a social impact when there is a reference to it outside the scientific community (Bornmann, Marx, 2014). Although intuitive, this definition does not imply how to demonstrate and measure, rather than presume social impact, and currently different approaches are being considered This study will contribute to the exploration of the complex concept of social impact of scientific work and possible approaches of measuring it by relying on narrative data gained in the evaluation process of HEI's in Croatia. The research is based on the framework of productive interactions, which understands productive interactions as exchanges between researchers and stakeholders in which knowledge is produced and valued that is both scientifically robust and socially relevant (Spaapen, van Drooge, 2011).

Methodological approach and research questions

This study is based on case study data derived from experts' reports produced in the process of reaccreditation of Croatian HEIs and public research institutes. The evaluation procedure included written reports and recommendations made by expert panel indicating societal impact of scientific work. These reports are further analysed as textual material and used as a benchmark to assess the objectiveness of measurement, and thus reduce a possible bias in the results.

The study was framed by the conceptual challenge of understanding social impact in scientific work; within this broader context, research question that drive this study are:

1. Can societal impact of scientific work be measured according to the framework of productive interaction in the expert reports produced in the process of reaccreditation?
2. Are different types of institutions (faculties of public universities and public scientific institutes) putting emphasis on different types of social impact?
3. Are different scientific fields (biomedicine and social sciences) highlighting different examples of social impact?

The analysis will be was conducted by using a qualitative data processing tool (QDA Miner Lite). A document-category matrix was used in order to detect and identify instances of productive interaction in text reports. Following types of interactions where considered direct interaction (DI), indirect interaction (II) or financial interaction (FI).

The key terms for *direct interaction* are; participation in professional bodies and conferences, meetings with stakeholders, membership in management bodies and collaboration with public services.

The key terms for *indirect interaction* are; professional papers, presence in the media, social media presence and reporting to local governance.

The key terms for *financial interaction* are commercial and professional contracts and financing of students and teachers.

Through this qualitative analysis, we will gain an overview of the conceptual framework recognized by the expert committees as the social impact of scientific work at faculties, which are constituents of public universities, and public scientific institutes in the social and biomedical scientific field.

Case study of Croatia

In the Republic of Croatia, the social impact of scientific work has not been evaluated so far in the above-mentioned categories.

The process of re-accreditation of higher education institutions includes five phases: *self-assessment of higher education institutions*, *visits of the expert committee to the higher education institution*, *preparation of the final report of the expert committee*, *adoption of the Accreditation Recommendation* and *subsequent follow-up*. Each expert committee report must contain an analysis based on evidence gathered through the self-evaluation document prepared by the institution and evidence gathered during a site visit. It is equally important for the Croatian model of external evaluation that each report must have recommendations for improvement for each evaluation criterion. In addition to the above-mentioned analyses and recommendations, the expert committee also provides ratings for each evaluation criterion. In this study, final reports of expert committees for higher education in the scientific field of biomedicine and for higher education in the scientific field of social sciences were analyzed.

The research includes faculties of public universities since higher schools and polytechnics in the Republic of Croatia according to the Law on Scientific Activity and Higher Education are not obliged to carry out scientific activity or do not have to have a scientific accreditation (ASHE, 2009). In the scientific field of biomedicine, reports were analyzed for seven faculties that are part of public universities. In the scientific field of social sciences, reports were analyzed for faculties that are part of public universities that have a scientific accreditation for this area, apart from the faculties of economics. This field is not analyzed as part of the social science area because it is separately evaluated in the process of re-accreditation precisely because of its specificity. The scientific field of economics' societal impact is similar to the economical component of social relevance in the scientific field of technical science. In the social sciences, the reports for eighteen faculties that are part of public universities are analyzed.

The evaluation was carried out in accordance with the Criteria for assessment of quality of higher education institutions within universities (ASHE, 2013a). which were used in the 2010-2016 period.

Evaluation of public scientific institutes was carried out in accordance with the Principles and criteria for evaluation of scientific organizations in the Republic of Croatia (ASHE, 2013b). The analysis was carried out in the social science field and in the field of biomedicine sciences, so that the results are comparable with the results for higher education institutions within the public university system. Seven reports in total were analysed for four public scientific institutes in the field of social sciences and three public institutes in the field of biomedicine sciences.

Criteria from the self-evaluation report that assess the social impact of the scientific work are analyzed according to the adopted framework.

Results

All reports were analyzed according to the principle of productive interaction. A text is marked according to the number of instances when a coded interaction (grouped further in three categories) is detected in the body of text.

In the software we used, key terms are shown as categories encoded in the report texts (which are cases in this terminology). The table shows the number and percentage of codes, that is, how many times a code has appeared and what fraction (as a percentage) does it constitute of all recognized codes. The second column shows the number and the relative frequency (as a percentage) of cases, measuring in how many reports did the category appear out of the total number of reports analyzed. In both cases we restricted the number of reports by the scientific filed and the type of institution.

Table 1. Number and percentage of codes and cases - public scientific institutes in social sciences

TYPE	CATEGORIES	CODES		CASES	
DI	Professional conferences	0	0	0	0
DI	Professional bodies	3	14.30%	3	75.00%
DI	Management bodies	1	4.80%	1	25.00%
DI	Meeting stakeholders	4	19.00%	4	100.00%
DI	Collaboration with public services	4	19.00%	4	100.00%
II	Professional publications	0	0	0	0
II	Media presence	2	9.50%	2	50.00%
II	Social media presence	1	4.80%	1	25.00%
II	Reporting to local governance	2	9.50%	2	50.00%
FI	Commercial contracts	1	4.80%	1	25.00%
FI	Professional contracts	3	14.30%	3	75.00%
FI	Financing of students	0	0	0	0
FI	Financing of teachers	0	0	0	0

Table 1. shows the number and percentage of detected coded interactions and cases in the fields of social sciences.

The majority of codes for direct interaction refers to participation in meeting stakeholders and collaboration with public services in 100 % of cases, the majority of codes for indirect interaction are reporting to local governance and media presence of institution participation in 50 % of cases. The dominant codes for financial interaction are professional contracts in 75% of cases.

Table 2. Number and percentage of codes and cases- faculties in public universities in social sciences

TYPE	CATEGORIES	CODES		CASES	
DI	Professional conferences	2	2.70%	2	11,10%
DI	Professional bodies	13	17.80%	12	66.70%
DI	Management bodies	4	5.50%	4	22.20%
DI	Meeting stakeholders	11	15.10%	10	55.60%
DI	Collaboration with public services	3	4.10%	3	16.70%
II	Professional publications	2	2.70%	2	11.10%
II	Media presence	4	5.50%	4	22.20%
II	Social media presence	1	1.40%	1	5.60%
II	Reporting to local governance	12	16.40%	12	66.70%
FI	Commercial contracts	4	5.50%	4	22.20%
FI	Professional contracts	16	21.90%	16	88.90%
FI	Financing of students	0	0	0	0
FI	Financing of teachers	1	1.40%	1	5.60%

Table 2. presents results for the group of faculties of public universities in the scientific fields of social sciences.

The majority of codes for direct interaction are participation in professional bodies in 66% and meeting stakeholders in 55% of cases, the majority of codes for indirect interaction are reporting to local governance in 66% and media presence of institution in 16% of cases. The dominant codes for financial interaction are professional contracts in 88% of cases.

Table 3. Number and percentage of codes and cases - public scientific institutes in biomedicine

TYPE	CATEGORIES	CODES		CASES	
DI	Professional conferences	0	0	0	0
DI	Professional bodies	3	14.30%	3	100.00%
DI	Management bodies	0	0	0	0
DI	Meeting stakeholders	3	14.30%	3	100.00%
DI	Collaboration with public services	3	14.30%	3	100.00%
II	Professional publications	2	9.50%	2	66.70%
II	Media presence	2	9.50%	2	66.70%
II	Social media presence	1	4.80%	1	33.30%
II	Reporting to local governance	4	19.00%	3	100.00%
FI	Commercial contracts	2	9.50%	2	66.70%
FI	Professional contracts	1	4.80%	1	33.30%
FI	Financing of students	0	0	0	0
FI	Financing of teachers	0	0	0	0

Table 3 shows results for the group of public scientific institutes in the scientific fields of biomedicine.

All the codes for direct interaction are participation in professional bodies, meeting stakeholders and collaboration with public services in 100 % of cases. The majority of codes for indirect interaction are reporting to local governance in 100% of cases, media presence, and professional publications in 66 % of cases. The dominant codes for financial interaction are commercial contracts in 66% of cases.

Table 4. Number and percentage of codes and cases - faculties in public universities in biomedicine

TYPE	**CATEGORIES**	CODES		CASES	
DI	Professional conferences	1	3.30%	1	16.70%
DI	Professional bodies	5	16.70%	5	83.30%
DI	Management bodies	1	3.30%	1	16.70%
DI	Meeting stakeholders	5	16.70%	5	83.30%
DI	Collaboration with public services	3	10.00%	3	50.00%
II	Professional publications	3	10.00%	3	50.00%
II	Media presence	0	0	0	0
II	Social media presence	0	0	0	0
II	Reporting to local governance	0	0	0	0
FI	Commercial contracts	5	16.70%	5	83.30%
FI	Professional contracts	5	16.70%	5	83.30%
FI	Financing of students	0	0	0	0
FI	Financing of teachers	2	6.70%	2	33.30%

Table 4 indicates for the group of faculties of public universities in the scientific fields of biomedicine.

The majority of codes for direct interaction are participation in professional bodies and meeting stakeholders in 83% of cases, all the codes for indirect interaction refer to professional publications in 50%. The dominant codes for financial interaction are commercial and professional contracts in 83% of cases.

Conclusion

Conceptions of impact in science are nowadays reconsidered and broadened in order to reflect the influence of scientific work on society. Societal impact is an elusive concept, which is very hard to measure, and different approaches are being examined to deal with this challenge. The aim of the study was to determine the volume of scientific results recognized by the expert committees as having societal impact for the faculties of public universities and public scientific institutes in the social and biomedical scientific field, based on the conceptual framework of productive interaction. Answer to our first question is that societal impact of scientific work can be measured according to the framework of productive interaction in the expert reports produced in the process of reaccreditation.

The results of the study indicate differences between public institutes and faculties in field of social sciences. The majority of codes for direct and indirect interaction are recognized in public institutes in the area of social sciences. The percentage of codes on faculties is smaller. Both public institutes and faculties in social sciences have recognized codes in financial interaction, especially in professional contracts.

Similar to results in social sciences, data for public institutes and faculties in biomedicine are different. More codes for direct and indirect interaction are recognized in public institutes for biomedicine. Both public institutes and faculties in biomedicine have recognized codes in financial interaction, especially in professional contracts.

In public institutes direct and indirect interaction is more recognized than on faculties. On faculties professional bodies and professional contracts were recognized as important.

In addition to differences by type of institution (institute vs. faculty), results suggest differences in perceptions on what is recognized as social impact in different scientific fields (social sciences vs. biomedicine).

The majority of codes for direct interaction are recognized in the field of biomedicine. The percentage of codes in the field of social sciences is smaller but still relevant, especially in the case of public scientific institutes. Both fields have recognized codes in indirect interaction, but different elements were recognized as important (professional publications for biomedicine and reporting to local governance for social sciences).

Finally, both fields have recognized codes in financial interaction, and the same elements (commercial and professional contracts) were recognized as important.

The findings of the presented study highlight the potential usefulness of the concept of productive interaction as a framework for analyzing social relevance of scientific activity from narrative data, and lay groundwork for further research of differences in evaluating societal impact between different scientific fields and type of institutions.

References

ASHE. (2009). (Agency for Science and Higher Education, Act_on_Scientific_Activity). https://www.azvo.hr/images/stories/o_nama/Act_on_Scientific_Activity.pdf

ASHE. (2013a). (Agency for Science and Higher Education,Criteria for the assessment of quality of higher education institutions within universities). https://www.azvo.hr/en/evaluations/evaluations-in-higher-education/re-accreditation-of-higher-education-institutions-2010-2016

ASHE. (2013b). (Agency for Science and Higher Education, Principles and criteria for the evaluation of scientific organisations in the republic of Croatia). https://www.azvo.hr/en/evaluations/evaluations-in-science/re-accreditation-of-scientific-organisations/re-accreditation-of-public-research-institutes

Benneworth, P., Pinheiro, R., Sánchez-Barrioluengo, M. (2016). One size does not fit all! New perspectives on the university in the social knowledge economy. // Science and Public Policy 43, 6, 731-735. doi: 10.1093/scipol/scw018

Bornmann, L., Liakata, M., Clare, A., Duma, D.. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements // PLoS ONE 12, 3, 1-18. doi: 10.1371/journal.pone.0173152

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? a literature survey. // Journal of the American Society for Information Science and Technology 64, 2, 217-233. doi: 10.1002/asi.22803

Bornmann, L., Marx, W. (2014). How should the societal impact of research be generated and measured? a proposal for a simple and practicable approach to allow interdisciplinary comparisons. // Scientometrics 98, 1, 211-219. doi: 10.1007/s11192-013-1020-x

Hicks, D. (2012). Performance-based university research funding systems. // Research Policy 41, 2, 251-261. http://doi.org/10.1016/j.respol.2011.09.007

Holmberg, K., Didegah, F., Bowman, T., Kortelainet, T. (2015). Measuring the societal impact of open science - Presentation of a research project. // Informaatiotutkimus, 34, 1-4, 119-123. http://search.ebscohost.com/login.aspx?direct=true&db=llf&AN=114559577&site=ehost-live.

De Jong, S. P. L., van Arensbengen, P., Daemen, F., van der Meulen, B., van den Besselaar, P. (2011). Evaluation of research in context: An approach and two cases // Research Evaluation 20, 1, 61-72. doi: 10.3152/095820211X12941371876346

De Jong, S. P. L., Smit, J., Van Drooge, L. (2016). Scientists' response to societal impact policies: A policy paradox. // Science and Public Policy 43, 1, 102-114. doi: 10.1093/scipol/scv023

Samuel, G. N., Derrick, G. E. (2015) Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014 24, April, 229-241. doi: 10.1093/reseval/rvv007

SIAMPI et al. (n. d.). SIAMPI final report Executive summary, 1-36. doi: 10.1017/CBO9781107415324.004

Spaapen, J., Dijstelbloem, H., Wamelink, F. (2007). Evaluating research in context, A method for comprehensive assessment, 2nd edition, The Hague: COS. http://www.eric-project.nl/files.nsf/pages/NWOA_73VH8D/$file/eric_book_internet.pdf.

Spaapen, J., van Drooge, L. (2011). Introducing "productive interactions" in social impact assessment. // Research Evaluation 20, 3, 211-218. doi: 10.3152/095820211X12941371876742

Wilsdon, J., Bar-Ilan, J., Frodeman, R., Lex, E., Peters, I., Wouters, P. (2016). Next-Generation Metrics: Responsible Metrics & Evaluation for Open Science, STI 2016: Peripheries, frontiers and beyond, (September), 1-4. doi: 10.2777/337729

Wilsdon, J. (2017). Responsible Metrics // Higher Education Strategy and Planning: A Professional Guide, (July), 247-254. doi: 10.13140/RG.2.1.4929.1363

https://doi.org/10.17234/INFUTURE.2019.8

Preliminary communication

Intergenerational Knowledge Sharing in Business Settings
A Comparative Case Study between Germany and South-Korea

Lea Wöbbekind
University of Hildesheim, Germany
woebbek@uni-hildesheim.de

Christa Womser-Hacker
University of Hildesheim, Germany
womser@uni-hildesheim.de

Dowan Kim
Pai Chai University, South Korea
dwkim@pcu.ac.kr

Summary

Researchers view knowledge as an essential organizational asset for corporate success. Over the past decades, many studies explored the management of knowledge in multinational companies. Nowadays in the context of demographic change, intergenerational aspects and age have become crucial factors for knowledge sharing. However, it remains unclear how cultural attitudes affect knowledge sharing within companies. The purpose of this paper is to give new insights into the impact of culture and age on knowledge sharing in German and South-Korean companies. To understand how culture influences an individuals' perception of age and generation in the context of knowledge sharing in an organization, a qualitative approach including semi-structured interviews is chosen.

Key words: knowledge management, intergenerational knowledge sharing, demographic change, cultural aspects

Introduction

Nowadays, many companies face relevant changes within their economic environment and regarding their future. At the organizational level, knowledge becomes an important asset. Moreover, knowledge sharing is seen as a strategic advantage of companies for success and competition (Nonaka, Takeuchi, 1995; Hwang et al., 2015). Workforces are becoming older and more age-diverse. It is necessary for organizations to avoid the loss of knowledge through the retirement of older employees. However, this is not an easy task due to the intangible character of knowledge. Individual and organizational knowledge differ in several characteristics as well as different distribution mechanisms. As several studies point out, culture, embodied in language, rules or concepts, affects the knowledge sharing of individuals and groups (DeLong, Fahey, 2000; Ardichvili et al., 2006). Thomas and Utler (2013: 41) define culture as a universal orientation system for societies or groups. This system influences the perception, thinking and norms and defines an individual's membership of the society. Moreover, many studies in the research area of knowledge management and knowledge sharing focus on the effects of cultural diversity in multinational companies on communication and knowledge transfer (Wikström et al., 2018).

Purpose

An overall aim of this article is to explore cultural issues on knowledge sharing in national companies. Besides individual, social and organizational approaches on intergenerational knowledge sharing (Wíden, 2017) more research needs to be done to explore the impact of culture on intergenerational knowledge sharing. The basis for this study is the KNOWISH project (Knowledge sharing between age groups) and previous master theses from Brinken and Kock (2017) as well as Ivantsova and Sivén (2016). Both theses explore knowledge sharing in German and Finnish companies. Semi-structured interviews with employees of different age groups were conducted. Using this knowledge this case study aims to:

- Provide an in-depth understanding of what generational effects and cultural issues appear in a South-Korean research institute.
- Compose this knowledge and compare the results with the two German case studies to understand the impact of culture on intergenerational knowledge sharing in different companies.

Generation model in South Korea

Kuyken (2012) portraits the generation model of baby boomers, generation X, and generation Y in the context of knowledge sharing in an organization (Zemke et al., 2000; DeLong, 2004). For the new case study, it is important to adapt the generation model to South-Korean culture.

Park and Park (2018) use the mentioned model above and add South-Korean characteristics for each particular generation and also focus on the effects on workplaces. They present a unique generation, called generation 386, which can be only found in South-Korea. Members of generation 386 are born between 1960 to 1969. The term was introduced in the 1990s to describe people in their 30s who attended universities in the 80s. The group of generation 386 witnessed a military dictatorship and therefore stands up for democratization in Korea. Moreover, they experienced the economic recession and the Asian financial crisis that continued in 1997 (Park, 2007). They appreciate traditional values and collectivism. Nowadays, individuals of this generation work as heads of departments or directors, generally higher positions in society. Consequently, this generation is responsible for the education of future leaders. It is important to understand the generation 386, as they have a large responsibility in the South-Korean society (Park, Park, 2018). Individuals born between 1970 and 1980 are called Generation X or *Shinsedae* in South-Korea and refers to the term *new generation*. Individuals of this group were raised in a better economic situation than the previous generation and experienced more freedom and individualism. They agree that work and career belong to themselves rather than their employers. Moreover, members of this cohort are supposed to be treated individually and focus on work-life balance (Park and Park, 2018). The third generation is called *Millennials* and refers to generation Y. It comprises of individuals born between 1980 and 1994. This generation was raised among information technologies and is, therefore, more skilled in the use of digital technologies, media and instant communication (Park, Park, 2018). In contrast to generation 386, this group is supposed to be freer from political or ideological tendencies (Park, 2007). Members of generation Y value personal life more than individuals from older generations. Another interesting generational aspect occurs in the context of over-parenting. It signifies that members of generation Y are used to receive solutions and caring from their parents. They may fail to develop their own identity and adequate independence until reaching adulthood. As a consequence, the hierarchical structures and values that older generations have in South- Korea are not likely to promote the strengths of generation Y at work (Park, Park, 2018). Because of rapid social transformation during only a few decades, South-Korean society seems to experience wider gaps between older and younger generations.

Several studies explore the organizational impact of added values for mixed-age teams. Bratianu and Leon (2015) emphasize the positive effects regarding the sharing of deep knowledge of more experienced employees, including know-how, working morale and quality, while younger employees include broad knowledge, regarding a high ability to learn. In theory, it encourages mutual learning and stimulates knowledge increasing. For our study, based on the year of birth, we recruited two participants from generation 386, four participants from generation X and five employees from generation Y in the South-Korean research institute.

Research method

In order to answer the research questions, an interdisciplinary approach from an information science perspective is chosen. We employed an empirical qualitative research model, as it allows for exploring unknown phenomena (Rosenthal, 2008: 18). In contrast to quantitative methods, this enables an in-depth examination of attitudes and values, as this study is not intended to test hypotheses nor statistical analysis (Raithel, 2006: 8). Eleven interviews were conducted with South-Korean employees in a business setting. The interview guideline from Brinken and Kock (2017: 175 ff.) is re-used. It includes 30 questions in 5 categories about communication, knowledge and learning aspects, organizational learning, systems, and virtual communities. The selection of young and old interviewees was not random, as the age of interviewees is between 26 and 54. The qualitative approach employed interviews from 30 to 60 minutes. Table 1 shows the demographic data of participants. In general, semi-structured interviews contain open-ended questions and were significantly useful for our approach. Other questions that arise during the interview can be addressed and discussed. Individual in-depth interviews can dip deep into the past of the interviewee and examine social and personal matters.

Table 1. Demographic data of the South-Korean participants

Number	Gender	Job description	Time in the company	Year of birth
1	m	Senior researcher	5 years	1975
2	m	Senior researcher	4 years	1976
3	m	Ph.D. student	4 years	1992
4	m	Ph.D. student	3 years	1986
5	m	Senior researcher	2 years	1977
6	m	Senior researcher	12 years	1981
7	m	Senior researcher	4 years	1983
8	m	Professor, Head of Lab	8 years	1972
9	f	Ph.D. student	1.5 years	1980
10	m	Team manager	22 years	1965
11	m	Department leader	28 years	1965

The German case studies (Brinken, Kock, 2017) include one middle-sized and one small business company: An Optician store and software development company. The interviewees (m=7, f=3) from the first company include one trainee (born 1994), three opticians (born 1992, 1994 and 1992), three shop managers (born 1955, 1976 and 1953), one optometrist (born 1984) and two employees from the upper management (born 1964 and 1952). The demographic data revealed a particularly broad age range from 22 to 64 years. The software and IT company is relatively young with only a few older employees: Three software developer (born 1981, 1986 and 1974), one student employee (born 1995), one team leader (born 1982), two project managers (born 1965 and 1980), two consultants (born 1969 and 1986) and one executive board member (born 1952). There were nine males and one female interview participants. The qualitative data analysis involved several research steps. To get an overview of the rich data, several reading sessions were conducted to explore initial ideas and interesting, related empirical data characteristics. Potential patterns were gathered and categorized for interpretation. After the data analysis follows the interpretation of surveyed data. Passages are formed into paragraphs of the interviews that are grouped based on their similarity (Mayring, 2015: 18).

Challenges

A literature review shows that more empirical research needs to be done to explore the advantages and disadvantages of conducting qualitative research, like interviews, in South-Korea. Park and Lunt (2015) conducted a qualitative study with Korean participants in the public sector to examine practical issues between Anglophone techniques and countries with a Confucian background. The authors highlight different issues between Anglophone research techniques, like interviews, with South-Korean participants. The main difficulties occurred with problems of the comprehensibility of the research question and trust issues between the interviewees and interviewer. Participants expressed difficulties in answering open questions and expressing their individual opinion. Attention must be paid to the impact of indirectness. Indirectness is a communication skill that expresses politeness or self-protection and is typically used in Asian countries. It is also utilized to avoid embarrassment or misunderstandings. The main impact on the interview may be that participants use proper words or behaviors instead of expressing honest statements or opinions (Zhang, You, 2009). As the Confucian background approves hierarchy and social mores in South-Korea, the researchers were worried that social networks of the participant’s influence on the recruitment process (Park, Lunt, 2015). This may result in a too homogeneous group of interviewees. These insights were consulted to understand the limitations of Western research methods in an Asian setting. The framework was taken into account to deal with arising problems during the interviews, as the interviewer had a non-Korean background.

Results

To present the findings a qualitative approach with semi-structured interviews is conducted with South-Korean employees to explore the impact of culture on intergenerational knowledge sharing. Quotes from the interviews will be shown to explain the findings. The data revealed insights into age structures, influence of status and (work) experience as well as competencies of each cohort in the organization. The most interesting outcome shows that sharing know-how and expertise between different age groups is not an important issue in the context of knowledge management and demographic changes. In greater detail, maintaining an orderly and harmonious society, based on Confucian philosophy, is highly relevant at the workplace and forms the relationship of generations in the South-Korean institute: *Yes, in that culture because younger people are just educated to obey to older people first. So not asking something. […] compared to US or Deutschland, in classroom we usually do not ask something. Just listen the teacher's message and classes and memorize something. That is our culture. That is an obstacle in knowledge sharing. [...]* (P5, Generation X).

The appreciation of seniority is integrated into a three-stage age system: The classification of young, old and elder employees in the organization. This hierarchical system is based on chronological and professional age, status and position in the institute and shows the impact of hierarchy on work life. The study also revealed interesting insights into the characteristics of each generation, which support the theoretical framework of generations by Kuyken (2012) and Park and Park (2018). Younger employees with less status and experience are excluded from decision-making processes in departments, with the following explanation: *Because older researchers have the more experience. Usually, in realistic case, the solution is focused on the experience* (P7, Generation X).

Younger employees demonstrate advanced competencies in technology and self-studying to gain knowledge. Regarding knowledge sharing, they emphasize learning especially from older and elder employees. Older employees are characterized as knowledge carriers, especially for knowledge about the company. Their professional age signifies high social competence. Based on the competencies and inabilities of young and older employees, a teacher-student relationship between these generations at the workplace could be revealed. The third identified group in the organization are elder employees. From the age of 50 up, they have leading positions in the departments. They, as several participants explained, have a higher status in society than younger employees that include qualified communication skills and experience. In comparison to generation Y, they lack openness for innovation and new technologies. As teams are organized with employees of different generations, the influence of hierarchy is shown by more responsibilities of a member of generation X and generation 386. Remarkably, younger employees mention an important communication style when working or communicating with elder employees: Showing respect, manners and the use of polite language. Moreover, respect is shown in general agreement with work decisions that might lead to communication and generation issues: *Korean society community they have a hierarchy structure. Many younger people bad situation. Sometime older people supervise provide. So, we can feel difficult communication with the supervisor* (P11, Generation 386).

Therefore, the data analysis revealed a business structure of top-down communication as well as a hierarchical structure in the organization that is identified by members of different generations. Apart from generational aspects, knowledge retention is not related to the success and competition of the company. Besides the findings of generational factors and barriers for intergenerational knowledge sharing, knowledge hoarding might be an important issue. As already mentioned, each generation acquired specific expertise and skills in different work areas. Interestingly, the analysis revealed that sharing expertise is not correlated with benefits and advantages for the management and individual employees. Reasons for this outcome include the lack of good quality of knowledge, human factors and trust: *To get the other people´s experience, some knowledge, it is very difficult. We are not a machine, we are human. It is difficult to communicate with unknown people* (P1, Generation X). A younger employee describes communication issues with elder colleagues in greater detail: *So that is the reason why I told you the communication is usually top-down. But that is very, how can I say… it is very… it seems like to slow. They are very far, it is very hard to communicate with them directly* (P3, Generation Y).

The interviewee´s statements support the result that no official channels or guidelines for sharing knowledge are foreseen. Therefore, no intrinsic or extrinsic motivational factors for knowledge sharing in the organization could be revealed.

The impact of culture and intergenerational aspects in the German case studies (Brinken, Kock, 2017) lay on young employees being better in handling technologies. They are open-minded towards new trends or trying new things out. This relates to the flexibility and motivation of young employees.

The competencies of older employees lay in their work experience and proficiency in communication skills and less in learning and using new systems. The competencies both of young and old employees demonstrate that knowledge sharing happens mutually, based on skills. Both groups can learn from each other and are willing to share their best practices or expertise. More similarities can be found by an open atmosphere in both companies, the prioritization of teamwork and a flat hierarchy structure.

Comparison of knowledge sharing and generational aspects

The analysis of generational aspects in Germany and South-Korea revealed two different concepts of age: The German interviewees describe a non-hierarchical system of young and old employees. Both groups are supposed to be knowledge carriers. Therefore, knowledge sharing happens mutually between them at the workplace. Generally, German employees show an open-minded, critical thinking and rational attitude towards knowledge sharing.

According to the conducted interviews, knowledge sharing happens mutually in the German case studies (Brinken, Kock, 2017). Especially tacit knowledge is identified as an important asset for sharing with colleagues. Helping each other at work is natural and happens on a daily basis. Moreover, the atmosphere has a very positive impact on communication and knowledge sharing. As already mentioned, the German interviewees describe a knowledge sharing-friendly environment characterized by a familiar atmosphere, team spirit and strong relationships between all employees. In both German companies, this atmosphere is created by the appreciation of open communication as a corporate value and community-building events. Team spirit is also reflected in the way employees learn within the organization. The German interviewees mention that they learn by reflecting past situations together with colleagues, by exchanging experiences in training and meetings and by observing colleagues. This behavior demonstrates flat hierarchical structures in the German companies and openness for direct communication and discussions of experienced and inexperienced employees. Nevertheless, this structure is not necessarily associated with culture. The generation model, introduced by Kuyken (2012) and Park and Park (2018), is characterized by shared values and characteristics of individuals based on its socialization. However, these results differ greatly from those in the South-Korean organization. Table 2 shows the outcomes of both studies in greater detail.

Table 2. Comparison of the results

Feature	South-Korea	Germany
Age structure	3-stage age model	2-stage age model
Hierarchy	Particularly pronounced in the company	Non-existent in the companies
Communication style	Formal and respectful language style, top-down communication	Direct and open, feedback culture
Working environment	Self-studying	Teamwork, helping each other
Individual motivation	No intrinsic motivation could be identified	Intrinsic motivation (responsibility and joy)
Experience	Age and status in the company form experience	Professional age (work experience)
Technologies	Use of virtual communities	Intranet, social networks
Understanding of generations	Teacher-student relationship	Equal relationship, knowledge carriers
Support by the management	No official support	To some degree
Knowledge loss	Knowledge hoarding, required knowledge is gained by experience	Less knowledge hoarding

Social and cultural differences between younger and elder employees (for example shown in communication styles and the general understanding of generations) explain the mentioned generation gap in the South-Korean workplace. For example, old and elder employees are responsible for decision-making. Members with a lower chronological age are excluded from management decisions. To balance this gap, younger employees are very well in self-learning and studying. Therefore, in both cases, a broad range of expertise and know-how can be identified.

The impact of culture on intergenerational knowledge and knowledge sharing

Teams of mixed ages provide diverse knowledge and expertise. Transferring knowledge in intergenerational teams or departments can overcome the knowledge gaps of individuals. In the South-Korean company, older and elder employees are less innovative and open to changes and lack the motivation to learn new skills. This knowledge gap is compensated by the advanced skills of younger colleagues regarding information technology. A broad range of ages shows a high diversity of skills, capabilities, and strengths for each cohort. Each generation contains knowledge carriers and expertise. Especially South-Korean culture revealed its influence on intergenerational knowledge sharing. Its impact is shown in the behaviors and values of individuals. In detail, the influence of cultural aspects is demonstrated by hierarchy and high appreciation for seniority.

Many barriers to knowledge sharing could be revealed in both studies. Reasons for hoarding knowledge include trust issues, limited support by the management, hierarchy structures and lack of good quality knowledge. Nonetheless, knowledge sharing is mainly a process in one direction only within the South-Korean organization and is characterized by hierarchy and a top-down communication style. Whenever expertise is shared, it is often limited to a small team or single colleagues and does not affect the entire organization or a larger team. Interestingly, teamwork is considered to be less important in the organization. Based on the theoretical background, Confucian philosophy has an important role. It shapes the relationship between young, old and elder employees. Intergenerational knowledge sharing is negatively influenced by formal language styles. Younger and also older generations behave politely and considerately towards individuals with a higher status and elderly. South-Korea is an example of a country with pronounced collectivism and strong hierarchies. Additionally, there is no official support or system to share or gain knowledge in the company. The phenomena of knowledge barriers revealed that knowledge hoarding, as well as knowledge loss as a result of the retirement of elder employees, is a possible scenario.

Conclusion

The companies differ largely because of domains, working environment, age structures and work routine. Nevertheless, interesting insights into the impact of culture on intergenerational knowledge is illustrated. The presence of some aspects of social norms, values, and behaviors leads to the assumption they might be important to intergenerational knowledge sharing. Based on these concepts, intergenerational knowledge sharing is not actively pursued or encouraged within the South- Korean company. On the contrary, the German case studies show how working environments and intrinsic motivation of individuals favorably affect knowledge sharing between employees of different age groups. These differences in the results show that culture in Western and Asian countries shapes assumptions about what knowledge is important and how it will be used and shared within an organization or department.

Several individual and organizational reasons for knowledge hoarding and knowledge loss could be revealed. Age structures, like a three-stage age model in South Korea, demonstrate the influence of culture.

References

Ardichvili, A., Marer, W., Wei, L., Wentling, T., Stuedemann, R. (2006). Cultural influences on knowledge sharing through online communities of practice. // Journal of Knowledge Management 10, 4, 541-554

Bratianu, C., Leon, R. D. (2015). Strategies to enhance intergenerational learning and reducing knowledge loss. // VINE 45, 4, 551-567

Brinken, H., Kock. H. (2017). Exploring intergenerational Knowledge sharing in Organizations. https://edoc.hu-berlin.de/handle/18452/20725 (2019/08/01)

DeLong, D., Fahey, L. (2000). Diagnosing cultural barriers to knowledge management. // Academy of Management Executive 14, 4, 113-127

DeLong, D. (2014). Lost knowledge. Confronting the threat of an aging workforce. Oxford: University Press

Hwang, E. H., Singh, P. V., Argote, L. (2015). Knowledge sharing in online communities: Learning to cross geographic and hierarchical boundaries. // Institute for Operations Research and the Management Sciences 1, 2, 1593-1611

Ivantsova, E., Sivén, T. (2016). Intergenerational knowledge sharing within two case studies. http://www.doria.fi/bitstream/handle/10024/131066/Intergenerational%20knowledge%20sharing%20within%20two%20case%20studies.%202017.pdf?sequence=1&isAllowed=y (2019/08/02)

Kuyken, K. (2012). Knowledge communities: towards a re-thinking of intergenerational knowledge transfer. // Journal of information and knowledge management systems 42, ¾, 366-381

Mayring, P. (2015). Qualitative Inhaltsanalyse. Grundlagen und Techniken. Weinheim/Basel: Beltz Verlag

Nonaka, I., Takeuchi, H. (1995). The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. New York: Oxford University Press

Park, S., Lunt, N. (2015). Confucianism and qualitative interviewing. Working Seoul to soul. // Forum Qualitative Sozialforschung 16, 2

Park, S., Park, S. (2018). Exploring the generation gap in the workplace in South Korea. // Human Resource Development International 21, 3, 276-283

Park, S. Y. (2007). Shinsedae: Conservative attitudes of a new generation in South Korea and the impact on the Korean presidential election. // EWC Insights 2, 1, 1-4

Raithel, J. (2006). Quantitative Forschung. Ein Praxisbuch. Wiesbaden: VS Verlag für Sozialwissenschaften

Rosenthal, G. (2008). Interpretative Sozialforschung: Eine Einführung. Weinheim/München: Juventa-Verlag

Thomas, A., Utler, A. (2013). Kultur, Kulturdimensionen und Kulturstandards. In: Genkova, P.; Ringeisen, T.; Leong F. Handbuch Stress und Kultur. Wiesbaden: Springer

Wíden, G. (2017). Individual, social and cultural approaches to knowledge sharing. // Journal of Information Science Theory and Practice 5, 3, 6-14

Wikström, E., Eriksson, E., Karamehmedovic, L., Liff, R. (2018) Knowledge retention and age management – senior employees' experiences in a Swedish multinational company. // Journal of Knowledge Management 22, 7, 1510-1526

Zemke, R., Rains, C., Filipczak, B. (2000). Generations at work. New York: American Management Association

Zhang, F., You, H. (2009). Motives of indirectness in daily communication. An Asian Perspective. // CCSE Journal 1, 2, 99-102

https://doi.org/10.17234/INFUTURE.2019.9

Original scientific paper

Event-based Modelling of a Major Historical Government Source
Ministerratsprotokolle 1848–1918

Stephan Kurz
Austrian Academy of Sciences, Vienna, Austria
stephan.kurz@oeaw.ac.at

Wladimir Fischer-Nebmaier
Austrian Academy of Sciences, Vienna, Austria
wladimir.fischer@oeaw.ac.at

Summary

Our paper showcases a critical-historical document edition with a long tradition and of considerable size, the "Ministerratsprotokolle" (MRP). We are currently transferring the MRP to a digital-edition paradigm, based on the XML markup scheme proposed by the Text Encoding Initiative (TEI). Our paper starts out by presenting the corpus and discussing the workflows that lead to the present state of the MRP data. Our main task is to edit, but also to disseminate this important digital Cultural Heritage resource. In order to open access to a broader public quickly, our choice fell on the easiest to process and most general category in our code: events. Events in our case include first of all the dates of ministerial council sessions and the agenda items discussed during these sessions. For these two kinds of events, we propose a markup strategy that is compatible to RDF statements, linking documented text and facts which the text is referring to. To insure reusability from across all disciplines, we are using a prototype eXist-db application that serves the data both as TEI XML and via API. The aim of our paper is twofold: To theoretically discuss event-based modelling of textual resources, and to describe the corpus unlocked by this type of modelling.

Key words: digital edition, data modelling, event-based modelling, Linked Open Data, mass data, event, Text Encoding Initiative

Introduction

The *Ministerratsprotokolle der Habsburgermonarchie und Österreich-Ungarns* (MRP) are a major corpus of governmental documents stemming from the Habsburg Monarchy's administrative legacy.[1] Covering nearly sixty years of government, the minutes (protocols) of the Ministerial Council are one of the few edited resources that on the one hand display the inner workings of the Monarchy's governments, and represent a huge data mine full of prosopographic, political, administrative, economic, cultural, and social information in general, on the other. Structurally, the MRP are organised as a series of session events that each include agenda item events.

Our research institution is responsible for the MRP's scholarly edition under the proposition that cultural heritage must be steadily curated to keep it in circulation. To all historians, preserving historical heritage and making it accessible to the public in a scientifically prepared form is what basic research means. For historians in a digital paradigm, the on-line availability and functionality of historical textual resources has become key, as "[h]istorians consider the content of the text 'data', and they want to use this data in their research to gain knowledge about the past." (Vogeler, 2019: 309)

In our paper, we are presenting the data and underlying event-based data model that has been implicit in the MRP edition's text. We want to make this underlying model explicit in order to make the data more accessible to research outside of the historical disciplines. We are discussing two converging, but distinct data sets within our overall edition project: one comes from a large amount of retro-digitized material, the other one derives from our current "hybrid edition" research and editing

[1] Throughout the paper, we will reference the material we are discussing by the siglum "MRP." For general information on the edition series and on the editing guidelines that have not substantially changed since their inception (Rumpler, 1970). For details on the digital edition workflow (Kurz et al., 2019).

process. But before focussing on the editorial workflow and specifically on entities and events implied, allow for a brief overview on the edition's contents and contexts.

The edition corpus and data

The MRP minutes document the Council of Ministers of the Austrian Empire (which included Hungary at the time) from its advent in 1848 up to 1867 (Series 1). The 1867 Austro-Hungarian Compromise transformed the Austrian Empire into the "dual" Austro-Hungarian Monarchy. Three bodies emerged from the hitherto unified Council of Ministers: the Joint Council of Ministers of the Austro-Hungarian Monarchy (Series 2), a Council of Ministers for the Kingdom of Hungary, and one government for the remaining countries, abbreviated as Austria or Cisleithania, all of them spanning the years 1867 to 1918. The minutes of the latter government's sessions are being edited as "The Minutes of the Cisleithanian Council of Ministers, 1867−1918" (Series 3).

The Council of Ministers or *Ministerrat* was the central body of government. The minutes of its sessions reflect all aspects of political life, from issues concerning the state's structure and organization to social, economic and technical developments as well as cultural and social problems. The protocols were journalized by dedicated recording clerks and later on circulated among the participants, before being presented to the emperor for his formal decision ("Allerhöchste Entschließung"), which put their content into immediate effect.

The structure of the minutes is relatively uniform. Each one starts with a list of the members of government present or absent, and of domain specialists invited at certain occasions. Then follows a table of contents and, as the main part, a detailed summary of the propositions and discussions having taken place during the meetings. When particularly controversial topics were on the agenda, the exchange of opinions is well graspable, but usually, the texts are concise summaries of propositions, arguments and outcome. It is important to note that the minutes are not recorded in direct speech, but rather written in the third person.[2]

Who is the document edition good for? Classical historians have thus far used it for intellectual information extraction, mostly with regard to political decision making. But the minutes are full of information both of encyclopedic value and with a potential for quantitative processing. Names of persons involved in political affairs as well as lower employees of the state, extending also to foreign officials, are abundantly present. Attached to those persons are information on their lives and careers, titles and decorations, as well as their relations to other persons. A plethora of institutions are being mentioned, which over time changed both their names and responsibilities. The same is true for place names and regional entities from all over the Monarchy, from present-day Montenegro to Poland, from Ukraine to Italy. Finally, most of the agenda items refer to at least one law or decree from the Austrian legal codes. The language of the minutes and the arguments made are large-scale specimens of late 19th-century administrative German and of the political discourse of Austria-Hungary's elites. And this is only the text of the minutes themselves. The scientific comment in footnotes and the extensive introduction of each volume, enriches all this with references to contemporary news articles, legal codes and to all relevant other minutes connected to the respective agenda item. To sum up, the MRP combine "text from the source with interpretation by and for historians" (Vogeler 2019: 312) and are thus good for classical historians and also for cultural, social, economic, and quantitative historians, as well as corpus linguists, historical linguists and discourse analysts (among others).

All these raw materials that have been buried between book covers (for general notes on related challenges see Piotrowski, 2012), are waiting to be transformed into the largest historical data mine of the political, cultural, legal, economic, and technological history of the Habsburg Empire.

[2] This modal difference keeps us from directly using teiParla, the recently developing standard for parliamentary minutes: Erjavec/Pančur recently organized a workshop for a proposed teiParla standard, see https://www.clarin.eu/event/2019/parlaformat-workshop. At the Austrian Academy of Sciences, T. Wissik will be applying this to the ParlAT corpus of contemporary Austrian parliament debate transcripts.

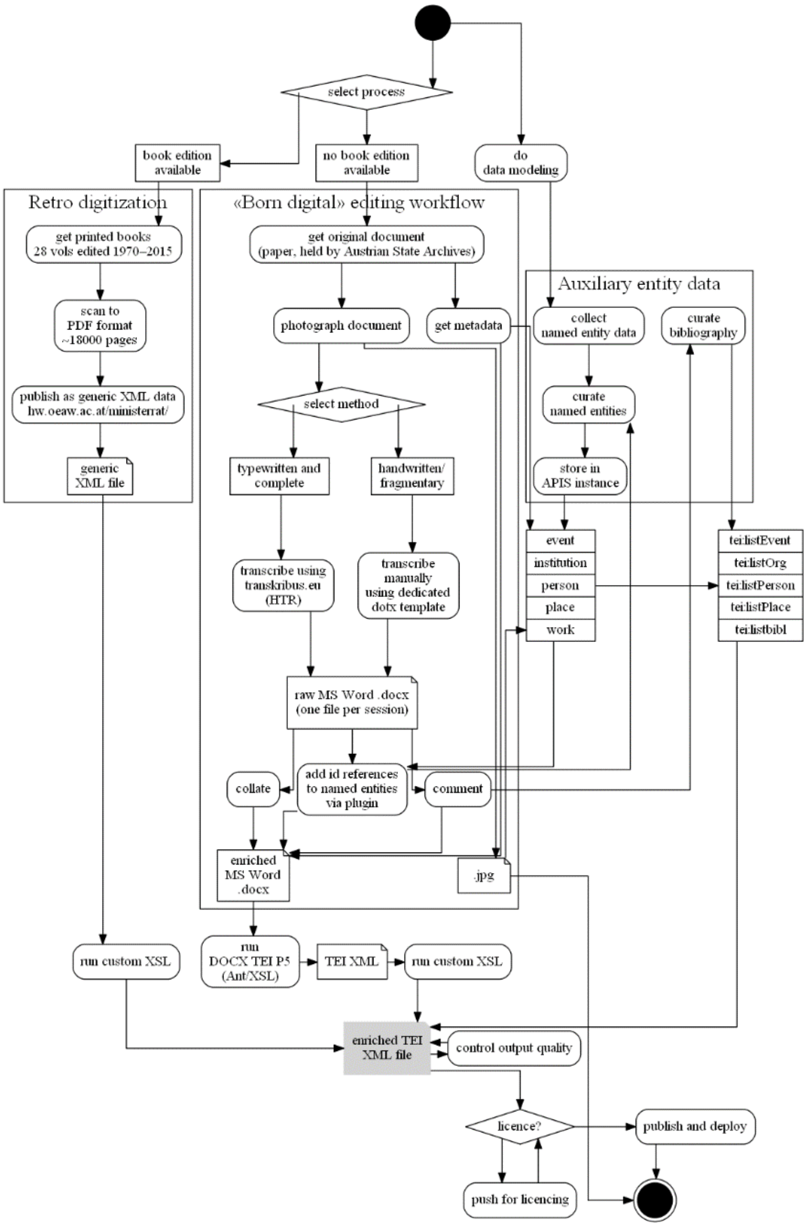


Figure 1. Ministerratsprotokolle 1848–1918: Digitisation Workflow Activity Diagram. Source: own work. Graphviz source files for diagrams are available in the [GH02] repository

In the following, we present the detailed editing process in order to explain the important issues of this edition (Fig. 1, for the deployment infrastructure cf. Fig. 2 below). It starts with “sources,” i.e. thousands of more than one-hundred-year-old documents in the Austrian State Archives: handwritten and typewritten minute forms. The minutes of the 1848–1867 period have been transcribed in the past forty years by hand, using a word processor. Editors have then annotated, commented and supplemented them with lists of terms to be used in indices in the appendix. After a manual typesetting step done in Adobe InDesign, editors then highlighted terms intended for the index by felt marker and added them to the index manually, before the books were printed. The resulting 18,000 pages are the data we are retro-digitizing.

From Print to Hybrid

The transformation of the print edition to a full-text source has met some difficulties. The first volumes of the edition were still typeset in the pre-digital age. Subsequently, machine-readable original files of the transcripts were used, but many of them have been destroyed. The same is true for the InDesign files of the more recent volumes. Therefore, the 28 already edited print volumes had to be scanned and re-keyed and have been made available in a generic XML format by the Austrian

Academy of Sciences' in-house publisher. This did not include any additional semantic markup, although the actual content is both structured and provides semantic clues. In other words, the print edition is currently "only" available as flat text data.

To remedy this situation and provide better access to the existing data, our research institution have decided to convert the data into the widely-accepted XML format proposed by the Text Encoding Initiative (TEI Consortium, 2019), for which different tools are already available in order to enrich the data semi-automatically. Further reasons to opt for TEI encoded XML as target format include: It is -) standardized (Bunard, 2019), -) both human and computer readable, -) highly accepted as an exchange and archiving format, -) able to accommodate a superset of our editorial needs, given the limited markup depth that we can afford to produce within the given time constraints, and -) it fits both of our data sets, which saves development effort and enables us to think about the MRP as one common data source. Due to unsolved licencing questions at the Austrian Academy of Sciences Press, we currently cannot publish the TEI full-text derivates of the retro-digitized material. Generic XML data is available at [HW] under legally unclear "Open Access" conditions.

Since 2018, we have been using an XML-based hybrid edition workflow that moves the semi-automatic semantic markup a step closer to the transcription phase. For the three concurrent volumes that we are editing at the time of writing, we are using an MS Word .docx workflow (cf. Fig. 1, 'Born digital editing workflow'). This includes linking entities via a JavaScript plugin and automates part of the structural parameters of the edition text by applying paragraph and character styles context-sensitively.[3] The resulting .docx XML data is then pre-processed through standard OxGarage OOXML .docx to TEI templates. The resulting documents are further customized using XSL transformations, especially regarding genre-specific structural markup and other details. The TEI output of both the "born digital material" and the "retro digitization" pipelines are equivalent, which is important for the next steps of processing and for the deployment of the output.

When assessing the distinct qualities of our edition as a whole, we concluded that we are dealing with a trove of dated (or dateable) facts that have left their traces in the records. Although in the TEI world, the use of the event element has not been very prominent,[4] for our edition, events turned out to be a central structural element, which we use both for the structural organisation of textual elements and for providing new discovery tools on the application level.

What is the background of this? For enriching both the retro-digitized and the born-digital MRP data, we look for automatically detectable items that yield date strings identifiable by regular expression string matching. Those either refer to 1) acts of reference or citation, or to 2) events outside the textual sphere ("facts"). In the first case, depending on the context before and after a date string, we are able to add links to external data sources, such as the digitized newspaper (ANNO) and legal documents (ALEX) archives held by the Austrian National Library, and also to reference points within the MRP corpus itself (such as a mention of preceding or following sessions). In the second case, and even more prominently, we model the minutes themselves as describing events on the level of agenda items: The session event of a particular day contains references to all points on the agenda, which we also understand as events. Each of the agenda items has one or more actors whose presence and utterances are recorded in relation to the agenda item event. Both the events and relations have already been present in the MRP paper edition and digitized full text in the "short regesta" or abstracts, but only implicitly. Taken together, we are in a position to not only hyperlink existing references and dates but to also make the structure of events explicit in the markup of the edition data. Thus, it is safe to say that the whole MRP edition hinges on events.

Entities and Events

Most of the entities can be modelled with obvious and well-used elements of the TEI vocabulary: dates in time, named entities, such as individual persons, geographical places, bibliographical entities such as laws and newspaper articles.[5] Yet, where possible, the mere occurrence of a date has to be properly rewritten as an event to gain meaning.

[3] For examples of work-in-progress .docx documents, cf. http://mrptestapp.acdh-dev.oeaw.ac.at/.

[4] Since 2010, there have been multiple interventions on the TEI mailing list that aim toward referencing events from within the text, but so far, no single solution has been adopted by the TEI community.

[5] MRP related challenges concerning named entities and building indices from them are discussed in Kurz/Zaytseva 2019.

In order to collect the event structure and link the session's agenda item texts to the events via a stable URI for the born-digital part of our data, we are using a relational database and web front-end (dubbed "Auxiliary entity data" in Fig. 1). This is done with an instance of the APIS database system,[6] in which we are creating entries for the event, person, institution, place and work entity types, including authority file identifiers where applicable. For the majority of textual mentions such as person names, this is done semi-automatically, as names of ministers and other government officials are recurring in the texts; other entities do not even exist in any of the usual Linked Open Data sources (WikiData, VIAF/GND, GeoNames); we have to manually add those to our database while editing the source texts.[7]

However, it is a different case with the pre-existing texts from the printed books: We cannot re-edit the textual data from scratch since our time resources are limited. Therefore, we can only apply automatic information extraction strategies one step at a time, and we have opted to start out with events:

Since dated facts are equally common in both our data sets, we attempt to at least match the majority of possible dateable facets, and wrap them in date/@when-iso elements for further analysis. The respective xsl:analyze-string regex solution targets contemporary writing styles for dates, including some abbreviated forms employed in the MRP text that would otherwise need domain specific expertise and/or manual reading (e.g. "1. l. M." refers to the first of the current month ['laufenden Monats'], in this case 1864-07-01).

Although we originally only have a string representation, we can infer that we are indeed dealing with a dated event that is relevant to the MRP corpus, and for which we understand "event" to be a change of conditions related to a subject and an object that took place at a specific point in time. Consequently, any event can be expressed as a subject–predicate–object triple with a date attribute. This is purposefully compatible with a triple-based logic at least for the description of events that are linked to the edition text.[8] Currently, we are only applying this logic to the textual facets mentioned, as they present extra-textual *facts* that form interpretational additions to the underlying *text* on the editor's behalf. Hence, we separate them from the latter: The fact that a council session took place (a particular topic was discussed) is extraneous to the minute's text – it is a (well-sourced) observation by the editor.

In modelling the data structure for the born digital workflow, our starting point were the textual units we refer to as "agenda items." These are the basic units of the digital edition's layout and data model. A ministerial council session is not tied to a certain date in a one-to-one relation, it could be intermitted and last for several days. Therefore, we decided to use the agenda item as the basic unit as it can in almost all cases be tied to a particular singular date. After this decision we could define which TEI element best fits the agenda item. Our choice was to encode the textual content as div type="agenda_item" in the document's body, and to additionally replicate the label assigned to the agenda item in the "Protokollbuch" (book of protocols, a second source that is physically distinct to the actual minutes) into the event element in the profileDesc/abstract/ listEvent of each XML file's teiHeader.

In the model created while transforming generic retro-digitized XML data into TEI, we replicated the session event and the contained agenda item events directly via XPath selection. These are the only events where this is possible without manual intervention with all necessary data regarding the *who*, *what*, *when* and *where*. For *who*, we construct a listPerson with role attributed person elements. *When*

[6] For more on the project during which this Austrian Prosopographical Information System was developed, cf. https://apis.acdh.oeaw.ac.at/. Python and a PostgreSQL database drive its backend, with a Django frontend in place for manual curation and an API e.g. for autocomplete plugins like the one we are using: https://github.com/dariok/officeEntityPlugin.

[7] Named entity recognition and the variety of challenges that come with it is beyond the scope of this paper. Among other strategies, we have been experimenting with the Pelagios Recogito system that uses Stanford NLP tools for additional NER markup, cf. https://recogito.pelagios.org/. For the retro-digitised part of our corpus, named entities that form the base of person and place indices will have to be added at a later stage.

[8] It would even be possible to remodel the edition completely based on RDF triples, as the Swiss "Nationale Infrastruktur für Editionen – Infrastructure nationale pour les éditions (NIE – INE)" (https://fee.unibas.ch/de/nie-ine/) are successfully proposing. Apart from the necessary resources for such a transition, the humanities environment community, in this case the MRP editors, are favoring a more human-readable TEI approach over a pure LOD approach for the time being. For discussion of the encoding of RDF relationships within the TEI, cf. https://github.com/TEIC/TEI/issues/1860.

and *where* can be inferred by string parsing; they are constructed from the source XML during XSLT processing together with the list of agenda items. The all-important *what* is populated from the session's formal heading in the minutes and the agenda item's description in the "Protokollbuch," respectively.

The Guidelines of the TEI Consortium propose the use of the event element as "data relating to any kind of significant event associated with a person, place, or organization." Thus, it can be placed in the descriptions of said entities, e.g. to list events in a biography that are related to one natural or legal person, or that took place at a particular place. In addition, the element may be used on its own, as long as it is either nested, or grouped in a listEvent container within most of the analytical descriptors the TEI guidelines have on offer. For the MRP edition, we chose to accommodate a listEvent within the teiHeader, wrapped in a profileDesc/abstract construct. This provides convenient stand-off markup that models the event as categorically different from the text that is describing the event; in other words, our approach avoids mixing up interpretation and textual or documentary evidence (Vogeler 2019: 318).

Moreover, our proposed usage of event does not call for any adaptations of the current TEI definition of event.[9] For the time being, we only cover the macro structure of historical facts considered events. There are numerous examples for other events that may be extracted in the future, e.g. we could also lend event status to a text passage giving evidence that a particular minister said something within the scope of one agenda item (of type "utterance," which might include one of type "quotation" etc. ad libitum) like "Der Eisenbahnminister erinnert daran, dass …."

We see the following advantages of our decisions, so far:

- Our model adopts a TEI based workflow for both retro-digitised and newly edited content and thus contributes to the broader aim of making Cultural Heritage available in the digital age through standardisation and reusability.
- Putting events in the centre of our data model allows for easier dealing with the problem of interlinking textual documents and extra-textual facts.
- By using events as hinges between fact and text, we contribute to easier and sustainable accessibility of the documents we are editing, and to the development of new discovery tools.

While implementing the workflow outlined above, we also had to create a working environment not only for data input and curation, but also for the eventual publishing of the output. In Figure 2, we show the deployment components involved.

[9] This converges with the fact that source editions from the historical disciplines tend to use abstracts (short regesta) to sum up the content of the given text. As this practice is not universal across the disciplines, there are efforts to make listEvent more interchangeable while keeping the distinction between evidence and interpretation. A joint paper of one of the authors with scholars of other disciplines has been presented at TEI 2019, cf. the "Recreating history through events" paper by C. Fritze, H. Klug, S. Kurz and C. Steindl, https://www.conftool.com/tei2019/sessions.php. In a nutshell, this paper proposes an additional eventName element (in parallel to pers|place|orgName) for the purpose of inline referencing, extending the attributes of event with @who and @dur, and further promoting the use of listEvent with the goal of providing a "calendar" webservice that interlinks existing TEI-based digital scholarly editions.

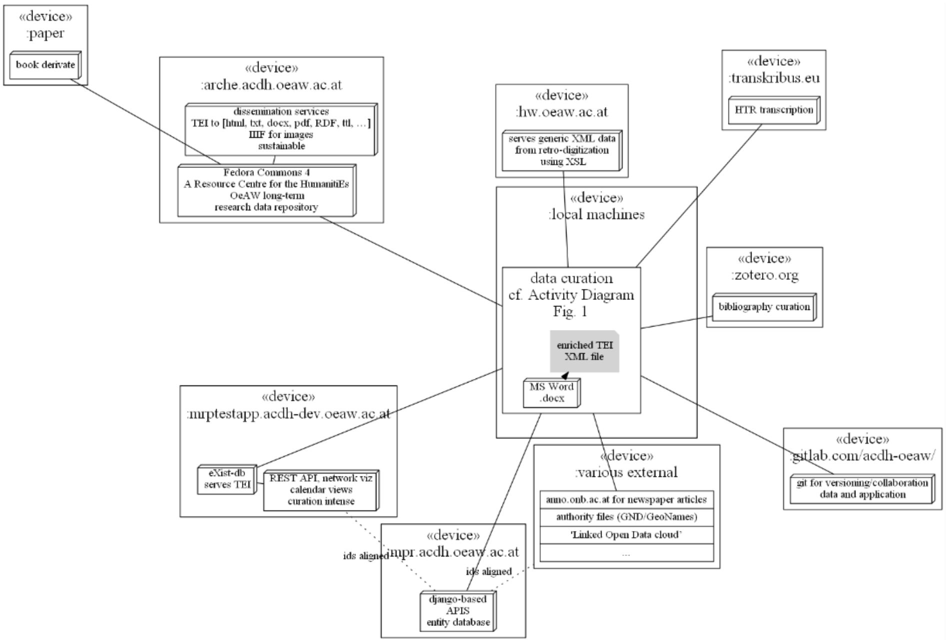
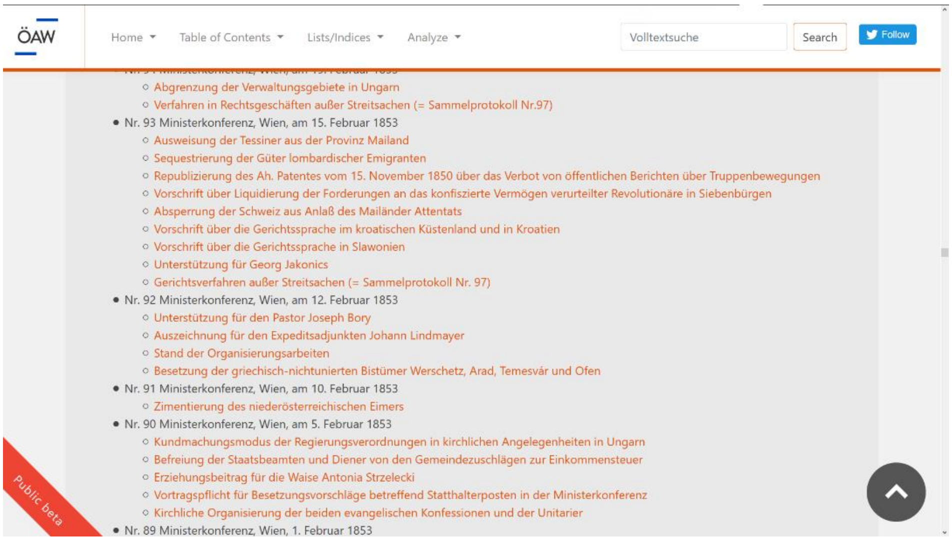


Figure 2. Ministerratsprotokolle 1848–1918: Deployment Diagram. Source: own work

A listEvent of 27 volumes of the Ministerratsprotokolle 1848–1867 edition as outlined above is available at [MRPTESTAPP] under CC-BY licence; it features both the single sessions (2301 entries) and the respective agenda items (10959 entries) which are modelled as events. This is the first time that a complete "table of contents" of all agenda items in the whole first edition series is available to the public (Fig. 3).

The same showcase application also displays a selection of full-text protocol XML files from our current editorial work on three volumes of the 1867–1918 series, which include listEvent data in their teiHeaders.

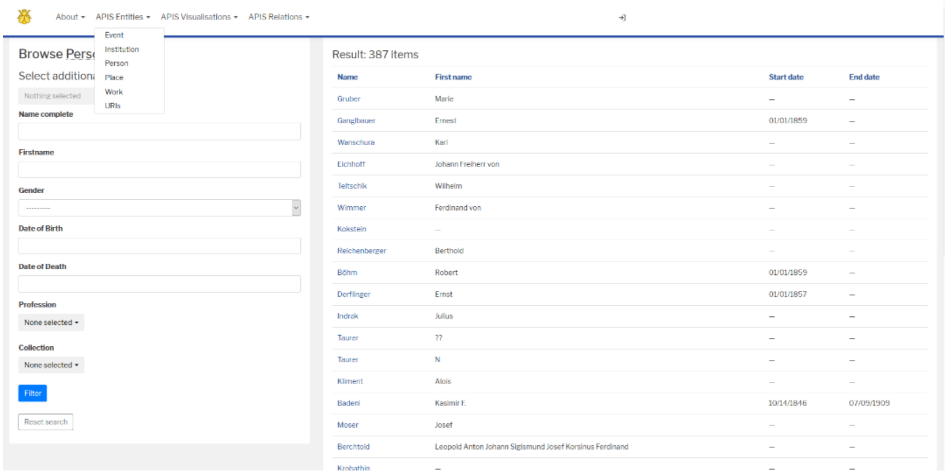
Sources for the eXist-db application based on the KONDE dsebaseapp are available under [GH01], showcase data are kept in [GH02]. A screenshot of the APIS-based entity database is provided in Fig. 4.



- Abgrenzung der Verwaltungsgebiete in Ungarn
- Verfahren in Rechtsgeschäften außer Streitsachen (= Sammelprotokoll Nr.97)

- Nr. 93 Ministerkonferenz, Wien, am 15. Februar 1853
  - Ausweisung der Tessiner aus der Provinz Mailand
  - Sequestrierung der Güter lombardischer Emigranten
  - Republizierung des Ah. Patentes vom 15. November 1850 über das Verbot von öffentlichen Berichten über Truppenbewegungen
  - Vorschrift über Liquidierung der Forderungen an das konfiszierte Vermögen verurteilter Revolutionäre in Siebenbürgen
  - Absperrung der Schweiz aus Anlaß des Mailänder Attentats
  - Vorschrift über die Gerichtssprache im kroatischen Küstenland und in Kroatien
  - Vorschrift über die Gerichtssprache in Slawonien
  - Unterstützung für Georg Jakonics
  - Gerichtsverfahren außer Streitsachen (= Sammelprotokoll Nr. 97)
- Nr. 92 Ministerkonferenz, Wien, am 12. Februar 1853
  - Unterstützung für den Pastor Joseph Bory
  - Auszeichnung für den Expeditsadjunkten Johann Lindmayer
  - Stand der Organisierungsarbeiten
  - Besetzung der griechisch-nichtunierten Bistümer Werschetz, Arad, Temesvár und Ofen
- Nr. 91 Ministerkonferenz, Wien, am 10. Februar 1853
  - Zimentierung des niederösterreichischen Eimers
- Nr. 90 Ministerkonferenz, Wien, am 5. Februar 1853
  - Kundmachungsmodus der Regierungsverordnungen in kirchlichen Angelegenheiten in Ungarn
  - Befreiung der Staatsbeamten und Diener von den Gemeindezuschlägen zur Einkommensteuer
  - Erziehungsbeitrag für die Waise Antonia Strzelecki
  - Vortragspflicht für Besetzungsvorschläge betreffend Statthalterposten in der Ministerkonferenz
  - Kirchliche Organisierung der beiden evangelischen Konfessionen und der Unitarier
- Nr. 89 Ministerkonferenz, Wien, 1. Februar 1853

Figure 3. Ministerratsprotokolle 1848–1918: Screenshot of the eXist-db based application [MRPTESTAPP] displaying the 1848–1867 listEvent table of contents. Source: own work



Result: 387 items

Name	First name	Start date	End date
Gruber	Marie	–	–
Ganglbauer	Ernest	01/01/1859	–
Wanschura	Karl	–	–
Eichhoff	Johann Freiherr von	–	–
Teltschik	Wilhelm	–	–
Wimmer	Ferdinand von	–	–
Kokstein	–	–	–
Reichenberger	Berthold	–	–
Böhm	Robert	01/01/1859	–
Derflinger	Ernst	01/01/1857	–
Indrak	Julius	–	–
Taurer	??	–	–
Taurer	N	–	–
Kliment	Alois	–	–
Badeni	Kasimir F.	10/14/1846	07/09/1909
Moser	Josef	–	–
Berchtold	Leopold Anton Johann Sigismund Josef Korsinus Ferdinand	–	–
Krobathin	–	–	–

Figure 4. Ministerratsprotokolle 1848–1918: Screenshot of the APIS-run entity database. Source: own work

Conclusion and Outlook

The intention of our project is to open up new access paths to a pre-existing historical textual resource, in a sustainable manner. Much effort has been put into the transcription and commentary of the paperwork that puts the inner workings of post-1848 Habsburg empire governance on historical record. The intention to publish the outcomes of multiple decades of historical research has to include new access routes through non-serial (non-book) representation of the textual data, if we are to follow the digital paradigm that has been articulated e.g. by Sahle 2013, but already by Shillingsburg 2006. In short, this paradigm is defined by the notion that a digital scholarly edition differs from a paper-bound scholarly edition in that it does not merely constitute a digitized version copying the features of a printed book. Instead, it defines a digital edition by its essential incompatibility with a (broadly speaking) linear concept of text-as-book – such an edition cannot be losslessly converted into book form. This includes linking to and/or drawing from external resources like in the case of authority files e.g. with the use case of disambiguation of personal names.

In this sense, the digital edition of the minutes of the Council of Ministers goes beyond the functionalities of the volumes published to date; a print version is nevertheless produced in the chosen hybrid edition approach; it offers a reading typography in a similar fashion to the previous volumes, including all paratexts. In many respects, the future MRP digital edition transcends the paradigm of the book, as it offers:

- the whole set of features from the printed book series, i.e.
    - scientific introduction
    - lists of outdated expressions and participants of the council
    - minute texts, including their abstracts, and addenda
    - indices or persons, places and institutions
- multi-volume facetted full-text search,
- enhanced display and filtering options compared to the print product,
- supplementary facsimiles,[10]
- extensibility in the case of new source finds,
- the addition of authority file data and linked data,
- the development of new audiences through extended visibility,
- also in conjunction with access to the underlying data via API,
- enhanced workflow and data transparency in comparison to previous book production.

Following the claim that "indices of persons, places, and events and calendars and maps are becoming default components for historical digital editions" (Vogeler 2019: 313), our suggestion is to provide a maximum of additional discovery tools with a minimum of additional editorial effort. With listing

[10] As the MRP edition establishes its texts from various different sources, it will not be possible to keep a complete track of source facsimiles within the given time constraints. Still, we publish facsimiles at least of damaged sources ("Brandakten") with the goal of showing the extent of missing source parts (tei:gap and tei:damage).

almost 11,000 agenda item events with the related governmental staff, a valuable partial data set has been made available already.

Currently, we are not permitted to publish the results of our efforts to standardize and formalize the full-text data of the retro-digitized MRP corpus in TEI markup. Still, we hope to contribute to a discussion on Digital Cultural Heritage that enables researchers not only from the humanities, to make use of the entire corpus that spans over roughly 5 million tokens.

Over the upcoming years, the MRP edition project, one of the long-term projects the Austrian Academy of Sciences has committed itself to, will continue to provide both governmental documents and supporting auxiliary data, thus contributing to historical fundamental research, while also opening up the source data to the public. We are convinced that the administrative history data we provide will spur further research – given the new method and technical dissemination we even hope that it will transgress traditional disciplinary boundaries.

References

Burnard, L. (2019). What is TEI Conformance, and Why Should You Care? // Journal of the Text Encoding Initiative 12, 1-22. http://journals.openedition.org/jtei/1777 (03.07.2019)

Dumont, S. (2015). correspSearch – Connecting Scholarly Editions of Letters. // Journal of the Text Encoding Initiative: Selected Papers from the 2015 TEI Conference 10, 1-21. http://journals.openedition.org/jtei/1742 (03.09.2019)

Grishman, R. (2015). Information Extraction. // The Oxford Handbook of Computational Linguistics. 2nd ed. / Mitkov, R. (ed.). https://doi.org/10.1093/oxfordhb/9780199573691.013.009 (03.09.2019)

Kurz, S., Fischer-Nebmaier, W., Kampkaspar, D., Lein, R., Schmied-Kowarzik, A. (2019). Die Edition der Ministerratsprotokolle 1848–1918 digital: Workflows, Möglichkeiten, Grenzen. // 5. Digital Humanities Austria Konferenz DHA 2018 Conference proceedings / Zeppezauer-Wachauer, K., Hinkelmanns, P., Ernst, M. (eds.). Salzburg/Wien (in print)

Kurz, S., Zaytseva, K. (2019). Herausforderungen für Thementhesauri und Sachregister-Vokabularien zur Erschließung im Kontext des digitalen Editionsprojekts Cisleithanische Ministerratsprotokolle. // DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts / Sahle, Patrick (ed.). Frankfurt/Main

Piotrowski, M. (2012). Natural Language Processing for Historical Texts. // Synthesis Lectures on Human Language Technologies 17. https://doi.org/10.2200/S00436ED1V01Y201207HLT017

Rumpler, H. (1970). Die Protokolle des Österreichischen Ministerrates 1848–1867, Einleitungsband; Ministerrat und Ministerratsprotokolle 1848-1867. Behördengeschichtliche und aktenkundliche Analyse. Wien: Österreichischer Bundesverlag

Sahle, P. (2013). Digitale Editionsformen. Norderstedt:BoD. // Schriften des Instituts für Dokumentologie und Editorik 7-9

Shillingsburg, P. L. (2006). From Gutenberg to Google: electronic representations of literary texts. Cambridge: Cambridge University Press

Parthenos. Standardization Survival Kit. https://ssk.readthedocs.io/ (03.09.2019)

TEI Consortium (2019). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.6.0, July 16, 2019. https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf (03.09.2019)

Vogeler, G. (2019). The ‘assertive edition’. On the consequences of digital methods in scholarly editing for historians. // International Journal of Digital Humanities 1, 2, 309-322. https://doi.org/10.1007/s42803-019-00025-5

https://doi.org/10.17234/INFUTURE.2019.10

Review paper

The Intermesh of Records Management Principles and Enterprise Architecture
A Framework for Information Governance in the Swedish Context

Anneli Sundqvist
OsloMet - Oslo Metropolitan University, Norway
annsu@oslomet.no

Tom Sahlén
Sahlén Arkivkonsult AB, Sundsvall, Sweden
info@arkivkonsulthb.se

Mats Andreasen
Blitt, Sweden
mats.andreasen@gmail.com

Summary
To capture and control records generated in complex processes, involving many different actors in fluid constellations, organizations need tools that extends the range of traditional records management practices. Records management needs to be incorporated in more encompassing information governance regimes and integrated with the organizations' overall management systems. The authors argue that the intermesh of records management principles and Enterprise Architecture is a fruitful approach in the development of coherent information governance regimes. The paper presents a framework for information governance based on records management principles and Enterprise architecture and a methodological approach how to develop and implement information governance solutions by integrating records management with Enterprise Architecture using agile methods, Design thinking and User Experience Design (UX). The work is based on literature reviews and modelling workshops and is a result of an on-going development project aiming at methodological development to improve records management practice.

Key words: information governance, records management, enterprise architecture, agile methods

Introduction

The development of information technology in interplay with social and political changes has brought about a considerable shift in communication practices, production and work processes. One effect of these changes will be vast amounts of information. A fair amount of this information will constitute records and object to records and archives management. Records is here used as an inclusive concept that covers both transactional information, and information created and captured during business performance not as evidence of transactions, but because they represent some value or use for the organization. As lately has been more widely recognized, records are *assets* (ISO15489:2016; ISO/TR21965:2019). Assets that have to be managed and could add value to the organization (and to society), but also could be added value through appropriate management. In contemporary organizations records are not just the documentary result of business processes, they are integral to the processes, and, in some instances, the creation of records constitutes the actual performance of business activities. The management and control of records could thus not be undertaken as separate support functions external to the control of the business processes, but has to be integrated with the organizations' overall management systems. This means that holistic governance models are required, as well as methodological tools. *Information governance* (IG) is a concept that has been propagated within the field of information management and corporate governance the last 15 years, and it has also gained recognition within the records management community (Brooks, 2019; Hagmann, 2013). The alignment of RM with IG is advocated by several representatives of the RM sector, for instance

Franks (2013), Lomas (2010) and ARMA International, whose *Generally Accepted Recordkeeping Principles* is promoted as a framework for good IG practice.

The purpose of this paper is to present a framework for information governance based on records management principles, and a methodological approach how to develop and implement information governance solutions by integrating records management with Enterprise Architecture using agile methods, Design thinking and User Experience Design (UX).

The paper is based on literature reviews and modelling workshops. It reports the work of an on-going development project aiming to produce guidelines for the implementation of information governance regimes in a Scandinavian context and to contribute to methodological development. The project is performed as a practical action research project (Denscombe, 2014), that is it addresses a specific problem within a particular community, records management professionals, to improve practice. The project group consists of a team including academic researchers, records professionals, IT-architects, all contributing with experience from their respective field of competence and with vast experience of Scandinavian records management practice.

A framework for information governance

A common definition of IG is Gartner's (2019) "the specification of decision rights and an accountability framework to ensure appropriate behavior in the valuation, creation, storage, use, archiving and deletion of information. It includes the processes, roles and policies, standards and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals", which also is embraced by ARMA International. Usually IG involves compliance, information security, risk management, privacy, data management, big data, e-discovery, *and* archives and records management (e.g. Reed, 2017). The established definition of RM is a "field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposition of records, including processes for capturing and maintaining evidence of and information about business activities and transactions in the form of records" (ISO 30300:2011, clause 3.4.3). RM could thus be seen as a subset of IG, with a more specific target and mandate (e.g. Hagmann, 2013; Lomas, 2010; Reed, 2017; Saffady, 2015). However, with a more inclusive conceptualization or records as suggested above, it could be argued that most information handled by IG would qualify as records. RM requirements could thus also apply to the management of other information resources (cf. ISO/TR 21965:219(E), p. v), in order "to ensure appropriate behavior in the valuation, creation, storage, use, archiving and deletion of information". The crucial notion here is, however, *governance,* here defined as "[t]he method by which and enterprise ensures that stakeholder needs, conditions and options are evaluated to determine balanced, agreed-on enterprise objectives are achieved. It involves setting direction through prioritization and decision making; and monitoring performance and compliance against agreed-on direction and objectives", which could be contrasted with *management,* which is about the planning, building, running and monitoring of "activities in alignment with the direction set by the governance body to achieve the enterprise objectives" (ISACA, 2019). Governance is about setting goals, deciding strategies, and defining roles and responsibilities, while management is about the control and execution of business activities. Governance is a matter for top management, but it can only be carried out with help of executive functions, e.g. RM functions, and management systems.

IG is today a well-known concept, but there is no established standard concerning general IG. Thus, the implementation of IG has to rely on the implementation of related frameworks and standards, e.g. ISO/IEC 38500, COBIT, ISO/IEC 27000, the aforementioned GARP, and not the least the ISO 30300 series - Management systems for records. A management system is "a set of interrelated or interacting elements of an organization to establish policies and objectives, and processes to achieve those objectives" (ISO 30300:2011, clause 3.4.1), and a management system for records (MSR) aims "to direct and control an organization with regard to records (ISO 30300:2011, clause 3.4.2). The implementation of a MSR would provide a basis for a comprehensive IG framework, requiring RM to be linked to the objectives of the organization. The following section will present a generic framework for RM adapted to Scandinavian and particularly Swedish conditions, developed as an outflow of continuing RM practice during more than 25 years (Bergbom et al., 1994; Bodin, Sahlén, Sjögren, 2000; Sundqvist, 2005; Sahlén, 2016). The framework takes its stance in the Swedish conceptualizations of records and archives and the established division of labour and responsibilities. By linking the framework to the requirements of ISO15489:2016 and the ISO30300-series, but also to some extent the ISO27000, a comprehensive IG regime could be established. Five generic[1] RM functions are identified, that together with a management function perform the planning, execution and control of RM activities in an organization. Those functions should not be regarded as administrative units, and in practice there are no Chinese walls between them. They represent the fundamental measures that have to be taken to implement a good practice of RM - to guarantee the production, maintenance, preservation and reuse of records. In an analogue environment, those will usually be performed sequentially following a life-cycle perspective, but in the digital world they proceed in a continuum, partly parallel with each other. The management function has a central role, responsible for the control of the operational RM functions and the implementation of policies and strategies, organization, and processes of change decided by the top management (ISO 30301:2011, section 5). That is, the management function upholds the relationship with the overall governance system of the organization. See figure 5 below.

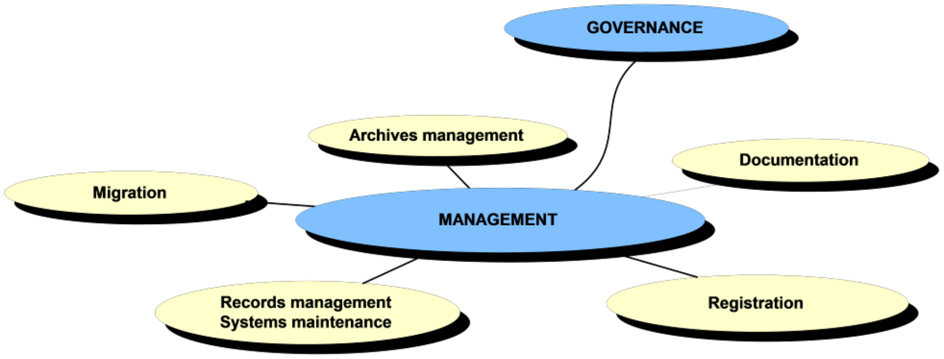


Figure 1. The RM Governance Model (Adapted from Sahlén, 2016: 180)

The role of the five generic RM functions could briefly be described as follows:

- documentation: identification of processes that should be documented; deciding what records that should be created or received; establishing requirements on records (cf. ISO 15489-1:2016, clause 5.2); establishing requirements on records systems (cf. ISO 15489-1:2016, clause 5.3).
- registration: capturing records, registration and journalizing; metadata records management and systems maintenance: administrative and technical maintenance; use and disposition of records and records system as long as they are in active use, including e.g. retention, access, storage and information security
- migration: the process of controlled transfer of records between systems and between the business environment and archival platforms
- archives management[2]: long term management and preservation of semi-active or in-active records system requirements; archival description.

The functions are implemented and managed with help of four continuing activities that can be broken down to a work procedure, developed and maintained by the management function, figure 2.

[1] Functions that exist in every records-creating organization, but not necessarily explicitly defined. Business processes will always be documented, records will be created and captured, managed, transferred and preserved, however not always in a planned and controlled manner or in compliance with legislation or standards.

[2] According to Scandinavian tradition, records management and archives management are closely connected and often performed by the same functions within the organizations.

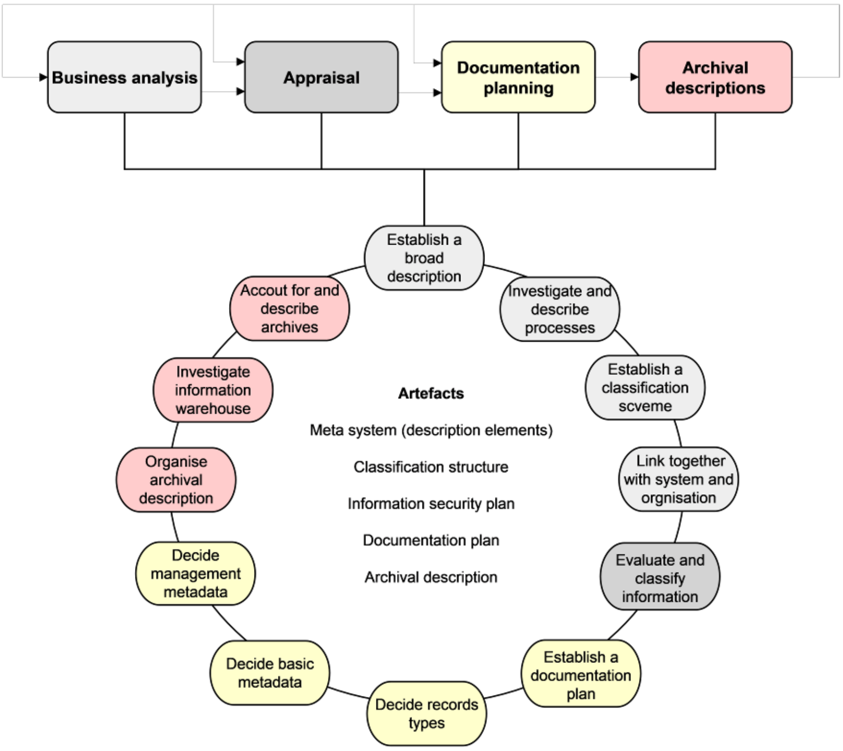

Business analysis
Appraisal
Documentation planning
Archival descriptions
Establish a broad description
Investigate and describe processes
Establish a classification scveme
Link together with system and orgnisation
Evaluate and classify information
Establish a documentation plan
Decide records types
Decide basic metadata
Decide management metadata
Organise archival description
Investigate information warehouse
Accout for and describe archives
Artefacts
Meta system (description elements)
Classification structure
Information security plan
Documentation plan
Archival description

Figure 2. The RM Implementation Model (Sahlén 2016: 93)

The model can be applied both as a means for development and improvement, and for operational RM activities. It has a rather high level of abstraction, and has to be broken down, adapted to local needs, and supplemented with operational tools.[3]

The business analysis – analysis of business context, functions and processes – is the most vital activity that lays the foundation for the other steps in the process. The identification of records, the continuing need for them and the requirements to maintain authenticity, reliability, integrity and usability according to ISO15489:2016 should then be assessed, and the result systematized. The aim of EA is to survey and evaluate the current status, identify fields for improvement and transform the organization to a better state. This is also a component in a MSR, but from a RM perspective it is also necessary to capture the "now". This is a condition for establishing provenance and the evidential properties of records, but it is also a means for the continuing control and usability of records. Describing and cataloguing records is thus an integral component of the process. A result of those activities is a set of *artefacts*[4], instruments for description and control of records, of which the most important are:

- the metasystem – a systematic description of business functions including data about type, organizational affiliation, relation to other functional entities, records requirements, security classification, retention periods, relation to system and architecture entity (cf. International Council on Archives, 2007).
- a classification structure based on functions, processes and activities (ISO 15489-1:2016, clause 8.3)
- an information security plan based on a risk assessment (cf. ISO/IEC 27000)

[3] A prototype of a web application where the Implementation model forms basis for a RM toolbox, i.e. supplementing each step in the procedure with methodological guidelines, can be found at http://www.arkivkonsultab.se/manual-for-informationsforvaltning/

[4] Within EA and related disciplines, artefacts are products describing different aspects of the architecture (TOGAF 9.2, 2018, clause 3.20), and could "range from range from high-level principles to low-level technical diagrams (Kotusev, 2019: 103).

- a documentation plan[5] - a compilation, based on the business classification scheme, documenting all types of (intra-organizational) records required in the business and rules for their management.
- an archival description (catalogue)

Those will then provide tools for the continuing RM cycles, and object to regular revision and improvement.

Enterprise architecture

A prerequisite for IG, and the above presented framework, is the incorporation of RM in more encompassing governance regimes and the organizations' overall management systems, which requires a purposive strategy and course of action. A viable strategy is to integrate RM with Enterprise Architecture (EA). The congruence of RM and EA has been acknowledged rather recently, and there is a limited, but emerging research on the topic advocating a closer connection between the two fields. The research generally concerns how RM principles and requirements could be incorporated into, or supplement EA to enable the management of authoritative information assets (e.g. Becker et al., 2011; Sprehe, 2005; Vieira et al., 2011; Vieira, Valdez, Borbinha, 2011), or how EA can support all-encompassing RM strategies and procedures (e.g. An, 2009; Katuu, 2018a; Katuu, 2018b; Katuu, Ngoepe, 2015; Svärd, 2013).

EA is a tool for analysis, planning and change of organisation and business processes, with the ultimate purpose to meet desired organizational objectives and deliver value to the organization. To achieve this, control and coordination of the organizations' resources and processes is required - to "ensure that the business and IT are in alignment. The enterprise architect links the business mission, strategy and processes of an organization to its information and technology strategy" (ISO/TR 21965:2019(E), p. v). This includes both the fulfilment of direct goals such as providing a certain service to customers, but also non-functional goals such as business-agility – the capability to react to changes. A common definition of architecture is "[t]he fundamental concepts or properties of a system in its environment embodied in its elements, relationships, and in the principles of its design and evolution" (ISO/IEC/IEEE 42010: 2011(E), clause 3.2). EA is characterized by a holistic view of the properties of an enterprise[6] in its environment; its goals and strategies and its components and their relations to each other and to the whole.

EA is usually considered to have originated in the 1980s (e.g. Sessions, 2007), however it has also been argued that the roots of EA, if not the concept, could be traced back to the early 1960s (Kotusev, 2016). The concept of EA is attributed to the at the time IBM employee John Zachman, who in a couple of articles during the 1980s (Zachman, 1982; 1987) appropriated architechural principles in the planning and development of organizational information systems. The resulting so called *Zachman Framework* is a structure for describing an enterprise from different views, which has served as a model for many later EA approaches. Following the initial initiatives, several models and frameworks have been developed.[7]

An architecture framework is defined as "conventions, principles and practices for the description of architectures established within a specific domain of application and/or community of stakeholders" (ISO/IEC/IEEE 42010: 2011(E), clause 3.4.). Such could have a more or less narrow focus - technological or more comprehensive.[8] However, several EA frameworks define four basic domains (e.g. TOGAF 9.2, 2018, clause 2.3):

[5] Work title, should not be confused with Hans Boom's concept concerning appraisal from a societal perspective. The document goes under different descriptions in Scandinavia. It has some similarities with a records retention plan, but is more comprehensive including for instance classification code, security class, access level, medium, format, storage, system affiliation, disposal etc.

[6] An enterprise could be an organization (private, public, commercial, or non-commercial), units of an organization or a group of organizations - "[t]he highest level (typically) of description of an organization and typically covers all missions and functions. An enterprise will often span multiple organizations" (TOGAF 9.2, 2018, clause 3.38). In the following *organization* will be used as a generic term covering all forms of enterprises.

[7] The presumption of this paper is that different EA frameworks or models could form the basis for integration with RM, and none particular is advocated here.

[8] One of the most widely known and applied frameworks is TOGAF (the Open Group Architecture Framework), an American industrial standard first issued in 1995. TOGAF is used as the reference framework in ISO/TR 21965:2019(E), the recently issued standard on records management in EA.

- *business architecture e*– business strategies, governance, organization, and key business processes
- *data* or *information architecture* – the structure of an organization's data assets and data management resources
- *application* or s*oftware architecture* –the individual applications, their interactions, and relationships to the business processes
- *technology* or *infrastructure architecture* – hardware, software and infrastructure required to support the other architectures

It is also common to identify a layer of *solution architecture* on a tactical level, which focuses on a particular problem or business operation and how IS/IT supports that operation (TOGAF 9.2, 2018, clause 3.69). The business and information domains are not primarily about technology, but about how organizations work and how information is understood, modelled and put to effective use. A central element in EA is thus the control of the organization's information assets, why there is a natural link to RM.

Architecture work is distributed between several coactive roles, in principle corresponding to the domains above. A common role setup is the following (IASA, 2019):

- the enterprise architecture function – a function unifying architecture work in the organization with a holistic perspective
- business architect – participates in the development of business strategies to accomplish specific business goals and secures the relationships between business processes, information flows and systems
- information architect – controls storage, retrieval and integration of information needed to carry out business processes
- solution architect – plans the delivery of IT-solutions based on business needs in order to optimize the value of the solution for the organization
- software architect – realizes solutions by structuring and designing software system applications
- infrastructure architect – creates and delivers technology strategies to optimize the organization's use of technology resources, that is hardware, network, technological platforms and physical systems

The roles could be combined, and for instance the Swedish branch of IASA integrates the information architect with the business architect function, and correspondingly the business and information architecture domains.

The relationship between the roles is shown in figure 3. The EA function works at a high level, with the overall scope of the organisation in mind. The Business Architect reaches out down to technical details and mainly acts as a bridge between the business and technology, and works in pair with the Solution Architect who is more concerned with technical details. Software and Infrastructure Architects normally works together with the Solution Architect. In short, the Business Architect identifies business and user needs to make sure that those are fulfilled, while the Solution Architect provides the required technical services.

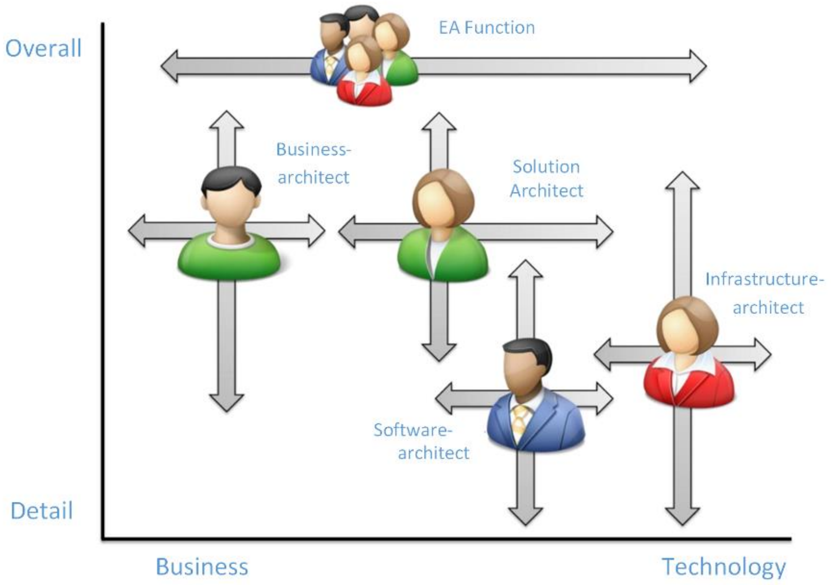


Figure 3. Architecture roles (Akenine et al., 2014: 12, with permission)

The intermesh of RM and EA - a methodological approach

The problems that RM addresses in contemporary organizations, could be characterized as “wicked problems”. That is, they are difficult to pinpoint due to their complexity and interdependences, and one single solution may not be at hand. RM has to take the interests of various stakeholders into consideration (cf. ISO15489-1:2016: vi), navigate within the “new paradigm”, and provide continuity in organisations that continuously are forced to adapt to changing circumstances. Those types of problems are not solved by a linear approach, instead a design process way of working can be more successful. Design thinking involves context analysis, identifying problems, ideation and generation of solutions generating, creative thinking, sketching and drawing, modelling and prototyping, testing and evaluating (Cross, 2011). The design thinking process does not take the outset in a problem, but in the users and user needs to understand and assess possibilities before focusing on solutions. Design thinking is closely related to User Experience Design (UX), the process of enhancing user satisfaction by improving the perceived usability and accessibility of a product or a service (e.g. Hassenzahl & Tractinsky, 2006). These approaches have similarities with agile methods for systems development. Traditional systems development dominating during the 1970s and 1980s, e.g. the waterfall model, was based on a linear and sequential logic where changes should be avoided. However, personal computers and network based work processes, flexible organizations and new market relations demanded more flexible solutions and an adaption to more or less continuous change. The agile movement emerged as a response to this development in the end of the 20th century, advocating an evolutionary, incremental and iterative approach (Fowler, Highsmith, 2001). Agile is now an umbrella term for different systems and software development methods, among which the most recognized are SCRUM and Kanban, characterized by among other things a high level of flexibility and a close collaboration with users/customers (Akenine, 2014: 208-210).

The following will describe a model for collaborative work based on agile methods and design thinking, integrating RM and EA. The model is based on the architecture roles (figure 3 above), and shows how RM could collaborate with those and how work tasks and responsibilities could be divided. Figure 4 below shows how RM could work together with EA, Design thinking/UX and agile methods, to develop and implement IG solutions.

The departure is the need for RM in an organization guided by certain rules according to legislation, standards or other commitments. The RM models, described above, contributes with knowledge of RM requirements and the disciplinary artefacts. The Business architect contributes with business analysis and information analysis, and produces relevant artefacts from that perspective. Design thinking identifies users and produces user scenarios, use cases and drives the work according to method of insight, hypothesis, prototyping, workshops etc. Agile methods provide a way of doing collaborative work in small steps with agile development teams to build shared knowledge, and help with requirements in the form of scenarios, epics and user stories. The agile team follow the process and secure transparency of the work through a shared Kanban or Scrum *backlog*, a list of all new features and changes required to reach a particular outcome.

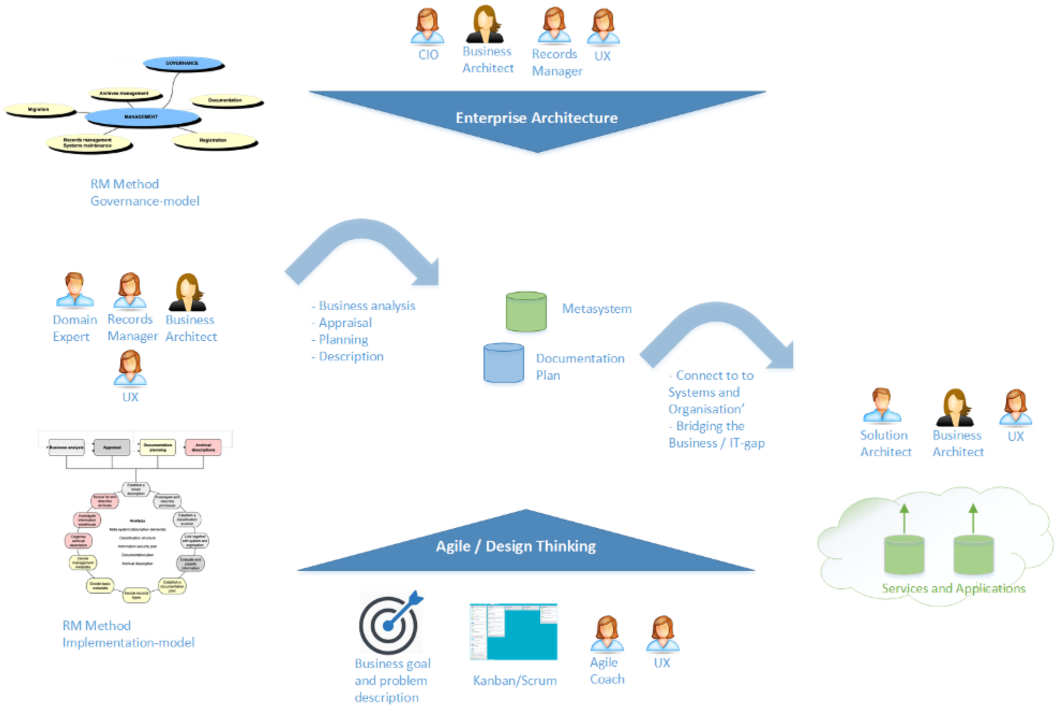


Figure 4. A dynamic work model integrating RM and EA

The figure gives an overview of the method with EA at the top governing the process and agile methods and Design thinking supporting it, while a broad group consisting of several roles or functions perform the steps in the RM method and the transfer to the Solution architect/development team for realization of needed services and applications. This could be done by the agile concept of writing requirements as ‘user stories, typically done by the Business Architect with help from UX experts. The starting point is to define the goal and scope of the work, often an underrated activity. Agile methods and Design process methods aid to the building of a common view of the problem and communicating it to the organisation, as well as providing management with the necessary input to plan and follow up work. This can be done in the form of a *Design Sprint*, a time-constrained, five-phase process used to suggest a solution for a product or service (Banfield et al., 2015), where the Business architect, the Records manager and the Agile and Design process roles participate. Next step is the business analysis, where the Records manager and Business architect work in close collaboration, using an agile approach to address the problem in small steps, and a Design thinking approach using workshops and visualization to collect and analyse large amounts of information in a structured and prioritized way, and feedback to validate the understanding via rapid prototyping. An efficient method in this stage is *Capability modelling*, which corresponds to the first step in the RM Implementation model. A capability is “[a]n ability that an organization, person, or system possesses” (TOGAF 9.2, 2018, clause 3.30), emerging from the goals and objectives and realized by business processes. Capabilities are stable, high-level representations of *what* the organization aims to do. Capabilities form the top layer of a business architecture model, and can then be broken down to business process and functions, systems and applications. The capabilities are visualized in a “City map”, a graph showing the most important elements in the architecture and why they exist by illustrating value streams (Akenine et al., 2014: 47). Next follows the *Business Process analysis*, and the city map is here used to map, home in and prioritize the most important activities, and to give an overview and context of the work. The city map is also used as input to an agile approach of development, providing a high-level roadmap that helps the agile organisation to prioritize, identify expected applications and the resources needed. This is a typical task for a Business architect with support from domain experts and the Records manager to focus on the information. Typical output is process charts, information models and a first sketch on services that can provide users with necessary information. A design process approach could be used to build early prototypes to verify the findings. This step is followed by the establishment of a business classification, which is performed by the Records manager with input from the previous work process analysis. The business processes should then be aligned with the relevant systems and organizational units, which is the primary domain of the Solution Architect in close collaboration with the Business Architect. This is one of the most crucial parts of any IT development project, bridging the gap between Business and IT (IASA, 2019). Business architecture shows *what* shall be done, while Solution architecture shows *how* it shall be done. The agile approach will here help by building work packages that supply development teams with both requirements and priorities for planning. Especially building and maintaining the metasystem is important and should be done using input from the Business architect. Here the Design process way of working contributes with prototypes in the form of templates and feedback from users, to quickly get services and solutions correct. The following step, the *appraisal* and assessment of information, is primarily RM work, but deliverables and support particularly from the Business architect is needed. *Documentation planning* is a mix of activities performed by the RM roles and Business architect using design process methods. The last step, *Archival description*, is a typical RM domain.

Concluding remarks

The premise of this paper is that RM should be a central feature of IG, and that RM requirements could apply to the management of other organizational information assets. The intermesh of RM principles and EA is a fruitful approach in the development of coherent IG regimes The aim of EA is to support business performance in order to fulfil the organizations' overall mission and enable change and development. A focus on supporting the fulfilment of organizational goals and missions, and the requirements from other stakeholders, is also characterizing RM. EA and RM thus meet in a common mission, aiming at the same basic goal. Modern RM is also concerned with change and continuous improvement (e.g. ISO 30300:2011, clause 2.4.8) and service development, an essential feature of EA. However, the object of EA is the management of the organization as a whole and its assets, while the object of RM is the management of the information assets regarded as records. RM has thus a different mandate, but constitutes an indispensable element if the organization's goals should be fully realized. RM requirements need to be embedded in the domains of EA and the phases of the development of an architecture (ISO/TR 21965:219(E), clause 12.1). RM, on the other hand, needs to adopt methods and tools already developed in EA, which would enhance professional knowledge and performance. The suggested approach is therefore to merge RM work with architecture work, identify roles that work together, and align methods and corresponding deliverables. Of central importance in this process is the Business architect role, which spans over or has connections to almost every other role. The collaboration should start here. An example is business process analysis, which is main concern for contemporary RM, but also the foundation of the Business Architecture. The Records manager and the Business architect should work in close collaboration performing analyses and producing deliverables as city maps, process maps etc., in conjunction with UX experts that drives the process in a design thinking way, and by capturing input from domain experts. EA and RM thus co-function very well in this particular area, but RM also interfaces with other architect roles.

The paper reports the current results of an on-going development project. The next phase will be testing and evaluating the work model. Further R&D activities could include a systematic mapping of the elements of RM principles and different EA frameworks in whole.

References

Akenine, D., Toftefors, J., Berg, C., Kammerfors, E., Folkesson, R., Olsson, S. H. (2014). Boken om IT-arkitektur. Helsingborg: HOI Förlag

An, X. (2009). The electronic records management in e-government strategy: case studies and the implications. // 2009 International Conference on Networking and Digital Society 1, 17-20

Banfield, R., Lombardo, C. T., Wax, T. (2015). Design sprint: A practical guidebook for building great digital products. O'Reilly Media, Inc.

Becker, C., Antunes, G., Barateiro, J., Vieira, R., Borbinha, J. (2011). Modeling digital preservation capabilities in enterprise architecture. // Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times. ACM, 84-93

Bergbom, R., Forsberg, J. E.. Hagstedt, R., Sahlén, T. (1992). Den goda visionen: om kontorsinformationssystem i offentlig miljö. Älvsjö: Kommentus

Bodin, S., Sahlén, T., Sjögren, C. (2000). Dokumenthantering i företag och organisationer: en kvalitetsfråga. Stockholm: Folkrörelsernas arkivförbund

Brooks, J. (2019). Perspectives on the relationship between records management and information governance. // Records Management Journal 29, 1-2, 5-17

COBIT 5: A Business Framework for the Governance and Management of Enterprise IT. ISACA. https://www.isaca.org/COBIT/Pages/COBIT-5.aspx (30.8.2019)

Cross, N. (2011). Design thinking: understanding how designers think and work. Oxford-New York: Berg

Denscombe, M. (2014). The good research guide: for small-scale social research projects, 5th ed. Berkshire: McGraw-Hill Education

Fowler, M., Highsmith, J. (2011). The agile manifesto. // Software Development 9, 8, 28-35

Gartner Inc. (2019). https://www.gartner.com/it-glossary/information-governance/ (30.8.2019)

Generally Accepted Recordkeeping Principles®: Information Governance Maturity Model. Overland Park, KS: ARMA International, 2013 https://rim.ucsc.edu/management/images/ThePrinciplesMaturityModel.pdf (30.8.2019)

Hagmann, J. (2013). Information governance–beyond the buzz. // Records Management Journal 23, 3, 228-240

Hassenzahl, M., Tractinsky, N. (2006). User experience-a research agenda. // Behaviour & information technology 25, 2, 91-97

IASA Global. (2019). https://iasaglobal.org/itabok-1/engagement-model/role-descriptions/ (30.8.2019)

International Council on Archives (2007). ISDF - International Standard for Describing Functions

ISACA. (2019). https://www.isaca.org/Pages/Glossary.aspx (30.8.2019)

ISO 15489-1:2016 Information and documentation — Records management — Part 1: Concepts and principles

ISO 30300:2011 Information and documentation – Management systems for records – Fundamentals and vocabulary

ISO 30301:2011 Information and documentation — Management systems for records — Requirements

ISO/IEC 38500:2015 Information technology — Governance of IT for the organization

ISO/IEC/IEEE 42010: 2011(E) Systems and software engineering - Architecture description

ISO/TR 21965:2019(E) Information and documentation — Records management in enterprise architecture

Katuu, S. (2018). Using Enterprise Architecture as a means of understanding institution technology ecosystems. // Proceedings of the ICMLG 2018 16th International Conference on Management Leadership and Governance. Academic Conferences and Publishing International, 130-138

Katuu, S. (2018). The Utility of Enterprise Architecture to Records and Archives Specialists. // 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2702-2710

Katuu, S., Mpho N. (2015). Managing digital records in a South African public sector institution. // INFuture 2015: e-Institutions Openness, Accessibility. and Preservation, 135-141

Kotusev, S. (2016). The history of enterprise architecture: An evidence-based review. // Journal of Enterprise Architecture 12, 1, 29-37

Kotusev, S. (2019). Enterprise architecture and enterprise architecture artifacts: Questioning the old concept in light of new findings. // Journal of Information Technology 34, 2, 102-128

Lomas, E. (2010). Information governance: information security and access within a UK context. // Records Management Journal 20, 2, 182-198

Reed, B. (2017). Recordkeeping in Information Governance. Information Governance ANZ. https://www.infogovanz.com/recordkeeping-in-information-governance (30.8.2019)

Saffady, W. (2015). Records management or information governance? // Information Management 49, 4, 38-41

Sahlén, T. (2016). Informationsforvaltning - i offentlig och privat sektor. Stockholm: Näringslivets Arkivråd

Sessions, R. (2007). A comparison of the top four enterprise-architecture methodologies. Houston: ObjectWatch Inc. http://www3.cis.gsu.edu/dtruex/courses/CIS8090/2013Articles/A%20Comparison%20of%20the%20Top%20Four%20Enterprise-Architecture%20Methodologies.html (30.8.2019)

Sprehe, J. T. (2005). The positive benefits of electronic records management in the context of enterprise content management. // Government Information Quarterly 22, 2, 297-303

Sundqvist, A. (2005). (ed.). Dokumentstyrning i processorienterade organisationer. Stockholm: Folkrörelsernas arkivförbund & Näringslivets arkivråd

Svärd, P. (2013). Enterprise Content Management and the Records Continuum Model as strategies for long-term preservation of digital information. // Records Management Journal 23, 3, 159-176.

TOGAF® Standard. (2018). Version 9.2, C182. The Open Group. https://pubs.opengroup.org/architecture/togaf9-doc/arch/index.html (30.8.2019)

Vieira, R., Valdez, F., Borbinha, J. (2011). An Analysis of MoReq2010 from the Perspective of TOGAF. // International Conference on ENTERprise Information Systems. Springer, Berlin, Heidelberg, 335-344

Vieira, R., Borbinha, J., Valdez, F., Vasconcelos, A. (2011). A reference architecture for records management. // Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times. ACM, 339-340

Zachman, J. A. (1982). Business systems planning and business information control study: a comparison. // IBM Systems Journal 21, 1, 31-53

Zachman, J. A. (1987). A framework for information systems architecture. // IBM systems journal 26, 3, 276-292

https://doi.org/10.17234/INFUTURE.2019.11 Review paper

Digital Preservation of Electronic Records in the Croatian State Archives, U.S. National Archives and Records Administration and Library and Archives Canada
The Importance of Education of Information Specialists

Lana Žaja
Croatian State Archives, Zagreb, Croatia
lanazaja@gmail.com

Summary
The aim of this paper is to give insight, by a comparative analysis, of the publicly available documentation on the digital preservation of records in the Croatian State Archives, U.S. National Archives and Records Administration and Library and Archives Canada, and through the process of establishment of digital archives, point to the necessity of scientific and operational knowledge on information technologies and digital preservation. Methodologically presented activities and projects for the development of tools for digital preservation of records in digital archives, point out, above all, the potential education of archivists, who must be able to provide quality in-personal and on-line information service. Planning, strategy and the establishment of a digital archive are processes that, besides being organized and systematically implemented, require highly educated experts from multiple scientific areas. The use of international standards for the long-term data preservation that need to be secured within the defined guidelines and measures for the preservation process plan, which are to follow all components of the electronic file processing system in the archives, is presented. The directives on solving problems related to long-term storage and access to electronic archive material in today's environment should also be in accordance with the terms set by the legal frameworks. Given the importance of this issue, as well as its goals, the role of digital archives within state administration, e-government development and online access to information in databases, is crucial for scientific, operational and functional management of the system by experts who know how to ensure plan for development of policy on institution documents management.

Key words: long-term digital preservation, digital records, digital archives, international standards, archivist education, information specialists

Introduction

The introductory part should answer two key questions, namely: what exactly does the term 'long-lasting digital archiving' mean and how does this term really differ from a regular backup of a document? The term 'archiving' nowadays covers several meanings in terms of the concept of sustainable digital archiving. The premise for what is not a long-term digital archiving, in the IT context, includes the following:

Long-lasting digital archiving is not a backup (it's not just about preserving the important file flow)

Long-lasting digital archiving is not a hierarchical storage manager (migrating files to some of the formats in order to make place on the hard disk)

Long-lasting digital archiving is also not the last stage of data storage before oblivion or permanent loss of records.

Permanent archiving of a digital document implies a mandatory digital preservation plan that describes in detail the process of long-term preservation and preservation of digital data and metadata. The establishment of such a plan is primarily used to establish data sustainability and assess the reliability of techniques and tools to preserve a virtual collection. Maintaining access to data in its original form is certainly the cornerstone of digital preservation plans due to the constant evolution of technology. Computer hardware and software is slowly getting outdated and it is important to consider not only data retention, but also the preservation of their “support”.[1] The digital preservation plan consists of a series of actions to be carried out by the institution responsible for digital objects. The conservation plan takes into account the policy of conservation, legal obligations, technical and organizational constraints, user needs, conservation objectives, conservation context, assessment of conservation strategies and its outcomes, including the logical decision-making process. It also determines every action, along with the responsibilities, rules and conditions for their performance on the collection (conservation action plan). In case that this action, their implementation and their technical environment permit it, the action plan must be implemented on a continuous basis. Each conservation plan must include the following elements[2]:

- identification which encompasses the definition of the plan and preparation of the project documentation
- the status of the plan (active, on hold, etc.)
- a description of institutional parameters
- description of the collection
- requirements for conservation management (budget, technical limitations, etc.)
- the cost of plan implementation
- the role and responsibility of the members responsible for the plan
- conservation action plan (in order to determine “when”, “what”, “where” and “how” is most commonly defined in the project itself).

There are several digital preservation rules and policies that can help institutions regulate and develop conservation strategies. Digital preservation policies are very conceptual, but they provide a practical guide in realization of conservation. The key to policy is, for example, the maintenance of manoeuvres that must be done to establish and enable free access. Besides, copyright compliance is an important policy in the implementation of the digital preservation plan. Digitized or born-digital material must be subjected to an assessment of the legal rights of reproduction and the law on secrecy. Although in many institutions, the development of preservation plan is often “ad hoc”, it is possible to support its development by means of tools such as standards and policies defined by the international reference offices entities. By adopting a well-documented and standardized planning process, the durability of digital objects greatly improves. The fragility of digital objects lies in the fact that their IT environments are constantly changing, which is a problem in some of the preserved activities, for instance, in case of data migration. Although there are many tools for converting the object’s format, it still possible that the initial appearance will change, for example, footnotes or hyperlinks could disappear. It is also possible to create an original environment of the object by emulation (process of a computer, device or program that mimics the function of another computer, device or program), but this method remains limited because it is difficult to extend it to a large number of digital objects. Moreover, this method is successful only during migration, but some object specificities may be lost due to inaccurate or incomplete mimicking competition, as well as the inability to restore essential aspects of the object by emulation. The main challenge is not only to preserve the data, but also to maintain access to that data in such a way as to preserve authenticity. Preserving object’s authenticity is therefore a key issue in the development of the digital preservation program. That is way documenting the process of preservation is necessary, while transparency ensures the reliability of the contents of digital collections.

Croatian State Archives

The strategic plan from 2016. to 2018. encompasses the objective of securing the conditions for long-term preservation and usability of the record of the Croatian State Archives.[3] The priority of this objective was to build the infrastructure and process of digital archive with relevant norms and existing practices, and the establishment of e-archives for the enforcement of legal regulations in relation to electronic records. For the aforementioned period, it was planned to prepare complete

[1] National Library of Australia. Guidelines for the Conservation of Digital Heritage. 2003. http://unesdoc.unesco.org/images/0013/001300/130071e.pdf. (20.7.5019.)

[2] Becker Hannes C., Guttenbrunner K.M., Strodl S. Systematic Planning for Digital Preservation: Assessment of Possible Strategies and Preservation of Building Plans. International Journal on Digital Libraries. 2009.

[3] Croatian State Archives. http://www.arhiv.hr. (25.7.2019.)

project documentation and to provide sources for funding the construction of the digital archive structure. Chapter 5 of the strategic program of scientific research from 2015. to 2020.[4], in the planned topics of research under the title of the topics "Archives in the Digital Societies", includes the planning of the preservation of digital records. This research area includes several narrowly defined topics or subject-related researches related to managing electronic records, either with digitalisation of archive services or operation of archives in the communication space that is opened by digital technologies. There are two main research goals in this area. The first one is the upgrade or the ability of archives and other organizations to long-term protect of electronic documentation and to preserve its integrity, credibility and usability, and the second one is to improve methodologies and tools for designing and communicating archival content and services. The research is focused on the question of metadata scheme, development of catalogue rules and ontology in management of digital information resources. The archive has, together with main institutions in the field of museum and library activities and faculties which host an information science study, concluded an agreement on the development of new cataloguing rules, which should be based on harmonized or integrated concepts of presentation, structuring, description and search of information sources. The research involves analysing and evaluating different conceptual models and ontologies that are being offered today, with emphasis on those that are more present in the archives, libraries and museums. It also cover the issue of their interoperability, relationship to linked data and open data principles, as well as issues of standardized terminology and normative control and questions related to visualization of data. The second topic of research in this area is long-term preservation of integrity, credibility and usability of documentation in electronic form. The archive is part of one of the leading international projects in this are InterPARES[5] Trust that is led by the University of British Columbia, within which technology, legal, business and social aspects are explored, as well as risks that can affect long-term preservation of electronic documents. Given that the long-term preservation of electronic documents largely depends on the characteristics of the information systems in which such documents arrive, the research is also focused on the issue of norms and procedures in the administration and the way in which they are implemented in information systems, that are used in business, with an emphasis on administration and public services. A particular goal in this segment of the research is the development of recommendations or guidelines for the revision of regulations and standards that currently regulate this area in Republic of Croatia and recommendations for the implementation of measures and activities relevant to the long-term preservation of integrity and credibility of electronic documents. The results of this research should enhance the ability of archives and creators of archival documents for long-term protection and preservation of the records in electronic form. Improving the quality of online catalogues and digital collections should facilitate their publication on relevant international portals. Contribution to the introduction of more modern norms and procedures in the office business of public service of the Republic of Croatia is expected. The importance of the strategic scientific research plan and its program has its main foundations in its own development by experts in the field of digital archives. Structuring this kind of program at the level of information studies leads to the creation of quality programs, the provision of information in real-time and the establishment of the necessary tools for managing the requirements of original digital records.

Library and Archives Canada

Library and Archives Canada[6] (hereinafter referred to as LAC) collects, manages, maintains and provides permanent access to Canadian documentary heritage, while also serving as permanent repository of Canadian government records as well as publications and records of historical or archival value. LAC is the only organisation in the Canadian government with a national mandate for long-term preservation of records. This mandate is contained in the LAC legislation which places librarians and archivists in a status where they must have a knowledge o long-term preservation and disposal of documentary heritage. The LAC library and archives contain a wide range of textual,

[4] Strategic program of scientific research from 2015 to 2020. http://www.arhiv.hr/Portals/0/Dokumenti/Planovi%20i%20izvje%C5%A1%C4%87a/Strate%C5%A1ki%20program%20znanstvenih%20istra%C5%BEivanja%202015.-2020.pdf?ver=2017-07-31-133703-747 (25.7.2018.)

[5] InterPARES Trust. https://interparestrust.org/ (23.4.2019.)

[6] Library and Archives Canada. http://www.bac-lac.gc.ca (6.8.2019.)

visual, audio-visual and web content that support various software, hardware and operating systems in a multitude of formats that are sensitive to technological obsolescence, media degradation and record decay in general. The LAC's digital archive at the "Conservation Centre" serves as a storage for permanent digital collection of LAC. "Conservation Centre" contains only a part of digital collections that LAC has acquired over the years. Much bigger task is securing digital collections of documentary heritage, which should be preserved in a way that meets the LAC's mandate. In its "Strategy for program of digital reservation"[7], LAC describes in three phases, a vision of digital document preservation:

Phase 1: Collecting information and factors of success of phase 1, Phase 2: Program development and factors of success of phase 2, Phase 3: Program implementation and factors of success of phase 3.

Digital preservation is defined as "active management of digital content over time, in order to ensure permanent access". If this approach should be enabled, professional digital preservation staff must proactively monitor the information system and respond according to requests, in order to protect the digital heritage contact from technological obsolescence.

LAC is in compliance with the international standard ISO 14721: 2012, Reference model for the Open archive information system. OAIS reference model[8] describes the functions and roles of the digital archive and helps to define the key elements that are required in the description of the digital preservation program site. LAC uses an OAIS reference model for defining the scope of its digital preservation program. Based on the definition of a model of lifecycle "Digital conditioning centre", digital preservation in LAC is part of a wider set of functions in the digital lifecycle of a record. The functions that are addressed in the DCC model are: downloading, preserving, storing and transforming. Although digital preservation is not directly responsible for the full digital lifecycle of a record, LAC provides guidance and advices for ensuring the integration of different components. The vision of digital preservation is that, within the end of 2024, LAC has a program of sustainable digital preservation that is in line with the international standard ISO 16363[9]: revision and certification of reliable digital reserves. ISO 16363 is an international standard based on ISO 14721[10]. This is actually a checklist of delivered programs for digital preservation that provides a number of criteria for organizations that must meet these criteria in order to implement a reliable digital storage. Key elements in developing digital preservation program are defined in these international standards. To achieve this vision, LAC works on implementing sustainable technical solutions, such as the establishment of a collection management framework, in order to ensure systematic storage of digital collections. By shaping the future of digital preservation in LAC, this strategy provides a framework for managing and coordination of activities that are needed by organizations for achieving advanced maturity levels of digital information preservation systems. Each phase is actually a strategy divided into a set of interrelated results with multiple executive checkpoints for confirming the correctness of access. Phase 1 "Information collection" is based on the program platform, developed by the developer in cooperation with information specialists, as the basis for the overall development of the information system. The main activities are researching, identifying and documenting business needs and digital preservation issues. These activities are described as a fundamental exercise in defining the problem and specifying it. The results of the first phase identify the magnitude and scope of the digital preservation challenges of LAC, they define the need for technology and human resources and they provide a final analysis of non-compliance documentation to meet the stringed requirements of ISO 16363. The digital preservation business requirements, which were completed in June 2017, were used as a starting point for discussion and results, for streamlining business processes and developing business technology solutions to support the long-term preservation of digital funds. This document provides information on the size and scope of the required technology and it serves as a catalyst for the evolution of software solutions. It also provides managers with information that should help them to make investment decisions on infrastructure digital protection solutions. As part of the inventory of

[7] Strategy for a digital preservation program. http://www.bac-lac.gc.ca/eng/aboutus/publications/Pages/strategy-digital-preservation-program.aspx (10.6.2019.)

[8] Open Archival Information System OAIS. https://www.oclc.org/research/publications/library/2000/lavoie-oais.html (23.4.2019.)

[9] ISO 16363. https://www.iso.org/standard/56510.html (23.4.2019.)

[10] ISO 14721. https://www.iso.org/standard/57284.html (23.4.2019.)

digital collections, a gathering of collections profiles was created to better understand digital collections that LAC has acquired in the 1970s. These profiles allow the LAC to: quantify digital content that includes backdrops of record collecting, determine who is responsible for them, detect where they are located, determine their file formats and other collection characteristics, determine whether or not they are conserved and evaluate the risk of their permanent loss. These collection profiles describe the scope of work and the profile of professionals who need to be employed in the preservation of these digital assets, helping to guide future conservation planning. During this phase an analysis of political “gaps” is also being carried out, which means that future operational plans, procedures and priority practices for further development in the future should be identified. This analysis is managed by the LAC in establishing a formal policy framework to preserve Canadian digital documentary heritage. One of the key outcomes of Phase 1 is the development of guidelines that provide guidance on future development, as well as on implementation of the program. With describing the planned steps for building digital preservation capacity, this paper sets out the main planning activities to be fulfilled, the roles and responsibilities of information specialists, the sequence of their implementation and the overall time frame. The guidelines range from gathering information, technical solutions, managing digital collections, practices, plans and operational policies to the decision point at the end of Phase 1. The purpose of the aforementioned decision is to identify the features required by the information system in determining the size and scope of the problem. Actions and deliveries of this phase correspond to the requirements of digital preservation and analysis capabilities. The program development follows in Phase 2, which aims to develop additional precision on the estimates obtained in Phase 1, on the design of conservation programs. Until beginning of Phase 2, the size and scope of the problem should be understandable and clear, are the next steps are the design and delivery of digital preservation program capabilities. Technical solutions, based on the requirements for digital preservation completed in Phase 1, which have established the appropriate technology and infrastructure in the solution of digital preservation, in Phase 2 focus on developing analysis and design solutions. It also includes aspects of technical design for infrastructure of records storage, networks, platforms, applications, as well as interoperability and integration with other LAC information systems. The financial cost and the invested human knowledge, which is involved in this new business solution, are analysed. Ultimately, the options for managing the final decisions are presented. Based on inventory of digital collections, LAC in this second phase develops a digital collections management plan to allocate priority to digital collection formats, requirements for their conservation and the level of investment needed for each collection. One of the key activities of this phase is the implementation of more detailed sampling of collections for the assessment of corruption and data loss. It also develops technical specifications and standards for preserving different types of content. The plan of phase development practice, procedures and rules is issued in Phase 2 on the basis of analysis of a “difference” that is completed in Phase 1. By analysing all these decisions, a policy that has implications on price can be identified, such as the number of copies, storage and diversity of geographical locations. Workflow and procedures are further defined to allow unobtrusive “track” of records from receipt to retention and retrieval. Factor of success of Phase 2 is achieving clarity in all its program directions. LAC defines what makes scalable services and infrastructure for the program. It sets the level of investment in the required technology, infrastructure and resources, as well as cost models. The knowledge base on the scope of digital content that needs to be preserved, the need for required staff, the billing priorities, the preservation of new collections, the cost of delivering the program and the appropriately allocated budget levels are implemented in this Phase 2. By summarizing the above mentioned, the following can be concluded: steps in Phase 2 are: development of programs, technical solutions, management of digital collections, practices, plans and operational policies and decision-making at the end of Phase 2. The purpose is to develop potential technological solutions based on business requirements developed in Phase 1, as well as to provide evidence on the stated concept and to evaluate the required technological investments. Actions and deliveries depend on technical analysis and design of technical solutions for defining appropriate hardware and software technologies and all other components required for long term preservation. Phase 3 is the implementation of the program, in which the LAC is fully ready to launch program implementation because it has a defined vision, established interinstitutional obligations, management support, organizational capacity, consistent workflow and approved funding for the program. By the end of the Phase 3, LAC achieves its goal of a sustainable digital preservation program and a reliable digital repository in line with national and international best practices and standards. Phase 3 success factors are project completion of technical infrastructure, with well-equipped and fully formed business information, with which the design will be implemented. The result of all of the above-mentioned phases is a fully functional technological solution for supporting digital preservation. Based on the decision made in the database management plan, in Phase 2 LAC is equipped with the necessary tools and decision-making infrastructures in order to perform priorities of digital preservation of all collections. The collection management document must be refreshed regularly, as other new and old collections arrive. Given that digital collections management is implemented, a collection management plan becomes a key part of conservation planning and administrative workflow. The LAC implements ISO 16363, auditing and certification of trusted digital repositories, in order to evaluate the progress of its program implementation. During this revision, the approach that should satisfy audit and certification organization team is determined. Due to the large amount of digital collections, at the level of the required documentation, the LAC must fulfil the formal certificate with the auditor from the “third” (neutral) side, who is accredited according to ISO 16919[11]. Other approaches may include informal certification by re-self-evaluation. Comprehensive Phase 3 steps can be said to be: implementation of programs, technical solutions, digital database management, practices, plans and operational policies. The ultimate purpose of this Phase is to implement business solutions. Action strategies and deliverables provide a digital preservation solution, as well as the implementation of a scalable digital preservation solution. The digital preservation program strategy is regularly refreshed, when delivering each of the specified phases that are completed, or when implementing a digital preservation program that is integrated within the organization. LAC’s strategic approach to digital preservation is a dynamic development program that is in the process of maturation. Program plans and conservation services are regularly reviewed and updated to continuously align with their organizational mandate and strategic goals. Permanent improvement of the services for solving inherited, new and original digital content is part of the organizational planning, streamlined according to technical and other models of service provision in the digital ecosystem. LAC on its website points out that, with this investment strategy, it is most important to show stability of purpose and professionally trained commitments, with its obligations of digital preservation in order to ensure the survival of the national documentary heritage. The consequences and costs of human resources in terms of inaction, would be a failure in nonfulfillment of their statutory mandate, as well as obligation towards Canada’s citizens, and ultimately, in the worst case, the loss of national content of digital heritage.

U.S. National Archives and Records Administration

U.S. National Archives and Records Administration[12] (hereinafter referred to as NARA) identifies, stores and provides access to large number of archive materials of the US government. NARA takes care of preserving these records in order to protect civil rights, ensure government accountability and documents. The archive fund covers more than 13 billion pages of unique documents, electronic materials, maps, charts, photography, artefacts, as well as film, sound and video records. Records stored by NARA belong to the public, and the main mission is to focus on openness to the public, educate the public in participating in various digital programs, and strengthen national democracy through public open access to all state records. The preservation of NARA’s digital records, including copies for public use and digital surrogates, was created through the project scope which conducts digitalization procedures. NARA advocates the preservation and maintenance of access to the contents of all original digital records and digital surrogates that are in its possession and for which archivists determine that they contain sufficient historical or other values that ensure continued conservation by the US Government. The digital preservation implies continuous use of records, which is considered essential and viable for the purpose of creating digital materials. NARA uses several key strategies to enable the effective preservation of digital content, knowing that these strategies have to be flexible and adapted to current changes in benchmarks, technology and standards. The main goal of all strategies is to reduce risk and achieve best practice for digital preservation and maintenance of access to digital content. The first strategy involves combining

[11] ISO 16919. https://www.iso.org/standard/57950.html (24.4.2019.)

[12] National Archives and Records Administration of the United State of America. https://www.archives.gov/. (6.8.2019.)

documentation of standards and procedures. By documenting internal standards for the creation of digital surrogates, guidelines for the creation of digital surrogate agencies, as well as the minimum metadata and the preferred file formats for electronic records transmitted to NARA are enabled. By encouraging the use of open standards and accepted on consensus of information and digital professionals, the future access and preservation of the entire lifecycle of digital preservation of records are made easier. The second strategy sets priorities and is focused on approaching the risks of setting priorities for digital protection and performing the schedule of default digital activities. The third strategy manages files that store digital content in a trusted repository of digital objects, thus allowing continuous management, as well as accessing content over the entire lifecycle of the record. NARA's repository is based (as is the Canadian archive, as previously described) on the Open Archival Information System Reference Model (OAIS): ISO 14721: 2012: A reliable digital repository whose main mission is to provide a reliable, long-term approach to managed digital resources to their separate community, now and in the future (OCLC, Trusted Digital Repositories: Attributes and Responsibilities, 2002). NARA reduces the number of file formats that need to be actively managed, by standardizing files in selected formats, while retaining authentic features of original format. The fourth strategy is the importance of authenticity, which refers to the credibility of record as an accurate representation of the original. Insurance of the authenticity is documented according to OAIS model. The fifth strategy provides guidelines for metadata of conservation. By joining permanent digital identifiers and capturing metadata about preserving each digital object, data is stored as computer files in application software that allows searching. This software also helps to keep digital records over time, with manual (semiautomatic) and automatic preservation procedures. Preserving metadata ensured the preservation of key contextual, administrative, descriptive and technical information, along with a digital object. The sixth strategy describes organizational relationships that emphasize the importance of active participation in local, national and international digital preservation meetings in order to exchange information and experience in seeking further guidance and establishing cooperation in addressing digital protection issues. This engagement helps identify new risks, practices, as well as a description of new standards that continually strive to improve the digital archives program. By engaging information professionals who know how to manage information technologies, understanding of the needs of digital preservation, development of new technical tools and program-information systems are secured. Digital conservation activities should continuously be the subject of ongoing evaluation, using appropriate audit-based assessment tools, certificates and/or national standards for conservation of digital governance, which measure the ability, stability and maturity of digital preservation programs. Stable digital preservation is achieved through the digital protection infrastructure, which ensures data integrity, formats and their sustainability and information security. Digital conservation infrastructure such as hardware, software, networks, storage, associated equipment with related development tools, development tools, testing, management, control, management and support for information technology are important constituents of digital preservation. Warehousing or storage of data is one of the most important steps in determining the system's network capabilities and tools for creating, processing and managing active files of original digital files and digital surrogates. The next step is to process regular professional supervision, to update the entire system and tools that NARA develops for the purpose of developing its own business processes, and which need to be accessible for the storage of the contents of original digital files and digital surrogates. These replications include one storage copy in another environment, preferably in a remote geographical location that includes replications that can be provided through an e-service "NARA Cloud". The tools[13] for determining forensic identification and formatting of records include the identification of file formats (identification of technical files), validation of the format (confirmation of the file compliance with documented format specifications) and extraction of technical metadata (documentation on file creation, including applications and operating systems) used for support in assessing the risk of outdated format and publishing files to users using the appropriate application or browser in the default context. Tools for transformation of file formats serve to execute migration of files over time, as formats become obsolete and therefore risky. Standardized workflow processes for linking original digital and digital surrogate files with

[13] NARA's Adoption and Management of Cloud Computing. https://www.archives.gov/files/oig/pdf/audit-report-17-08.pdf (20.10.2019.)

record identifiers and metadata, provide primarily appropriate IT storage spaces and cloud-based access servers. Data integrity is based on records of all incoming files with degraded events, as well as all subsequent life cycles, such as transformations of the format of moving files with different revisions. Downloading a file is a process that must involve scanning "malicious" software and checking the file's validity. Checking the validation of files refers to checking in which the file has not been changed from the previous state. Copying content from a physical medium must include the use of "blocker", i.e. devices that prevent accidental damage of content on physical media. Furthermore, it is necessary to perform an annual check of all samples of digital electronic records and digital surrogates stored in the storages, including checking if they are not supported, in order to correct and/or replace the errors with other files as fast as possible. It is necessary to implement a retrograde quarterly revision of the log, in order to verify that the files in the repositories have remained unchanged and have not been corrected over time. Ultimately, it is necessary to carry out an annual check of all media, which in its workflow contains permanent records on permanent storage. Before the media (which contain permanent records) "reach" ten years, they are copied to tested and checked new electronic media. Format and media sustainability are the main features for checking formats and files at the download site. Characterization refers to the identification and description of the technical characteristics such as the original file environment. Technical metadata is recorded. Validation refers to the confirmation that the file should match the expected characteristic in that type of record. By creating action plans for file formats that identify formats of other files, the necessary action is enabled, for those formats that are no longer viable or not available through the current software. Then, it is necessary to create standardized versions of the files that are not in a risky format, as it is defined with plan in the previous action strategies for file formats. Standardization is converting all files into certain types of file formats, as for example, e-mail, jpeg, which will be sustainable over longer period of time. By analysing the file formats and media formats, potential future obsolescence can be permanently determined. By performing migration of automatic and manual formats or other conservation activities, based on the file formats action plans, systematic monitoring of large digital conservation files in technological environment is set up. Action plans include warning signs for formats, media and/or equipment that become potentially outdated and are no longer viable. Information security should allow physical access to: media, information systems, and data entry and processing services, which include reading, writing, and authorization in folders and files on different servers. The importance of insurance is that no one can access all the files. A file action system log must also be maintained, including deleting and modifying actions. From these procedures and guidelines, it is evident that there are many key factors that enable and contribute to the ultimate success of this digital protection strategy. Certainly, the critical factors that NARA has to deal with to fulfil its goals, which is education and the level of education of professional staff, must be highlighted. With all these strategies, NARA confirms that the digital preservation is a significant business management process that goes through several layers of business units requiring business and professional excellence. A special human resources development plan has been designed to support all of the above-mentioned functions, which are the man backbone and infrastructure of all of the above-mentioned information technologies. The planning process is crucial here because it identifies the needs of the infrastructure to support digital preservation which includes information systems, tools, storage, network capacity, data integrity and safety of the information system. By documenting all of the mentioned relevant management processes, in the context of the management, including those for predicting record storage, network capability, planning and implementation of additional capacities and refreshment of technology, the main guidelines for creators and record holders are given. NARA continues to develop and publish guidelines for format and metadata standards, in order to ensure the sustainability of original digital data and digital surrogates. The document "Preservation of electronic records: Strategy for the preservation of digital archive materials"[14] was published on 8 June 2017.

[14] Preserving-Electronic-Records-History. https://www.archives.gov/files/Preserving-Electronic-RecordsHistory.pdf (6.8.2019.)

Conclusion

Digital preservation of electronic records for contemporary archivists and digital archives brings changes on multiple levels, and the differences are visible not only to the degree of development of information systems and financial possibilities, but also to the level of education of professionals who manage complex digital systems. Using standards in digital preservation helps archives to create information systems and services that are safe and reliable. It is transparent that standards help archives to increase productivity, while reducing and eliminating errors in digital archives systems. Also, they allow direct alignment of data with metadata of different types of records. Standards ensure the protection of end-users of products and services, thus allowing certified products to meet the minimal criteria of standard set at the international level. In the digital archives of the USA and Canada, active engagement in the development of e-government has been evident, and the archival strategic role is becoming increasingly important. Croatian State Archives should, through networking with state-level administrations, introduce new additional favourable digital services for private and legal persons, as well as to increase the interest in adopting new continuous knowledge in the field of digital technology development. According to the LAC's definition, digital conservation involves "actively managing digital content over time in order to ensure consistent access to records". From this definition, it can be concluded that in the organization of the archive all employees should be proactively and continuously educated, and within the archive there should be special professional services that monitor and, if necessary, intervene for 24 hours a day at all levels of the system, in seeking to protect the digital heritage content from technological obsolescence and/or accidental loss of records. By customizing traditional archival practices in the management of electronic records and manuscripts, information professionals need to be able to select the appropriate tools to use in analysis and processing of digital records and manuscripts. Creative design and development of work processes for accessing and processing digital records and manuscripts requires from repository managers and archivists to have professional expertise and scientific responsibility for the arrangement, description and availability of electronic records. The role of technological development dictates the gathering of world experts from different fields of knowledge in addressing for digital archives to keep pace with the numerous developmental reforms brought by the technological future.

References:

Bralić, V., Kuleš, M., Stančić, H. (2017). A model for long-term preservation of digital signature validity: TrustChain // INFuture2017 Proceedings: The Future of Information Sciences / Atanassova, I., Zaghouani, W., Kragić, B., Aas, K., Stančić, H., Seljan, S. (eds.). Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia, 89-103

Croatian State Archives. http://www.arhiv.hr. (25.7.2019.)

Duranti, L. (2000). Arhivski zapisi: teorija i praksa. Zagreb: Hrvatski državni arhiv. International Organization for Standardization (ISO). https://www.iso.org/standard (24.7.2019.)

ISAAR. (2011). ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families, 2nd Edition. (2011). http://www.arhiv.hr/Portals/0/ISAAR%28CPF%29_2_Izd_Hrv_1.pdf (1.8.2019.)

ISAD(G). (2011). General International Standard Archival Description - Second edition. http://www.arhiv.hr/Portals/0/ISAD_%28G%29_2_Izd_Hrv.pdf (1.8.2019.)

ISO 14721, ISO 16363, ISO 16919. http://www.iso16363.org/ (19.10.2019.)

Library and Archives Canada. http://www.bac-lac.gc.ca (6.8.2019.)

National Archives and Records Administration of the United State of America. https://www.archives.gov/ (6.8.2018.)

PREMIS Preservation Metadata Maintenance Activity. US Library of Congress. Retrieved. (2013). https://www.loc.gov/standards/premis/ (2.8.2019.)

Stančić, H. (2017). Obrazovanje arhivista // Arhivi u Hrvatskoj - (retro)perspektiva / Babić, Silvija (ed.). Zadar: Hrvatsko arhivističko društvo, 37-49

Stančić, H., Rajh, A., Brzica, H. (2015). Archival Cloud Services: Portability, Continuity, and Sustainability Aspects of Long-term Preservation of Electronically Signed Records = Les services d'archivage dans un nuage informatique: Portabilité, continuité et durabilité: Aspects de la conservation à long terme des documents signés électroniquement. // Canadian journal of information and library science 39, 2, 210-227

Stančić, H., Rajh, A., Jamić, M. (2017). Impact of ICT on Archival Practice from the 2000s Onwards and the Necessary Changes of Archival Science Curricula // Proceedings of the 40th Jubilee International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2017 / Biljanović, Petar (ed.). Rijeka: Croatian Society for Information and Communication Technology, Electronics and Microelectronics – MIPRO 918-923

Strategic program of scientific research from 2015 to 2020. http://www.arhiv.hr/Portals/0/Dokumenti/Planovi%20i%20izvje%C5%A1%C4%87a/Strate%C5%A1ki%20program%20znanstvenih%20istra%C5%BEivanja%202015.-2020.pdf?ver=2017-07-31-133703-747 (25.7.2019.)

Strategy for a digital preservation program. http://www.bac-lac.gc.ca/eng/aboutus/publications/Pages/strategy-digital-preservation-program.aspx (10.6.2019.)

https://doi.org/10.17234/INFUTURE.2019.12 Preliminary communication

Data Quality in the Context of Longitudinal Research Studies

Tonko Carić
Institute for Anthropological Research, Zagreb, Croatia
tcaric@inantro.hr

Kristina Kocijan
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
krkocijan@ffzg.hr

Summary

This paper discusses the concept of data quality in the context of longitudinal research. By deconstructing quality assurance process and data collection strategies through a case study of the „Croatian Birth Cohort Study", we try to define causes and sources of poor data quality in the context of longitudinal studies. Besides the problems discussed throughout the known literature (panel conditioning, sample attrition, recall bias, temporal and financial demands), we introduce single-source problems, multi-source problems, security problems, design questionnaire problems and QA workflow problems as important aspects in the domain of the possible sources of errors. Additionaly we propose models for eliminating the errors through prevention and detection in order to improve data quality

Key words: data quality, quality assurance, data collection, research data, longitudinal study

Introduction

Data may be defined as a representation of facts or concepts or instructions in a formalized manner, suitable for communication, interpretation or processing by manual or electronic means. Tayi & Ballou (1998) define data as a "raw material for the information age". An element of data is an item, idea, concept or raw fact (Abdelhak et al., 1996 as cited in World Health Organization, 2003).

Information is "useful data" that is processed by the end-user in such a way that the information received is manifested as "knowledge (McFadden et al., 1998). In the literature about research data, the term "information" is often used interchangeably with the term "data". In the context of research studies, data can also be referred as „population-based data" (Chen et al., 2014) and such data go through the processes of collection, storage, processing and compilation.

A longitudinal study is a research design that involves repeated observations of the same variables (e.g., people) over short or long periods of time (Young et al., 2007). Longitudinal studies share many similarities with transversal studies, while differences do exist. Key benefits of collecting data through longitudinal studies include analysis advantages and measures of stability or instability. Moreover, longitudinal surveys can help understand causality - only the longitudinal survey can provide information about cumulative phenomena, following changes over time in particular individuals within the cohort (Young et al., 2007).

This paper intends to define possible causes and sources of poor data quality particularly in the context of longitudinal studies. The structure of the paper will flow from the introduction to quantitative data collection, total survey error and data collection to the CRIBS project that was used as our case study, including the steps for its data collection and quality assurance, detecting sources of errors and eliminating the errors. The paper will conclude with some main discussion points.

Quantitative data collection

Quantitative data collection methods rely on random sampling and structured data collection instruments that fit diverse experiences into predetermined response categories. Primary longitudinal data can be collected by direct observation (e.g., interviews, field observation), survey (e.g., personal structured interview, mail questionnaires, telephone surveys, diaries), tests and instruments or retrospective measures (e.g., investigation of archived documentation, interviews) (Leedy, Ormrod, 2001). This paper puts focus on the usage of primary longitudinal data for longitudinal data collected through a survey.

There are several ways in which longitudinal surveys provide benefits in terms of data collection. These are mostly connected to either the quantity or quality of data that can be collected compared to alternatives such as the use of retrospective recall (Leedy, Ormrod, 2001). Longitudinal research can utilize either primary data or secondary data. With primary data collection, the principal investigator designs the measures and methods of data collection and supervises the data collection effort.

Total Survey Error perspective

Surveys are a common method in academic research to collect data. The Total Survey Error (hereafter TSE) approach has been established as a systematic framework to understand the various sources of error that are associated with each of these steps (Biemer, Lyberg, 2003; Jedinger et al., 2018). The term survey error refers to the deviation of an estimator from the true value in a population (Biemer, Lyberg, 2003).

According to Weisberg (2009), these potential errors can be divided into three categories of respondent selection (e.g., coverage error), response accuracy (e.g., item nonresponse error) and survey administration (e.g., mode effects). Current research that relies on the TSE approach, however, focuses on a narrow concept of survey data quality that involves errors that are induced by sampling, measurement and non-responses, but does not include other factors present while working with survey data.

Data quality – beyond the TSE

There is a need to analyse the quality of data (hereafter DQ) outside the TSE approach. However, there is no single definition of the quality in the context of research data accepted by researchers and those working in the discipline.

The World Health Organization (2003) defines research data quality as the ability to achieve desirable goals. Quality data represent what is intended or defined by their official source, are objective, unbiased and comply with known standards (Abdelhak et al., 1996 as cited by World Health Organization, 2003). Following on from this, World Health Organization (2017) has created a DQ review framework that, in addition to the known parameters from the TSE perspective, has introduced the following data quality indicators: bias and human errors in data entry and computation.

They also used the term “data quality dimension”. Why “data quality dimension”? According to the available literature, data quality in scientific studies is a multifaceted concept for which there is no precise or unique definition. One way of explaining DQ is through the concept of dimensions. Dimensions deconstruct data quality into practical, definable and measurable constructs (Tayi, Ballou, 1998; Bai et al., 2018).

Whitney et al. (1998) discuss data quality in longitudinal studies and emphasize the need for quality assurance and quality control procedures beyond the TSE approach. Quality assurance (hereafter QA) consists of activities undertaken before a data collection to ensure that the data are of the highest possible quality at the time of collection. Quality control takes place during and after data collection (Whitney et al., 1998).

QA is a process used to prevent problems in the data collection process and to support subsequent data quality. It plays an important role in the conduct of a research study by helping to ensure findings and conclusions are correct and justifiable (Yamanaka et al., 2016). According to the Szklo & Nieto (2014), QA activities before data collection aim to prevent or at least minimize systematic or random errors in collecting and analysing data. Traditionally, these activities have consisted of detailed protocol preparation, development of data collection instruments and procedures and their manuals of operation, and training and certification of staff. The development of manuals specifying quality control activities can also be considered as a quality assurance activity. QA therefore includes methods and procedures for preventing and correcting problems that may affect the quality of survey data (Biemer, Lyberg, 2003)

The available literature on QA focuses mostly on standardizing the protocols and personnel training (Sáez et al., 2012; Chen et al., 2014; Szklo, Nieto, 2014; Yamanaka et al., 2016). QA steps mentioned in research papers can be summarized into three steps: (1) developing a procedure manual for data collection, (2) developing a detailed recruitment and training plan to enforce the value of collecting accurate data and (3) monitoring and evaluating the process in the field and identifying areas of improvement to strengthen the study's protocol. However, the abovementioned steps lacked information science and computer science perspectives on data-related issues.

CRIBS case study

The project “Croatian Birth Cohort Study” of the Institute for Anthropology (hereinafter referred to as “CRIBS”) is a pilot of a longitudinal study aimed at the Croatian populations of the eastern Adriatic islands and the neighbouring mainland, in particular the population of pregnant women and their born children. It is a public health longitudinal study in which a sample of 500 pairs of mothers and their children will be examined, namely mothers' lifestyle, diet and health before and during the pregnancy, and growth and development of their children.

In the context of the study, 6 surveys are being collected: 3 surveys before pregnancy and 3 surveys from pregnancy to the child’s first year. The study expands over time and adds new data sources and collection methods such as allergy tests and additional surveys.

CRIBS quality assurance

Due to the unique characteristics of the longitudinal study, quality assurance is seen as an iterative process within the CRIBS study, where the characteristics of data collection processes and data handling are evaluated at each time point and analysed to improve the QA process for the following time point. In this way, methods and procedures for preventing and correcting problems that can affect the quality of the survey data are constantly being upgraded over time.

At the beginning of the study, the QA consisted of the following steps:

1. prevention - standard procedures were used to ensure accurate and consistent measurements throughout the study. Standardized training manuals were developed to document measurement protocols, detail procedures, and minimize errors. Data collection procedures for each registry were clearly defined and described. Manuals were presented in paper form;
2. detection - exploratory data analysis prior to data analysis was used in different software packages, depending on the researcher's preference (R, IBM SPSS, MS Excel);
3. correction - QA process concluded with a team debriefing of measurement activity to review results, discuss corrections and provide clarifications. The aim was to establish a continuous feedback mechanism between data sources and the research team to ensure consistency of data types, quantity, quality and origin.

CRIBS data collection

In the first wave of survey data collection, data were collected primarily through web surveys. CRIBS surveys do not contain HIPAA identifiers, and respondents are identified by a unique code. The advantages of web surveys are that they are very cost-effective. Relying on web surveys also has its drawbacks, such as excluding those participants who do not have a computer or are unable to access a computer.

For the purposes of conducting the CRIBS study web survey, we have opted for Google Forms as a commonly used survey data collection tool. Call for such surveys are sent by email. Respondents who did not have an email received hard copies of the surveys at their postal address. Upon receiving them, the researchers would manually enter such copies through the web form into the database. The collected data were then reviewed in the software package according to the preferences of the researchers, mainly IBM SPSS and MS Excel, in which Exploratory Data Analysis was performed to find possible errors.

Sources of errors

Through a case study and semi-structured interview with members of the research team, the following problems were identified in the research workflow. We detected expected data collection problems but also some less often discussed problems. Problems detected within the CRIBS study corresponding to problems discussed in other research papers (Yamanaka et al., 2016; Read et al., 2017; Young et al., 2007) were:

1. panel conditioning - the response may have been conditioned by previous experience of taking part in the survey;
2. sample attrition - continued loss of respondents from the sample due to nonresponse at each wave of a longitudinal survey;
3. recall bias;
4. generally-increased temporal and financial demands associated with these longitudinal studies.

Issues that were not found in relevant research papers were classified into the following categories: single-source problems, multi-source problems, security problems, design questionnaire problems, and QA workflow problems.

Single source problems are related to inconsistency and inaccuracy of collected data point and they do not reflect the quality of the database. Examples include *errors in data entry* (errors because of the interpretation of questions by the participant, unintentional errors such as misspellings, intentional distortion of data), *missing values*, *embedded values* (multiple values entered in only one field), *misplaced values* (values entered in the incorrect field), *duplicate entries* and *contradictory entries*.

Multi-source problems occur when multiple data sources, i.e. multiple surveys, have to be merged into a warehouse or aggregate database. Data sources often contain the same data but in different representations, which are often contradictory to one another. Such problems are a reflection of faulty survey design and are characteristic of longitudinal studies given the incidence of recurring questions. An example of a multi-source problems occurr when multiple data sources, i.e. surveys are designed with different names for the same variable which creates structural conflicts (e.g., survey “A” uses the term “customer” and source “B” uses the term “client”). Second example refers to a different representation of the same values when the variable (i.e. column in a tabular database) is called the same (e.g., survey “A” for a dichotomous “gender” variable uses “0/1” labels, while survey “B” for a variable of the same dichotomous “gender” variable uses different value labels such as “M/F”).

Security problems. We recorded a case where an anonymous employee changed the survey content. The data was recovered because we connected Google Forms survey data with Python script for backing up the data that was called twice a day via cron job at the beginning and the end of working time. However, data recovery was possible only because including data backup in our QA plan and making a custom backup script, since Google Forms do not have advanced backup features.

Problems related to the design of the questionnaire arise from the general design of the questionnaire and the chosen data collection tool, Google Forms. The general design of the questionnaire refers to the structuring of the question. Some of our surveys had a certain number of so-called “free text” questions that lead to single-source issues such as values entered in the wrong field. Abstracting data from free text is often a tedious process and it usually requires a human reader. The next decision in the questionnaire design concerns the decision of which questions to ask as mandatory. Specifically, after making questions mandatory, we have noticed a “trade-off” between missing values and the number of errors in data entry. In the case of mandatory questions, the number of missing values was kept to a minimum, but the number of single-source problems increased towards the end of the survey. In the case of optional questions, a noticeably smaller number of single-source problems was observed, but the number of skipped questions, which generated missing values, was increased. Furthermore, Google Forms does not contain the “save progress” feature, which is why the validity of such surveys may be in question as people might be in a hurry to complete it and so might not give accurate responses. Google Form also lacks advanced validation features for data input, making the “data cleaning” process extremely demanding.

Quality assurance workflow problems. A sustainable workflow model needs to be made. We detected some parts of our QA workflow lacking a reproducibility feature. For example, multiple software packages such as MS Excel, IBM SPSS and Statistica have been used for the same purpose. Since each of the following programs works with its proprietary file, as a result, a large number of heterogeneous files were created for the same data set which are not fully compatible with each other. That led to problems in creating a consistent workflow for working with the data, namely prevention and detection. Also, QA workflow required more comprehensive documentation of procedures in a more detailed manual.

Eliminating the errors

After a semi-structured interview, which examined the research team's attitudes towards the sources of errors found in the case study and discussed proposals to address them, a focus group was organized with the same members of the research team to provide a more thorough argumentation of the same topics and to obtain a wider range of information. The two sets of interventions (prevention and detection) were made according to the specified sources of error according to the steps of the QA process and will be discussed in the following sections. The third type of intervention (correction) will not be discussed at this time due to the length and complexity of the steps involved.

Prevention

Interventions that can be classified as a prevention step, can further be subdivided into four distinct models.

Managing attrition rate. Methods of email campaigns were used when sending web surveys to respondents for a more detailed insight into participants' behaviour. Of the last 353 web surveys submitted, we had a click rate of 71%, that is, 29% of respondents did not open an email within the span of 3 weeks. Of the 71% open emails, 78% of surveys were completed within three weeks. By stratifying respondents by their behaviour, we could elaborate campaigns tailored to a specific group of respondents to reduce sample attrition. A smaller group of respondents with a smaller click rate is devoted more time and is contacted by telephone.

Particular attention should be paid to sample attrition as a source of data quality problems. The problem of study attrition is unique to longitudinal designs and must be accounted for while presenting study results. From an analysis perspective, sample attrition is information about sample behaviour and can thus provide additional insight into the results. Still, sample attrition is not often talked about in the context of research (longitudinal) studies.

Project tailored tool. Downsides of a general survey tool have been revised that led to a decision to implement a new web survey collection tool, REDCAP (Harris et al., 2009). REDCAP is a fully compliant data collection tool with DPA & GDPR. User privileges and rights can be controlled and all interactions are logged and auditable. It has an advanced form with data validation features that eliminates certain single-source problems such as errors in data entry, values entered in the wrong field, and duplicated values. Moreover, it has the feature of saving data entry progress and resuming later. This reduces the trade-off effect between missing values and the number of single-source issues, but also attrition rates.

Live chat service. Demo live chat service is underway where the respondent can contact a research team member in real-time. In the demo version, such a feature proved to be extremely useful for reducing single-source problems such as errors in the interpretation of a question by the participant. However, maintaining the real-time help desk service is extremely challenging and time-consuming for a small research team.

Data documentation design. Documentation of procedures in a more detailed manual is under development. The manual now contains new information such as *a priori* specification of potential confounding variables. Documentation and manuals are in paper form and also in the form of self-hosted wiki, as wiki form turned out to be a great way to set up an in-house knowledge base. Wiki is being constantly updated. In addition to the manual, an interactive codebook following DDI Alliance instructions is under construction. The Data Documentation Initiative (DDI) (Data Documentation Initiative, n.d.) is an international standard for describing data produced by surveys and other observational methods for research data. Codebook also eliminates most multi-source problems, i.e. problems that occur when multiple data sources, i.e. multiple surveys, must be merged into an aggregate database or a data warehouse. Finally, the codebook can integrate within the new workflow written in the R programming language which makes it easier to export data to other analysis programs preferred by other members of the research team.

Detection

In addition to prevention models, we distinguish two observable cases of detection models.

Exclusive data analysis tool. Using multiple versions of software packages (R, IBM SPSS, MS Excel) for the same purpose (namely exploratory data analysis) is no longer possible. The ETL process within the R programming language is being made. R is not really designed for ETL - R by design loads data into memory, so it is limited by the amount of memory the user has available in his system. However, since the size of the data used is usually one of the major determinants of viable ETL, R's “Tidyverse” package has shown to be a good choice for narrow scope ETL – in our case, for small-scaled survey data. It should be noted, however, that R lacks high-level ETL process support and lacks features such as staging objects, manual logging and visualizing data pipes.

Traceability. An “R Markdown” in HTML format is created within the ETL process for an interactive report where each member of the research team can see the status of each ETL step and provide their feedback on the issues. The next step is to implement a more comprehensive data quality report.

Discussion and conclusion

The aim of this paper was to promote transparency and to share the insights about the errors inherent in most studies containing survey data. Those errors affect data quality, and data quality should be a key priority when planning a longitudinal study to guarantee appropriate results and conclusions from survey data. In practice, the commonly used TSE approach has not proven to be sufficient when working with survey data and analysing their quality. Biemer and Lyberg (2003) criticize TSE and say that it lacks a user perspective and should be complemented by a more modern quality paradigm.

Survey data quality is currently a vague concept with multiple definitions and sources, and according to Houston (2018), only a small body of academic research has described the use of data quality in research. Papers on quality assurance in research (longitudinal) studies predominantly talk about structures, processes, and policies that need to be a place to ascertain the quality of the data collected, but in-depth insights of information science and computer science approaches are rarely seen. Many of the data-centric topics such as data cleaning and transforming research data, building data pipelines, detection of errors in the data collection process and database-related challenges are rarely discussed - especially in social science and humanities.

Hence, it is necessary to reopen the methodological discussion about data quality and data in general research studies - a space where information experts can certainly find their place - especially when looking for new challenges on the horizon. For instance, setting up a causal frame according to observational data from the field was always a challenge. Consequently, researchers started to consider other data sources and integrating them into their survey-based analyses in order to work with innovative research questions (Spjuth et al., 2016). This entails challenges such as data linkage, i.e. merging survey data with other sources. Along with data linkage techniques, we can see the rise of harmonizing research data that refers to linking multiple different studies into one unified data warehouse (Spjuth et al., 2016). Finally, one cannot ignore the importance of FAIR guiding principles for research data management and stewardship which emphasises the capacity of computational systems to find, access, interoperate, and reuse data (European Commission, 2016). For this reason only, it is advisable to create a reproducible analytical pipeline - which opens a myriad of new challenges known in the IT sector but is rarely mentioned in the context of scientific research, such as usage of version control, dependency management, the need for good schema design and choosing the right tools in general.

Information scientist and computer experts (or as the trends suggest, „data scientists“) should play a more prominent role within the work of research (longitudinal) studies and be more open about their techniques and their advantages and disadvantages used for dealing with the research data and aim to integrate those insights into quality assurance as well as data management plans.

References

Bai, L., Meredith, R., Burstein, F. (2018). A Data Quality Framework, Method and Tools for Managing Data Quality in a Health Care Setting: An Action Case Study. // Journal of Decision Systems 27, 144-154

Biemer, P. P., Lyberg, L. E. (2003). Introduction to Survey Quality. John Wiley & Sons

Chen, H., Hailey, D., Wang, N., Yu, P. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. // International Journal of Environmental Research and Public Health 11, 5, 5170-5207

Data Documentation Initiative. Welcome to the Data Documentation Initiative. (2019). https://ddialliance.org/ (25.8.2019)

European Commission. (2016). H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 v3.0. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (1.9.2019)

Harris, P. A., Taylor, R., Thielke, R. (2009). Research Electronic Data Capture (REDCap) - A Metadata-Driven Methodology and Workflow Process for Providing Translational Research Informatics Support. Journal of Biomedical Informatics 42, 2, 377-381

Houston, L., Yu, P., Martin, A., Probst, Y. (2018). Defining and Developing a Generic Framework for Monitoring Data Quality in Clinical Research. // AMIA Annual Symposium Proceedings, 1300-1309

Jedinger, A., Watteler, O., Förster, A. (2018). Improving the Quality of Survey Data Documentation: A Total Survey Error Perspective. Data 3, 4

Leedy, P. D., Ormrod, J. E. (2001). Practical Research: Planning and Design. Upper Saddle River, N. J.: Merrill Prentice Hall

McFadden, F. R., Prescott, M. B., Hoffer, J. A. (1998). Modern Database Management. 5th ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co.

Read, K. B., LaPolla, F. W., Tolea, M. I., Galvin, J. E., Surkis, A. (2017). Improving Data Collection, Documentation, and Workflow in a Dementia Screening Study. // Journal of the Medical Library Association 105, 2, 160-166

Sáez, C., Martínez-Miranda, J., Robles, M., García-Gómez, J. M. (2012). Organizing Data Quality Assessment of Shifting Biomedical Data. Studies in Health Technology and Informatics 180, 721-725

Spjuth, O., Krestyaninova, M., Hastings, J. (2016). Harmonising and Linking Biomedical and Clinical Data across Disparate Data Archives to Enable Integrative Cross-Biobank Research. // European Journal of Human Genetics 24, 521-528

Szklo, M., Nieto, F. J. (2014). Epidemiology: Beyond the Basics. Jones & Bartlett Publishers

Tayi, G. K., Ballou, D. P. (1998). Examining Data Quality. // Communications of the ACM 41, 2, 54-57

Weisberg, H. F. (2009). The Total Survey Error Approach: A Guide to the New Science of Survey Research. University of Chicago Press

Whitney, C. W., Lind, B. K., Wahl, P. W. (1998). Quality Assurance and Quality Control in Longitudinal Studies. Epidemiologic Reviews 20, 1, 71-80

World Health Organization. (2003). Improving Data Quality: A Guide for Developing Countries. Manila: WHO Regional Office for the Western Pacific. https://apps.who.int/iris/handle/10665/206974 (25.7.2019)

World Health Organization (2017). Data Quality Review: Module 3: Data Verification and System Assessment. https://apps.who.int/iris/handle/10665/259226 (25.7.2019)

Yamanaka, A., Fialkowski, M. K., Wilkens, L. (2016). Quality Assurance of Data Collection in the Multi-Site Community Randomized Trial and Prevalence Survey of the Children’s Healthy Living Program. // BMC Research Notes 9, 1, 432

Young, A., Powers, J., Wheway, V. (2007). Working with Longitudinal Data: Attrition and Retention, Data Quality, Measures of Change and Other Analytical Issues. // International Journal of Multiple Research Approaches 1, 2, 175-186

https://doi.org/10.17234/INFUTURE.2019.13 Preliminary communication

Evaluation of a MOOC to Promote Information Literacy
First Evaluation Results

Stefan Dreisiebner
University of Hildesheim, Germany
stefan.dreisiebner@uni-graz.at

Maja Žumer
University of Ljubljana, Slovenia
Maja.Zumer@ff.uni-lj.si

Polona Vilar
University of Ljubljana, Slovenia
Polona.Vilar@ff.uni-lj.si

Thomas Mandl
University of Hildesheim, Germany
mandl@uni-hildesheim.de

Summary
The aim of this paper is to present the evaluation framework and preliminary results of the evaluation for a Massive Open Online Course (MOOC) on Information Literacy (IL) and its application within a Business Administration course. The aim of the evaluation was to assess user experience and progress of the students' knowledge of IL after completing the MOOC. The evaluation approach consisted of three phases: First, the students were asked to fill out a short self-assessment questionnaire and a shortened adopted version of a standardized IL test. Second, they completed the full version of the IL MOOC. Third, they were asked to fill out the full version of a standardized IL test and a user experience questionnaire. The evaluation results show that the MOOC was able to increase the IL skills of the students and was also perceived well. The evaluation approach worked well and can also serve as model for evaluations of other MOOCs, in particular on IL.

Key words: MOOC, information literacy, evaluation

Introduction

MOOCs (Massive Open Online Courses) are freely available online courses that have no entry limitations and are aiming at unlimited participation (Bozkurt et al., 2017). In November 2016, the European Union funded project *Information Literacy Online* (ILO) was started with the aim to develop, evaluate and disseminate a multilingual open-access MOOC designed to improve students' information literacy. Information literacy (IL) refers to the "set of integrated abilities encompassing the reflective discovery of information, the understanding of how information is produced and valued, and the use of information in creating new knowledge and participating ethically in communities of learning" (ACRL, 2016).

The content framework of the ILO MOOC is based on the SCONUL Seven Pillars of Information Literacy (SCONUL, 2011), on the ACRL Framework for Information Literacy for Higher Education (ACRL, 2016), on the Metaliteracy model (Jacobson, Mackey, 2016) and a good practice analysis in IL education (Robinson, Bawden, 2018). This led to the following MOOC content:

- Module 1: Orienting in an information landscape
- Module 2: Research is a journey of inquiries
- Module 3: The power of search
- Module 4: Critical information appraisal
- Module 5: Information use: the right and fair way
- Module 6: Let's create something new based on information and share it!

The content was developed in English first and later translated to Spanish, Catalan, German, Croatian and Slovenian. The final content was implemented into the *OpenEdX* platform following a pre-defined workflow (Libbrecht et al., 2019). An important final development stage of the ILO MOOC was the evaluation of two areas: a) the user experience and b) achievement of the planned learning outcomes (ie., progress of the students' knowledge of IL after completing the MOOC). While there are already several scientific contributions available that report about experiences and evaluation results of MOOC projects, there is a lack of such work regarding MOOCs on IL. The aim of this paper is to present the evaluation framework and preliminary results of the evaluation of this MOOC to encourage its further development which we hope would lead to the use of the MOOC in the area of business studies with its final goal of promoting IL and its application within this study area.

State of the Art

There are several scientific publications reporting on experiences and evaluation results of MOOC projects. A common approach is to use questionnaires on the user experience. A study about the MOOC experience at the Spanish National University of Distance Education used surveys covering 17 MOOCs offered by the university's own platform to analyse completion rates and the overall user experience (Gil-Jaurena et al., 2017). For an Australian study, conducted in cooperation with the OpenLearning platform, an evaluation form was embedded directly into the MOOCs to collect and analyse participants' attitudes and the perceived usefulness (Rawlings et al., 2017). A MOOC on literature searching for health libraries was evaluated by asking participants to fill out a feedback form at the end of the MOOC to derive recommendations for future projects (Young et al., 2017).

The most common approach to evaluate IL skills are standardized questionnaires (Beile, 2008). There are several tests that are optimized for specific target groups and educational levels, such as the Information Literacy Test of the James Madison University focusing on students (Cameron et al., 2007), the Beile Test of Information Literacy for Education focusing on future teachers (Beile, 2005), the Tool for Real-time Assessment of Information Literacy focusing on pupils (Kent State University Libraries, 2013), and the Information Literacy Test (ILT) for Higher Education (Boh Podgornik et al., 2016). The advantage of standardized IL tests is that they are quick and easy to administer and produce readily analysable and comparable data. However, no single test is able to capture the complexity of learning: While standardized IL tests in multiple-choice format are good at measuring lower-order thinking skills, they are not suited to measure higher-order thinking processes. Multiple forms of assessment would be needed to fully measure student performance and program effectiveness (Beile, 2008; Beutelspacher, 2014).

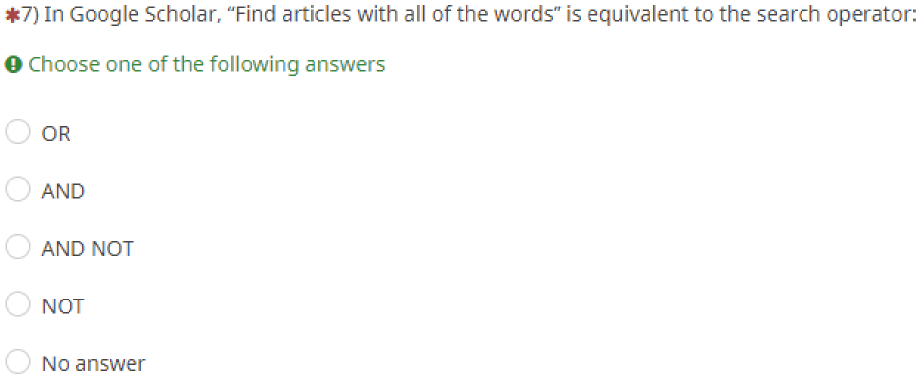
Methodology

The evaluation approach, applied in this study, consisted of three phases: First, the students were asked to fill out a short pre-test. Second, they completed the full version of the ILO MOOC. Third, they were asked to fill out a longer post-test. To allow matching of the pre- and post-test questionnaires, self-generated identification codes (Yurek et al., 2008) were used. Both questionnaires were implemented into LimeSurvey.

The pre-test consisted of four parts: 1) a questionnaire on personal background information, such as age, study program, and previous degrees; 2) a self-assessment of IL skills consisting of seven questions on previous experience and information needs which were rated on a three-point Likert scale, finishing with an open question on perceived problems regarding information needs; 3) a shortened adopted version of a standardized IL test (Boh Podgornik et al., 2016) consisting of 12 single-choice questions (Figure 1 gives an example of one question within this questionnaire); 4) a short questionnaire with three subject-related questions. Students were expected to complete the pre-test in 5-7 minutes.

The post-test consisted out of three parts: 1) the full version of a standardized IL test (Boh Podgornik et al., 2016) consisting of 39 single choice questions; 2) the same three subject-related questions as in the pre-test; 3) a user experience questionnaire. This questionnaire asked for the language setting used when attending the MOOC and allowed an open response answer on the overall experience with the MOOC. Afterwards, nine usability aspects were to be rated on a five-point Likert scale. For two of these aspects, participants had the possibility to leave comments. Finally, participants were asked how important they considered to have information in various formats within the course and had the possibility to leave open-response comments on anything particularly disturbing or anything particularly appealing when using the user interface. Students were expected to complete the post-test in 15-20 minutes.

20 students (14 male and 6 female) of the course *Business Intelligence* at the University of Graz, Austria, participated in the evaluation between March and April 2019. The course introduces students into web sources for competitive intelligence analyses and one business intelligence software tool. The age of the students ranged from 23 to 41 years (mean 26 years). All of them were enrolled in either the master program Business Administration or Business Education and Development, except one student, who was enrolled in the doctoral program of Economics and Social Sciences. The students were asked to participate in the German version of the MOOC outside of the regular class hours at home. They were required to register with their real name and active participation was encouraged and checked by their instructor.



*7) In Google Scholar, “Find articles with all of the words” is equivalent to the search operator:

Choose one of the following answers

○ OR

○ AND

○ AND NOT

○ NOT

○ No answer

Figure 1: Example of one question within the pre-test, source: Question adapted from Boh Podgornik et al. (2016).

Findings

The self-assessment before attending the MOOC showed that the students were confident about their skills, but were also aware of shortcomings (Table 1).

Table 1. Self-assessment of students

Item	Never	Sometimes	Frequently
I often search for information as part of my study activities	0%	15%	85%
	Not at all	**Some of them**	**Most of them**
I know the most important sources of information in my field	0%	85%	15%
	Never	**Sometimes**	**Always**
I know which source to use when I need a particular type of information	5%	80%	15%
	No	**Sometimes**	**Yes**
I can successfully use most sources to retrieve the information I need	0%	35%	65%
I usually find the information I need in the sources that I’m using	10%	25%	65%
I can compare and evaluate different resources	25%	25%	50%
I know how to use information appropriately to the task	20%	35%	45%

85% of the participating students frequently (weekly or several times a month) search for information as part of their study activities and 15% search at least sometimes (several times during a semester) for information. 85% believe to know some of the most important information sources in their field, but only 15% believe to know most of them. 80% sometimes know which source to use when a particular type of information is needed, but only 15% believe they always do. 65% of the students think that they can successfully use most sources to retrieve the needed information and acknowledged to usually find the information need in the used sources. The students appear to be less confident regarding comparing and evaluating different resources and using information appropriately to the task, where 25% and 20% answered with *no* and only 50% and 45% answered with *yes*, respectively. As main problems regarding their information needs the students reported to struggle with finding relevant information and information overload.

As Table 2 shows, the average result of the standardized generic questionnaire increased by 6.54% from 78.33% before attending the MOOC to 84.87% after the students have attended the MOOC. The worst test result increased from 50% to 61.10%, while the best test result decreased from 100% to 94.87%. A Wilcoxon signed rank test shows that the observed difference between pre- and post-test is significant (Z=-2.073; p=0.038).

Table 2. Results of the generic questionnaire before and after the MOOC

Testing point	Av. Result	Min Result	Max Result	Mean Points	Max Points	Std. Deviation
Pre-Test	78.33%	50%	100%	9.40	12	1.90
Post-Test	84.87%	64.10%	94.87%	33.10	39	2.86

The average result of the subject-related questionnaire increased by 28.33% from 31.67% before attending the MOOC to 60% after attending the MOOC (Table 3). However, before and after attending the MOOC there were students that gained nothing as well as those gaining 100%. A Wilcoxon signed rank test shows that the observed difference between pre- and post-test is significant (Z=-2.538; p=0.011).

Table 3. Results of the subject-related questionnaire before and after the MOOC

Testing point	Av. Result	Min Result	Max Result	Mean Points	Max Points	Std. Deviation
Pre-Test	31.67%	0%	100%	0.95	3	0.80
Post-Test	60%	0%	100%	1.80	3	1.12

As the previous results have shown, some students seem to have done worse in the test after the MOOC, as the best achieved test result decreased. An analysis of the questionnaires paired through the self-generated identification codes shows that indeed the test result decreased for 5 students in the generic test and for 3 students in the subject-related test. Nevertheless, the majority of 15 students in the generic test and 17 students in the subject-related test were able to increase their results. The maximum increase was 37.18% and the maximum decrease - 7.05% for the generic part. In the subject-related part students were able to both increase their result by 100% as well as decrease by -100% (Table 4). The fact that the results of a few students decreased from the pre- to the post-test might be explained through the fact that the generic post-test was much longer (39 vs. 12 questions), while the subject-related test was quite short (only 3 questions), where only one wrong answer already had a relatively high impact on the result.

Table 4. Change of the test results

Item	Generic part	Subject-related part
Average increase	6.54%	28.33%
Students increased	15	17
Maximum increase	37.18%	100%
Students decreased	5	3
Maximum decrease	-7.05%	-100%

The user experience questionnaire asked for ratings on a five-point Likert scale, with 1 meaning *very unsatisfactory* and 5 meaning *very satisfactory*. Table 5 provides an overview of the results. As can be seen, the highest satisfaction was reported for finding the next/previous navigation buttons (4.82), moving between the individual lessons (4.55) and the navigation in the user interface (4.50). The highest dissatisfaction was with the amount of information on the screen (3.00), clarity and general quality of the text of the lessons (3.27) and amount of material in the course (3.50). As reasons for their rating of the organization of the interface (buttons, menus, etc.) (3.77) the students commented positive ratings with “very intuitive” and “simple structure” and critical ratings with “slow video speed”, “usability issues with quizzes” and “animations would be nice”. As reasons for their rating of the language of the interface (3.64) students commented positive ratings with “clear” and “easy to understand” and critical ratings with “spelling errors”, “grammar errors” and “some elements are in English”.

Table 5. Reported user experience

Item	Mean	Min	Max	Std. Deviation
Navigation in the user interface	4.50	2	5	0.86
Finding the next/previous buttons	4.82	4	5	0.39
Moving between individual lessons	4.55	2	5	0.86
Amount of material in the course	3.50	2	4	0.60
Amount of information on the screen	3.00	1	4	0.53
Clarity and general quality of the text of the lessons	3.27	1	5	1.12
Layout of the text on the screen	3.73	2	5	0.98
Organization of the interface (buttons, menus, etc.)	3.77	1	5	1.15
Language of the interface	3.64	1	5	1.18

The students were additionally asked how important they considered to have information in various formats (e.g. text, videos) within the course. The mean of the answers on a five-point Likert scale was 4.23, which means *very important*. Finally, the students were asked whether they found anything particularly disturbing or anything particularly appealing when using the user interface. Not all students answered these open questions. As particularly disturbing, students named “hard to distinguish between exercises and content”, “multiple choice quiz, but only single answer selectable”, “progress bar not accurate”, “some content too detailed and some videos too long”, “too much text” and “different length of the single learning steps”. Five students answered this question with “no”. As particularly appealing students named “easy navigation”, “simple structure”, “helpful quizzes” and “lots of videos and examples”. Two students answered with “no” and one with “neutral”.

However, in the open answers regarding their overall experience with the MOOC the students gave mainly positive comments: students called the MOOC “very helpful and informative”, “well structured” and acknowledged “valuable references to external websites and information sources” and “different media formats”. Critical comments were given regarding “too detailed content”, “video speed” and “server connection issues”.

Conclusion

The results show an increase in the test result by 6.54% for the standardized IL test and 28.33% for the subject-specific questionnaire. A Wilcoxon signed rank test shows that the observed increase is in both cases significant. Thus, it seems that the MOOC was able to increase the IL skills of the students. However, the knowledge gain was lower than the increase of 13% (from 65.6% to 78.6%), that was achieved when a group of 163 students took the same standardized IL test as in this study after participating in an IL-specific study course (Boh Podgornik et al., 2016). A possible explanation is that the sample in this study consists of Masters students who already had a profound previous knowledge and already achieved good results in the test before attending the MOOC (78.33%). This is also supported by their reflected answers given in the initial self-assessment.

The evaluation of the user experience showed that the MOOC was generally perceived well. The students provided detailed feedback on issues like “multiple choice quiz, but only single answer selectable”, which enabled an immediate localization and fixing. Some of the mentioned issues could not be immediately fixed, like criticism of “too detailed content”. However, this point of criticism might be also explained with the profound previous knowledge of the participants, or even with general attitude of learners when learning something new.

The evaluation approach itself worked well and can also serve as model for evaluations of other MOOCs, in particular on IL. However, this work comes also with several limitations, that in turn provide avenues for future research: First, the evaluation included a relatively small sample of 20 students out of a single discipline (Business Administration) and on a similar level in their studies

(Master). Second, the students only attended the German MOOC, which is just one out of several available language versions of this ILO MOOC. Third, the evaluation of IL knowledge gains was based only on single-choice questions in the standardized IL test. Multiple forms of assessment would be needed to fully measure students' performance and the effectiveness of the MOOC. Further evaluations are planned, based on a more diverse and larger sample that will involve the MOOC in all available languages.

Acknowledgements

The production of this publication was co-funded by the Erasmus+ Programme of the European Union under the agreement number 2016-1-AT01-KA203-016762. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

ACRL (2016). Framework for Information Literacy for Higher Education. Chicago, Association of College and Research Libraries. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/issues/infolit/framework.pdf

Beile, P. (2005). Development and validation of the Beile Test of Information Literacy for Education (B-TILED). Doctoral Thesis. Kentucky, University of Kentucky. http://proquest.umi.com/pqdweb?did=1016019641&Fmt=7&clientId=18803&RQT=309&VName=PQD

Beile, P. (2008). Information Literacy Assessment. A Review of Objective and Interpretive Measures. // Society for Informationa Technology Teacher Education International Conference. Las Vegas, March 3, 1860-1867. http://hdl.handle.net/10760/15858

Beutelspacher, L. (2014). Erfassung von Informationskompetenz mithilfe von Multiple-Choice-Fragebogen. // Information - Wissenschaft & Praxis 65, 6, 341-352. http://dx.doi.org/10.1515/iwp-2014-0054

Bozkurt, A., Akgün-Özbek, E., Zawacki-Richter, O. (2017). Trends and Patterns in Massive Open Online Courses: Review and Content Analysis of Research on MOOCs (2008-2015). // International Review of Research in Open and Distributed Learning 18, 5, 118-147. http://www.irrodl.org/index.php/irrodl/article/view/3080/4284

Boh Podgornik, B., Dolničar, D., Šorgo, A., Bartol, T. (2016). Development, testing, and validation of an information literacy test (ILT) for higher education. // Journal of the Association for Information Science and Technology 67, 10, 2420-2436

Cameron, L., Wise, S. L.; Lottridge, S. M. (2007). The Development and Validation of the Information Literacy Test. // College & Research Libraries 68, 3, 229-236. http://dx.doi.org/10.5860/crl.68.3.229

Gil-Jaurena, I., Callejo-Gallego, J., Agudo, Y. (2017). Evaluation of the UNED MOOCs Implementation: Demographics, Learners' Opinions and Completion Rates. // The International Review of Research in Open and Distributed Learning 18, 7. http://dx.doi.org/10.19173/irrodl.v18i7.3155

Jacobson, T. E., Mackey, T. P. (2016). (eds.). Metaliteracy in practice. Chicago: Neal-Schuman

Kent State University Libraries. (2013). Tool for Real-time Assessment of Information Literacy Skills. http://www.trails-9.org/about2.php?page=about (13.8.2019)

Libbrecht, P., Dreisiebner, S., Buchal, B., Polzer, A. (2019). Creating Multilingual MOOC Content for Information Literacy: A Workflow. // Conference on Learning Information Literacy across the Globe. Frankfurt am Main, Germany, May 10. https://informationliteracy.eu/conference/assets/papers/LILG-2019_Libbrecht-et-al_Creating_ILO_MOOC.pdf

Rawlings, D., Miller-Lewis, L., Collien, D., Tieman, J., Parker, D., Sanderson, C. (2017). Lessons Learned from the Dying2Learn MOOC: Pedagogy, Platforms and Partnerships. // Education Sciences 7, 3, 67. http://dx.doi.org/10.3390/educsci7030067

Robinson, L., Bawden, D. (2018). Identifying Good Practices in Information Literacy Education; Creating a Multi-lingual, Multi-cultural MOOC. // Information Literacy in the Workplace / Kurbanoğlu, S., Boustany, J., Špiranec, D., Grassian, R., Mizrachi, F., Roy, L. (eds.). Cham: Springer International Publishing, 715-727

SCONUL (2011). The SCONUL Seven Pillars of Information Literacy. Core Model. London: Society of College, National and University Libraries. http://www.sconul.ac.uk/groups/information_literacy/publications/coremodel.pdf

Young, G., McLaren, L., Maden, M. (2017). Delivering a MOOC for literature searching in health libraries: evaluation of a pilot project. // Health information and libraries journal 34, 4, 312-318. http://dx.doi.org/10.1111/hir.12197

Yurek, L. A., Vasey, J., Sullivan Havens, D. (2008). The use of self-generated identification codes in longitudinal research. // Evaluation review 32, 5, 435-452. http://dx.doi.org/10.1177/0193841X08316676

https://doi.org/10.17234/INFUTURE.2019.14

Original scientific paper

Using LMS Activity Logs to Predict Student Failure with Random Forest Algorithm

Dejan Ljubobratović
Department of Informatics, University of Rijeka, Croatia
dejan.ljubobratovic@student.uniri.hr

Maja Matetić
Department of Informatics, University of Rijeka, Croatia
majam@uniri.hr

Summary

The paper presents a Random forest model in the task of predicting student success (grade) on the base of input predictors (lectures, quizzes, labs and videos) extracted from Moodle activity logs. Since 2010. University of Rijeka is using Moodle based Learning Management Systems (LMS) to complement traditional teaching. LMS is used for documents sharing, quizzes, assessments, video lecturing, tracking student progress and much more. When student access an LMS using his personal account, a digital profile is created that is saved in LMS log files. These logs were used to create a dataset with couple of hundreds of observations. However, building a prediction model using Random forest algorithm is relatively easy comparing to explaining the results. Interpreting Random forest and other machine learning black box models is a challenge regarding to complexity of their decision-making mechanisms. There are a number of new techniques allowing us to interpret such models, and couple of them is used in this paper for that purpose.
Another problem a researcher is facing using black box algorithms is GDPR. General Data Protection Regulation has a significant impact on many aspects of EU citizen's data collection and processing. This paper will highlight most challenging GDPR restrictions on data mining including GDPR's "right to explanation".

Key words: LMS system, random forest algorithm, educational data mining, predicting student success, interpretability, interpretable machine learning

Introduction

Mining and interpreting data collected in Massive Online Open Courses (MOOCs) is well researched and popular, giving a researcher huge database of different logs to deal with. For example, only one course “Learning How to Learn: Powerful mental tools to help you master tough subjects” offered by McMaster University at University of California San Diego enrolled more than 1.7 million people using Coursera platform.(Learning How to Learn, 2019)
LMS systems like Moodle are used to complement educational processes in universities and schools, with significantly smaller log database.
In this research we build a model to predict student success (grade) as a function of course activities using Random forest algorithm. Later in this work several methods were used to interpret the given model giving explanations to Random forest algorithm results. For data exploration, prediction model and result explanation in this work R language v. 3.6.1. is used, which is a freely available language and environment for statistical computing and graphics.
This work is divided in three logical parts. In the beginning are presented the basic ideas found in similar researches in order to compare them with our work, after which we highlighted the most challenging GDPR restrictions on data mining. Main part of this research is building a prediction model using Random forest algorithm, and explaining data used in the process. Interpreting results of our Random forest model using four different techniques is the main goal of the third part of this paper.

Related work

Educational data mining

Creating a precise model that can predict student future behaviour or student's final grade based on his activity is very appealing to any educational institution.

In order to classify the dropout student Yukselturk et al. (2014) used four data mining algorithms; k-Nearest neighbour, Decision tree, Naive Bayes and Neural networks. In their final results as the most important factors in predicting the dropouts were three variables; *online technologies self-efficacy*, *online learning readiness*, and *previous online experience* (Yukselturk, Ozekes, Türel, 2014).

In another conducted research, authors examined students' activity by gender, and by log time using LMS Moodle activity logs. They found there significant correlation; the female students were more active and successful in the course than are the male ones and the students were most active in the test weeks, specifically, on the day before the tests (Kadoic, Oreski, 2018).

Mishra et al. (2014) build performance prediction model based on students' social integration, academic integration, and various emotional skills. The key influencers to the *semester results* were *previous semester results,* followed by *good academic performance*. Out of all emotional attributes the *semester performance* was affected only by *leadership* and *drive of the students* (Mishra, Kumar, Gupta, 2014).

Using data mining methodology based on CRISP-DM methodology, Chalaris et al. (2014) found out that in the theoretical courses *student understanding* relates mainly with the instructor and teaching effectiveness, while in the laboratory practice courses, *lab facilities* are found to be the most correlated with the *achievement of learning objectives* (Chalaris et al., 2014).

Predicting student failure or revealing dropout factors in MOOC (Gupta, Sabitha, 2019) can help educators to redesign MOOC features (Xing, 2019), personalise teaching processes (Zhang et al., 2019), increase student performance (Ajibade, Ahmad, Shamsuddin, 2019) and finally keep the students from leaving the course. Of course, researching student data must be done in ethical way, respecting their privacy.

GDPR and data mining

The EU General Data Protection Regulation (2018), known as GDPR, is the most important change in data privacy regulations in 21st century. It has a significant impact on many aspects of EU citizen's data collection and processing, and affects not only EU companies but also multinationals which operate in EU. Machine learning models are fuelled by large amount of personal data. This means we need to respect the privacy of the individual in ethical way in order to overcome privacy risks (Ashford, 2019).

"Right to explanation" is another significant effect of GDPR on Machine Learning. According to Gregory Piatetsky GDPR doesn't really require an explanation of Machine Learning (ML) algorithms. Author distinguishes two explanations on those matters: Global explanation and Local explanation (Piatetsky-Shapiro, 2018).

Global explanation is mainly focused on how ML algorithm works. Some deep learning algorithms, so called black box algorithms, are almost impossible to interpret. Their complexity makes very challenging to understand exactly why, and how, a machine learning model has made a particular decision. On the other part, Local explanation deals with a question of factors contributed to a particular decision impacting a specific person. It is difficult to see how the meaningful explanation about the logic involved in some black box algorithms can be satisfied, especially in cases where a machine learning process involves multiple data sources, and elements that are not transparent or intuitive, whether for technological or proprietary reasons.

Revealing the full algorithm code and detailed technical descriptions of machine learning processes is unlikely to help. On the other hand a simple, non-technical, description of the process is more likely to be meaningful (Kuner et al., 2017).

Data set description

Database used in this research has 408 records collected from 5 generations of student activity in course "Programming 2". Dataset contains 6 variables: ID, lectures, quizzes, labs, videos and grade. Variable ID represents a student; although dataset is anonymized this variable was removed.

Variables *lectures*, *quizzes* and *labs* are total number of scores students received within the corresponding domain. Variable *videos* represent number of views of the video lectures and *grade* represents student grade on final exam. Research data was collected as described in previous research by Matetic using Interpretable neural networks in predicting student failure (Matetic, 2019). Sample of data used in this research is shown in Table 1.

Table 1. Sample of dataset used in the research

lectures	quizzes	Labs	videos	grade
0	19,33	32	15	D
5	22	27	7	D
5	15	7	10	F
5	27,66	27,5	13	C
3	28,66	0	50	F

Data exploration

First step, which precedes building a prediction model, is data exploration.

We're trying to predict grade, so we must pay attention to variables *labs* and *quizzes* which has the strongest relationship with grade. But, as the heatmap (Figure 1) suggests *labs* and *quizzes* has the strongest correlation between each other, while variables *videos* and *lectures* have the weakest correlation.

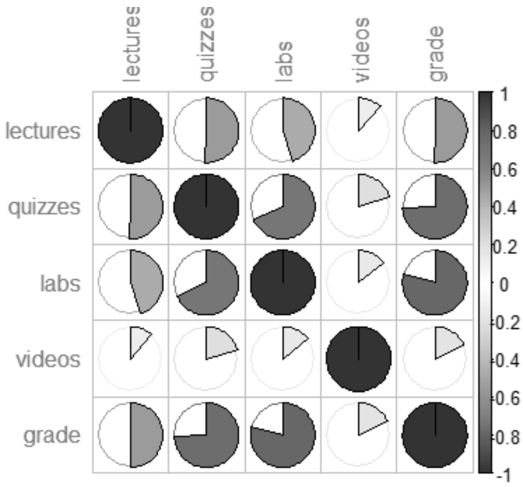


Figure 1. Data heatmap, showing correlation between variables

Plot in Figure 2 shows that FAIL grades were outnumbered by PASS ones. That means, for better results, data needs to be normalized.

Analysing plots on Figure 3, from distributions of student's grades (FAIL or PASS); we can see that lower scores in labs and quizzes mostly results in fail, giving us right skewed normal distribution. This is something that we expected.

Interesting fact to notice on quizzes plot is that FAIL distribution is slightly bimodal, showing us that certain number of students with relatively high scores on quizzes still manage to fail.

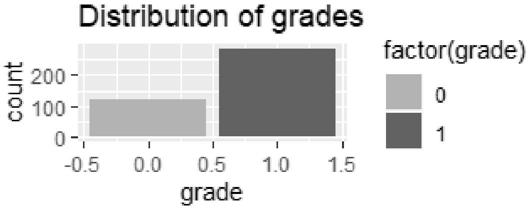


Figure 2. Distribution of grades (0-FAIL, 1-PASS)

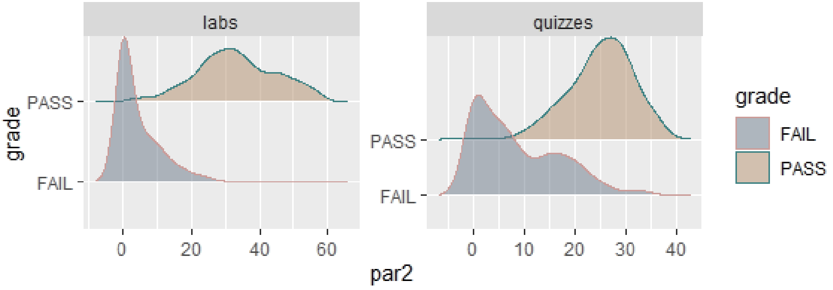


Figure 3. Distributions of students' grade (FAIL and PASS) by labs and quizzes variables; x axis (par2) shows student activity points

Building a prediction model using Random forest algorithm

Random forests algorithm constructs each tree using a different bootstrap sample of the data, and change how the classification or regression trees are constructed (Liaw, Wiener, 2003).

While in standard trees, each node is split using the best split among all variables, Random forest splits each node using the best among a subset of predictors randomly chosen at that node. This method performs very well compared to many other classifiers, including Discriminant analysis, Support vector machines and Neural networks, while it is robust against overfitting. It is very user-friendly method in the sense that it has only two parameters - the number of variables in the random subset at each node and the number of trees in the forest, and is usually not very sensitive to their values (Breiman, 2001).

For creating a Random forest model in this research, we used R language v. 3.6.1. with Caret package installed.

First step in our process was splitting our data into two sets: training data (80%) and test data (20%). We used 3 fold cross validation repeated 5 times, and then we build Random forest model (rf_model) with centered and scaled data. After the model was build, it was tested on test data, and model accuracy was 96.3%.

So, we build a model that predicts student failure using Random forest algorithm with high accuracy, but we have no clue on how this model makes prediction. Random forest algorithm is so called *black box* algorithm. Black box models, such as Random forest or Neural networks, give us little information regarding their decision-making processes, so we need an extra effort to explain it (Grigg, 2019).

Interpretation of Random forest model

Algorithms that hide their internal logic to the user, so called black boxes give us little information regarding their decision-making processes. This lack of explanation presents practical and an ethical issue. There are many approaches aimed at overcoming this weakness sometimes at the cost of reducing accuracy in benefit of interpretability (Guidotti et al., 2018).

On the other hand, models that are easy to interpret (whitebox) such as linear regression and decision trees tend to be inaccurate, as they often fail to capture complicated relationships within a dataset. In this work several methods to interpret the results of our Random forest model was used.

Variable importance

When training a Random forest model, it is normal to ask which variables have the most predictive power. High-importance variables are essential to model making and their values significantly affect the outcome values. On the other hand, variables with low importance can be left out from a model, and make it simpler and faster to fit and predict (Hoare, 2019).

The prediction error rate for Random forest classification model is calculated for permuted out-of-bag data of each tree and permutations of every feature. These two measures are averaged and normalized. As we can see on variable importance plots (Figure 4), variable *labs* is the most important variable in decision making process. Predictor *videos* is relatively important for class value PASS, but it's overall irrelevant.

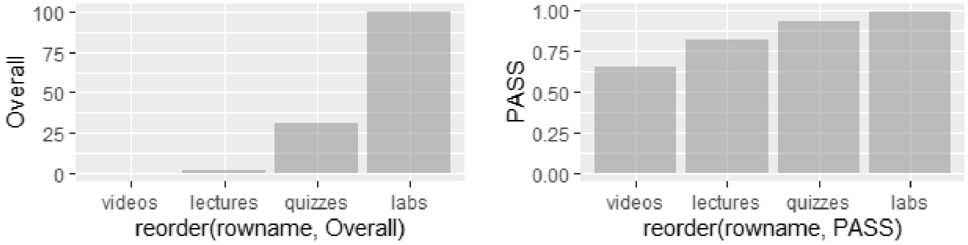


Figure 4. Overall variable importance and variable importance for PASS

Break Down model

The Break Down is a model agnostic method for decomposition of predictions from black boxes such as Random forest, Xgboost, Support vector machine (SVM) or Neural networks. As a result we get decomposition of model prediction that can be attributed to particular variables. Break down plot presents their contributions in graphical way (Figure 5).

Using R code with package breakDown, we detected variables that contributed the most to our final prediction. This method gives us the same variable *labs* as a most valuable predictor. That corresponds to result given by Variable importance tool.

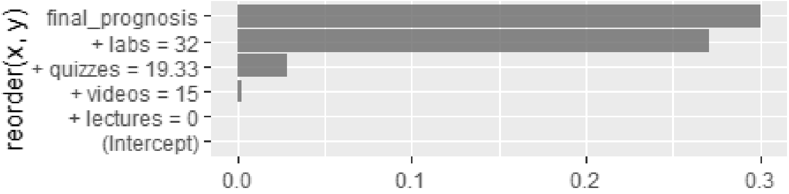


Figure 5. Break down plot visualise variables from Break down table

Tree surrogate

The tree surrogate method uses decision trees on the predictions where conditional inference tree is fitted on the predictions from the machine learning model and data. The R-squared value (variance explained) gives an estimate of the goodness of fit or how well the decision tree approximates the model. Our surrogate model has an R-squared of 0.836 which means it approximates the underlying black box behaviour quite well, but not perfectly. As we can see on Tree surrogate plot (Figure 6) *labs* is the most important predictor again.

On the right side of a plot, predictors *labs* and *quizzes* contributed the most to variable class PASS, while predictors *labs* and *lectures* (left side of a plot) are the most important to variable class FAIL.

The results are given in decision tree form, which is easy to interpret in contrast to Random forest lack of transparency.

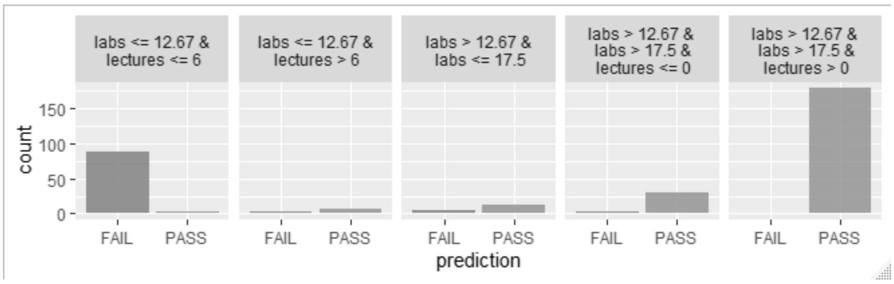


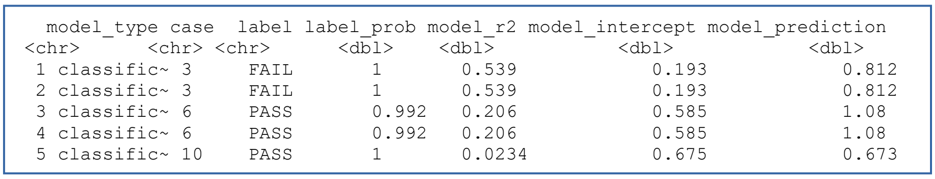
Figure 6. Tree surrogate plot

Local Interpretable Model-agnostic Explanations (LIME)

LIME is explanation technique that learns an interpretable model locally around the prediction, explaining predictions of any classifier in an interpretable and faithful manner (Guidotti et al., 2018).

LIME is explaining the predictions of black box classifiers in a way that for any given prediction and any given classifier it is able to determine a small set of features in the original data that has driven the outcome of the prediction. It creates a model agnostic locally faithful explanation set which helps us to understand how the original model makes its decision. By creating a representative sample set LIME provides to users' global view of a model's decision boundary.

The R code will give us output (Figure 7) with huge number of single outcomes which are individually explained by predictors in their own surroundings. These explanations can be visualized, but we will end up with enormous list of cases and their plots. Figure 8 show us just a sample of visualized explanation (cases from 3 to 25 out of 172).



```
  model_type case  label label_prob model_r2 model_intercept model_prediction
<chr>       <chr> <chr>       <dbl>    <dbl>           <dbl>            <dbl>
 1 classific~ 3     FAIL        1        0.539           0.193            0.812
 2 classific~ 3     FAIL        1        0.539           0.193            0.812
 3 classific~ 6     PASS        0.992    0.206           0.585            1.08
 4 classific~ 6     PASS        0.992    0.206           0.585            1.08
 5 classific~ 10    PASS        1        0.0234          0.675            0.673
```

Figure 7. R output - Sample of individual cases with corresponding predictors and their weights

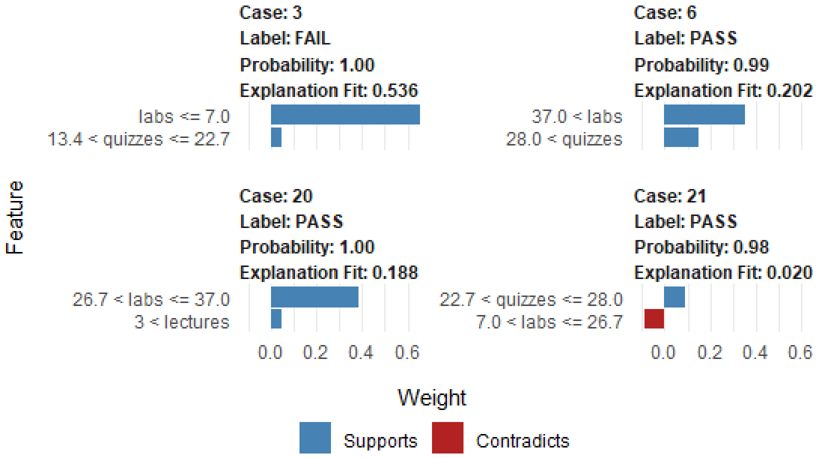


Figure 8. LIME sample of visualized explanation

Conclusion

If we need accuracy in predictions, we are usually forced to use machine learning models that are mostly black boxes. In other words, we cannot understand its learning processes or figure out logic behind its conclusions. But there are tools that explain our model's decision boundary in a human understandable way and for that purpose in this work we used several tools.

If we plan to take actions based on a prediction, or when we choose whether to deploy a new model or not, it is fundamental to understand the reasons behind predictions, and this is very important in assessing trust. Understanding the model, we can transform an untrustworthy model or prediction into a trustworthy one.

In order to create trust in our model, we need to explain the model not only to machine learning experts but also to domain experts which require a human understandable explanation.

In this work we used Random forest algorithm to build a model that can predict student failure with 96.3% accuracy what is quite good, but knowing almost nothing about which inputs contributed to that result. Using model interpreting tools, we revealed two most important variables; *labs* and *quizzes*. Variable *labs* is the strongest predictor in all our interpreting models and that understanding gives us the chance to intervene in educational process and make it better, what was our initial goal. We could use any given model to interpret our model predictions, but achieving same results with several techniques gives us trust in our model.

In our future work we plan to apply also problem domain appropriate time-series models investigating their interpretability.

Acknowledgment

This work has been fully supported by the University of Rijeka under the project number uniri-drustv-18-122.

References

Ajibade, S. S. M., Ahmad, N. B. B., Shamsuddin, S. M. (2019). Educational Data Mining: Enhancement of Student Performance Model Using Ensemble Methods. IOP Conference Series: Materials Science and Engineering 551, 012061. https://doi.org/10.1088/1757-899x/551/1/012061

Ashford, W. (2019). GDPR a Challenge to AI Black Boxes. // ComputerWeekly.Com. 2019. https://www.computerweekly.com/news/252452183/GDPR-a-challenge-to-AI-black-boxes

Breiman, L. (2001). Random Forests. // Machine Learning 45, 1, 5-32. https://doi.org/10.1023/A:1010933404324

Chalaris, M., .Gritzalis, S., Maragoudakis, M., Sgouropoulou C., Tsolakidis, A. (2014). Improving Quality of Educational Processes Providing New Knowledge Using Data Mining Techniques. // Procedia - Social and Behavioral Sciences 147, 390-97. https://doi.org/10.1016/j.sbspro.2014.07.117

Grigg, T. (2019). Interpretability and Random Forests. // Towards Data Science. https://towardsdatascience.com/interpretability-and-random-forests-4fe13a79ae34

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F. (2018). Local Rule-Based Explanations of Black Box Decision Systems. http://arxiv.org/abs/1805.10820

Gupta, S., Sabitha, A. S. (2019). Deciphering the Attributes of Student Retention in Massive Open Online Courses Using Data Mining Techniques. // Education and Information Technologies 24, 3, 1973-1994. https://doi.org/10.1007/s10639-018-9829-9

Hoare, J. (2019). How Is Variable Importance Calculated for a Random Forest? DisplayR. https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest

Kadoic, N., Oreski, D. (2018). Analysis of Student Behavior and Success Based on Logs in Moodle. // 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings, 654-659. https://doi.org/10.23919/MIPRO.2018.8400123

Kuner, C., Svantesson, D. J. B., Cate, F. H., Lynskey, O., Millard, C. (2017). Machine Learning with Personal Data: Is Data Protection Law Smart Enough to Meet the Challenge? // International Data Privacy Law 7, 1, 1-2. https://doi.org/10.1093/idpl/ipx003

Learning How to Learn (2019). Powerful Mental Tools to Help You Master Tough Subjects. https://www.coursera.org/learn/learning-how-to-learn

Liaw, A., Wiener, M. (2003). Classification and Regression by RandomForest. R News 2. // R News 3 (December 2002), 18-22

Matetic, M. (2019). Mining Learning Management System Data Using Interpretable Neural Networks, 1282-1287. https://doi.org/10.23919/mipro.2019.8757113

Mishra, T., Kumar, D., Gupta, S. (2014). Mining Students' Data for Prediction Performance. // International Conference on Advanced Computing and Communication Technologies, ACCT, 255–62. https://doi.org/10.1109/ACCT.2014.105.

Piatetsky-Shapiro, G. (2018). Will GDPR Make Machine Learning Illegal? // KDnuggets. https://www.kdnuggets.com/2018/03/gdpr-machine-learning-illegal.html

Xing, W. (2019). Exploring the Influences of MOOC Design Features on Student Performance and Persistence. // Distance Education 40, 1, 98-113. https://doi.org/10.1080/01587919.2018.1553560

Yukselturk, E., Ozekes, S., Türel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. // European Journal of Open, Distance and E-Learning 17, 1, 118-133. https://doi.org/10.2478/eurodl-2014-0008

Zhang, M., Zhu, J., Wang, Z., Chen, Y. (2019). Providing Personalized Learning Guidance in MOOCs by Multi-Source Data Analysis. // World Wide Web 22, 3, 1189-1219. https://doi.org/10.1007/s11280-018-0559-0

https://doi.org/10.17234/INFUTURE.2019.15

Preliminary communication

Music Information Seeking Behaviour among the Students of Humanities and Social Sciences at the University of Osijek

Darko Lacović
Faculty of Humanities and Social Sciences, Josip Juraj Strosmayer University of Osijek, Croatia
dlacovic@ffos.hr

Ivona Palko
Faculty of Humanities and Social Sciences, Josip Juraj Strosmayer University of Osijek, Croatia
ipalko@ffos.hr

Lana Horvatić
Faculty of Humanities and Social Sciences, Josip Juraj Strosmayer University of Osijek, Croatia
lhorvatic@ffos.hr

Summary

The aim of this paper is to give a short overview of some available studies about music information seeking and to present the research that identified music information seeking behaviour among university students. Quantitative methodology was used in the research. An online questionnaire was completed by the students from the Faculty of Humanities and Social Sciences at the University of Osijek (Croatia). The results of the research show that the majority of the respondents search for music information on the Internet (99.2%), and that they listen to music on smartphones (97.5%) or computers (90.8%), using the YouTube application (98.3%) several times per day (81.7%) on average. Most of the students possess their own personal music collection (72.5%). More than one-half of the respondents search for music by the performers (65%) or titles (63.33%) and prefer rock (60.8%) and pop (55%) music genres. Students indicate that the main reasons for listening to music are entertainment (93.3%) and decreasing boredom or loneliness (78.3%).

Key words: music information seeking, university students, Faculty of Humanities and Social Sciences, Osijek (Croatia)

Introduction

Since the beginning of the 21st century, many user studies have been conducted about music information needs, as well as the seeking, searching, browsing and retrieval of music information (for example Lee, Downie, 2004; Weigl, Guastavino, 2011; Hu, Kando, 2017), as well as storage or organization of music collections and music sharing over social media (such as Brinegar, Capra, 2011; Liikkanen, Salovaara, 2015; Lee, Cho, Kim, 2016). Although models of information behaviour are not mentioned in these studies since they are not directly related to the music information seeking, some general models, such as Wilson's models from 1981 and 1997 and Krikelas's (1983) model, can contribute to understanding of the context in which users are searching for music (for example, their motivation or factors for music information needs).

In this paper, authors will present some available studies on music information searching, as well as the research in which they have explored music information seeking behaviour of the students at the Faculty of Humanities and Social Sciences from the University of Osijek (Croatia). This study is rare in that the participants consist exclusively of university students.

Literature review

There are many studies in which authors examined music information seeking. Lee and Cunningham (2013) analysed the content of 198 user studies and found that in the majority of these studies, authors used various methods (qualitative and quantitative) on a smaller number of participants (in 45% of the studies from their sample, there were twenty people or fewer). For example, Lee and Downie (2004), in their online survey, revealed that the music information seeking of respondents was affected by recommendations or reviews from other people.

Some authors emphasized a user-centred approach for improving music information retrieval (MIR) systems in terms of recommendation tools or visualization interfaces and organization of collection (Schedl, Hauger, 2015; Lee, Price, 2016). Hu and Kando (2017) conducted an experiment on 51 participants who were engaged in MIR. According to their results, task complexity, user background, system affordances, task uncertainty and enjoyability were factors that led to difficulties in music searching. Therefore, these authors concluded these elements should be considered in designing better MIR systems. Moreover, Huron (2000), Dannenberg (2001) and Hofmann-Engl (2001) discussed some music perception models regarding MIR system design, while Schedl, Hauger and Schnitzer (2012) and Zhang et al. (2012) proposed models of serendipity in music retrieval systems.

Other authors were interested in music collections that users own and listen to when studying the process of their music information seeking behaviour. In the study conducted by Cunningham, Reeves and Britland (2003), results showed that the respondents were searching for and browsing popular music in physical collections at record stores or public libraries more than on the Internet. According to the research conducted by Brinegar and Capra (2011) among 184 music users, most of the respondents (75%) possess their digital music collection, but they also frequently transfer their music on external hard drives or optical mediums. Lee, Cho and Kim (2016), in their larger quantitative survey, revealed that 76.3% of the respondents owned a digital music collection, while 49.0% had a physical music collection organized mostly by artist (35.8% physical vs. 59.8% digital). 76.5% of respondents organized their collection using some music management software. Lee, Cho and Kim (2016) also discovered that more than half of respondents (74.5% from 2004 and 66.5% from 2012 survey) were avid listeners of music, while around one-third of them were casual listeners (21.3% in 2004 and 35.5% in 2012). The respondents reported that their preferred music genres were rock (18.0% in 2004), alternative (12.6% in 2004 and 36.2% from 2012) and blues (19.6% in 2012). Most of the respondents indicated that they listened to music on computers (98%); for entertainment (94.8% in 2004, 98.4% in 2012); to build their collection (89% in 2004, 85.4% in 2012); to search for online music multimedia (95.1%); to search the title of the works (91.1% in 2004, 92.1% in 2012); to find artist information (76.8% in 2004, 82.7% in 2012); to stream music or online radio (77.6% in 2004, 96.6% in 2012); to read any kind of music information (86.3% in 2004, 90.0% in 2012); to purchase and download music files (83.1% in 2012); and to visit online music stores (82.8% in 2012). The majority of the respondents (80.9% from the first survey and 82.8% from the second survey) consulted family or friends when they searched for music information. Bahanovich and Collopy (2009), in their large quantitative study, also revealed that respondents mostly listened to music on their computers every day (68%) for the purpose of entertainment.

Many authors found that people use music streaming services such as YouTube, Spotify or Pandora to listen to and share music online (Swanson, 2013; Cesareo, Pastore, 2014; Nguyen, Dejean, Moreau, 2014; Richardson, 2014; Hagen, 2015; Liikkanen, Salovaara, 2015). For instance, Liikkanen and Salovaara (2015) discovered that queries in YouTube searches were mostly related to music and that people often posted comments, voted or shared music. Lee, Cho and Kim (2016) found that 82.2% of the respondents searched for music they heard on streaming services, while 21.1% of the respondents used music-related apps (such as Pandora or Spotify) almost every day (13.2%), or a few times per week (7.9%). Lee et al. (2017), in their interviews (of 20 adults and 20 teen users), investigated music information behaviour in cloud-based systems and revealed that the most popular cloud services among respondents were Google Play Music, Apple iCloud, Amazon Cloud, Google Drive and Dropbox, which they mostly used for listening to their collections. Furthermore, participants also used streaming sites such as Spotify, Pandora and YouTube for discovering and listening to music for different purposes.

Research methodology

This study used quantitative methodology. A survey in the form of an online questionnaire was conducted in December 2018 at the Faculty of Humanities and Social Sciences (University of Osijek, Croatia). A total of 120 undergraduate and graduate students participated in the study. The research questions were:

- What kind of music do students search for and listen to in order to fulfil their information needs?
- What devices and applications do students use for listening to and organizing music?
- How do students search for and find music information online?

The questionnaire was made in Google Forms, posted on Facebook and sent to some participants' e-mail addresses through a Moodle online learning system. It was anonymous and consisted of sixteen mostly closed-ended questions, but the respondents also had the possibility to provide their own answers. The survey collected demographic data about respondents; their music preferences; devices and applications on which they listen to music; their personal music collections; frequency of listening to music; criteria for music searching; reasons for listening to music; and attitudes about music information.

Findings and discussion

More female (71.7%) than male students (28.3%) participated in the research. The majority of the respondents were 19 and 21 years old (20.8%), followed by students at the age of 22 (16.7%), 20 years old (12.5%), 23 years old (10%) and 18 years old (8.3%). Students between 25 and 29 years old were represented the least (2.5%). Less than one-half of the respondents were students of library and information science (47.5%); students of English language and literature (20.8%); students of Croatian language and literature (19.2%); students of history (17.5%); students of education (11.7%); and students of psychology (10%).

Regarding the sources of music information, as expected, most of the respondents indicated that they search for music information on the Internet (99.2%), which is in line with the results that Lee, Cho and Kim (2016) obtained and contrary to the findings of the study conducted by Cunningham, Reeves and Britland (2003). Around one-half of the students search for information about music from friends or colleagues (59.2%), while a small number of them look for this information in journals and magazines (10%), books (9.2%) or in libraries (1.7%). It can be observed that music has a socialization aspect for the respondents. Results are shown in Table 1.

Table 1. Sources of music information searching

Information sources	Percentage
Internet	99.2%
friends or colleagues	59.2%
journals and magazines	10%
books	9.2%
libraries	1.7%

In relation to music genres, more than one-half of the respondents stated that they listen to rock (60.8%) and pop (55%), while around one-third of the students prefer alternative (38.3%), hip hop or rap (31.7%), dance (30.8%) and old-time, R&B or soul (30%). This can be understood taking into account the presence of popular music throughout mass media. Around or less than one-quarter of the respondents are interested in classical music (27.5%); folk (25.8%); blues (22.5%); hard rock or heavy metal (20%); jazz (19.2%); reggae (17.5%); Latino (15%); or new age (14.2%). 8.8% of the respondents indicated that they listen to some other music types such as trash, techno, house, punk, experimental music and others, and surprisingly, 8.3% of the students prefer opera (Table 2). Similar results about music genres preferred by the respondents can be found in the research conducted by Lee, Cho and Kim (2016).

Table 2. Preferred music genres

Music genres	Percentage
rock	60.8%
pop	55%
alternative	38.3%
hip hop or rap	31.7%
dance	30.8%
old-time, R&B or soul	30%
classical music	27.5%

folk	25.8%
blues	22.5%
hard rock or heavy metal	20%
jazz	19.2%
reggae	17.5%
Latino	15%
new age	14.2%
other	8.8%
opera	8.3%

When asked about devices on which students listen to music, the majority of them answered that they use smartphones (97.5%) and computers (90.8%). Similar findings were obtained by Bahanovich and Collopy (2009) in their study. Less than half of the respondents listen to music on the radio (46.7%), and an even smaller number uses tablets (6.7%), mp3s or iPods (4.2%) for listening to music. Interestingly, 1.6% of the students listen to music on gramophones (Table 3). As expected, most of the respondents use the YouTube platform (98.3%) for listening to music – probably because it is available for free. This was also confirmed in the study conducted by Liikkanen and Salovaara (2015). Less than one-quarter of the students use music applications such as Soundcloud (24.2%), Spotify (17.5%) and Deezer (9.2%). These results are also in line with those of Swanson (2013), Richardson (2014), Lee, Cho and Kim (2016), and Lee et al. (2017).

Table 3. Devices for listening to music

Devices	Percentage
smartphone	97.5%
computer	90.8%
radio	46.7%
tablet	6.7%
mp3 or iPod	4.2%
gramophone	1.6%

Further results showed that 72.5% of the respondents have their own personal music collection and 27.5% of them do not possess one. Similar results are revealed in studies by Brinegar and Capra (2011) and Lee, Cho and Kim (2016). As can be seen from Table 4, students organize their music collections according to the following criteria: music artist (47.8%); frequency of listening to music (26.7%); album (23.3%); genres (21.1%); and domestic or foreign music (15.6%). Only 9.9% of the students organize their music collection in some other way (for example, according to year, playlist, personal mood, etc.). Most of the respondents indicated that they listen to music several times per day (81.7%), and a small number of them listen to music several times per week (17.5%) or several times per month (0.8%). It is obvious that students frequently listen to music, as also revealed by Lee, Cho and Kim (2016) in their study.

Table 4. Organization of music collection

Organization criteria	Percentage
music artist	47.8%
frequency of listening to music	26.7%
album	23.3%
genres	21.1%
domestic or foreign music	15.6%
other	9.9%

In the question about reasons for listening to music (Table 5), the majority of the respondents answered that they listen to music for fun and entertainment (93.3%) – which was also confirmed in the study conducted by Bahanovich and Collopy (2009) – as well as for decreasing boredom or loneliness (78.3%). Around one-half of the students listen to music when they want to pass the time while they wait for something (51.7%), or they use music as a background for exercise, such as running or other fitness activities (50.8%). A smaller amount (9%) of respondents listen to music for other reasons (for example, when they play games, as background noise or without a specific reason).

It can be concluded that students actually use music to relax themselves from their academic obligations.

According to research results, 80% of the students do not know how to play a musical instrument and 20% do know how to play a musical instrument (15.8% of the students play string instruments, 3.3% play drums and electronic instruments, and 1.6% play percussion). Similar findings are identified by Lee, Cho and Kim (2016).

Table 5. Reasons for listening to music

Reasons	Percentage
fun and entertainment	93.3%
decreasing boredom or loneliness	78.3%
to pass the time while waiting	51.7%
background for exercise	50.8%
other	9%

In relation to probability of searching for songs or music information by different elements, more than one-half of the respondents reported that they would certainly search for music by title (63.3%) and by performers (65%); or, as expected, that they would never search for music by publisher (52.5%). More than one-third of the students would rarely search for music according to appearance of the song (39.1%), would probably search by popularity of the song (36.6%) and would never (37.5%) or rarely (35.8%) search by country. Less than one-third of the respondents would probably search for music by title (28.3%); rarely by publisher (30.8%); probably by performer; rarely by music popularity (29.1%); and probably (25.8%) or never (25%) by appearance of the song. These findings are somewhat different than those from the study by Lee, Cho and Kim (2016). Other results are provided in Table 6.

Table 6. Probability of music searching by different elements

Elements	Never	Rarely	Probably	Certainly
title	2.5%	5.8%	28.3%	63.3%
publisher	52.5%	30.8%	10.8%	5.8%
performer	2.5%	5.8%	29.1%	65%
appearance of the song	25%	39.1%	25.8%	10%
popularity	17.5%	29.1%	36.6%	16.6%
country	37.5%	35.8%	18.3%	8.3%

Regarding the frequency of music information searching over the past month, more than half of the students indicated that they have never searched for information about music publishers (67.5%). Around one-third of the respondents answered that they searched for song titles several times per month (33.3%) and song publication date once per month (30.8%), but, also, that they have never searched for information about music genres (30.8%), albums (37.5%), date of publishing (39.1%) or song remixes (38.3%). Around one-quarter of the students searched for the following information: song performers several times per month (27.5%) and once per month or several times per week (20.8%); song titles several times per week (21.6%); music genres once (25.8%) or several times per month (20.8%); album information several times per month (22.5%) and once per month (19.1%); song lyrics several times per month (26.6%), several times per week (23.3%) and once per week (19.16%); and music remixes once per month (25%) or several times per month (21.6%). Other results are presented in Table 7.

Table 7. Frequency of music information searching

Music information	Never	Once per month	Several times per month	Once per week	Several times per week	Every day
song title	10.8%	10.8%	33.3%	15%	21.6%	8.3%
song performer	10.8%	20.8%	27.5%	10.8%	20.8%	9.1%
song publisher	67.5%	15.8%	10%	5.8%	2.5%	0.8%
genres	30.8%	25.8%	20.8%	13.3%	6.6%	2.5%

album	37.5%	19.1%	22.5%	7.5%	10%	5.8%
song lyrics	5%	10%	26.6%	19.1%	23.3%	15.8%
song publication date	39.1%	30.8%	17.5%	6.6%	2.5%	5.8%
song remixes	38.3%	25%	21.6%	8.3%	5%	1.6%

In the final question, respondents assessed some statements on music information and sources on the scale of 1 to 5 (1 = strongly disagree, 5 = strongly agree). According to the results (presented in Table 8), students mostly agreed with the statement that music information is easy available online (mean = 4.6), and that they encounter music information mostly by accident from different sources (mean = 4.5). Students agreed that they are very satisfied with the music information they find, and that they mainly use music information for fun and entertainment (mean = 4.1). The respondents agreed slightly less with the statement that they rarely find music information in print sources (mean = 3.3), and they agreed least with the statement that they need a lot of time to find music information that they are interested in (mean = 1.9). These results are in line with those previously obtained which are related to searching for music information.

Table 8. Agreement to the statements on music information and sources

Statements	Mean
Music information is easy available online	4.6
I encounter music information mostly by accident in different sources	4.5
I am very satisfied with the music information I find	4.1
I mainly use music information for fun and entertainment	4.1
I rarely find music information in print sources	3.3
I need a lot of time to find music information I am interested in	1.9

Conclusion

The findings of this research show that music information seeking is an important part of students' life activities. It is not surprising that the respondents often search for and find music information online, since music information is available (for free) through different applications. As expected, students mostly search for songs on smartphones by performers and titles for the purpose of entertainment. On the other hand, it is interesting that around 20% of the respondents look for music information in print sources (books, journals and libraries).

This small and preliminary study is one of the first to deal with music information seeking behaviour among university students in Croatia. It could be seen as a starting point for further investigation, in which a larger number of the respondents from different scientific areas or ages could participate, taking into account some specific aspects of their music information behaviour (for example, a narrow context in which music is retrieved, user requirements of the music information systems, etc.). Moreover, this research can serve as a framework for future studies which could also explore, among many other subjects, the impact of user experience for improving the design of some music applications. It is also recommended to use qualitative methods in order to gain deeper insight into music information behaviour of the users.

References

Bahanovich, D., Collopy, D. (2009). Music experience and behaviour in young people. London: UKMusic. http://infojustice.org/wp-content/uploads/2011/02/Music-Behavior-in-Young-People-Bahanovich-2009.pdf (24.7.2019)

Brinegar, J., Capra, R. (2011). Managing music across multiple devices and computers. // Proceedings of the iConference 2011. Seattle, WA: ACM, 489-495

Cesareo, L., Pastore, A. (2014). Consumers' attitude and behavior towards online music piracy and subscription-based services. // The Journal of Consumer Marketing 31, 6-7, 515-525

Cunningham, S. J., Reeves, N., Britland, M. (2003). An ethnographic study of music information seeking: Implications for the design of a music digital library. // Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries. Houston, TX: ACM/IEEE, 5-16

Dannenberg, R. B. (2011). Music information retrieval as music understanding. // Second International Symposium on Music Information Retrieval. Bloomington, IN: ISMIR, 139-142

Hagen, A. (2015). The playlist experience: Personal playlists in music streaming services. // Popular Music and Society 38, 5, 625-645

Hofmann-Engl, L. (2011). Towards a cognitive model of melodic similarity. // Second International Symposium on Music Information Retrieval. Bloomington, IN: ISMIR, 143-151

Huron, D. (2000). Perceptual and cognitive applications in music information retrieval. // 1st International Society for Music Information Retrieval Conference. Plymouth, MA: ISMIR, 2000. https://www.researchgate.net/publication/220723259_Perceptual_and_Cognitive_Applications_in_Music_Information_Retrieval (30.10.2019.)

Hu, X., Kando, N. (2017). Task complexity and difficulty in music information retrieval. // Journal of the Association for Information Science and Technology 68, 7, 1711-1723

Krikelas, J. (1983). Information-seeking behavior: Patterns and concepts. // Drexel Library Quarterly 19, 2, 5-20

Lee, J. H., Cho, H., Kim, Y-S. (2016). Users' music information needs and behaviors: design implications for music information retrieval systems. // Journal of the Association for Information Science and Technology 67, 6, 1301-1330

Lee, J. H., Cunningham, S. J. (2013). Toward and understanding of the history and impact of user studies in music information retrieval. // Journal of Intelligent Information Systems 41, 3, 499-521

Lee, J. H., Downie, J. S. (2004). Survey of music information needs, uses, and seeking behaviours: Preliminary findings. // Proceedings of the 5th International Conference on Music Information Retrieval. Barcelona: ISMIR, 441-446

Lee, J. H., Price, R. (2016). User experience with commercial music services: An empirical exploration. // Journal of the Association for Information Science and Technology 67, 4, 800-811

Lee, J. H., Wishkoski, R., Aase, L., Meas, P., Hubbles, C. (2017). Understanding users of cloud music services: Selection factors, management and access behavior, and perceptions. // Journal of American Society for Information Science and Technology 68, 5, 1186-1200

Liikkanen, L. A., Salovaara, A. (2015). Music on YouTube: User engagement with traditional, user-appropriated and derivative videos. // Computers in Human Behavior 50, 108-124

Nguyen, G., Dejean, D., Moreau, S. (2014). On the complementarity between online and offline music consumption: The case of free streaming. // Journal of Cultural Economics 38, 4, 315-330

Richardson, J. H. (2014). The Spotify paradox: How the creation of a compulsory license scheme for streaming on-demand music platforms can save the music industry. // UCLA Entertainment Law Review 2, 1, 45-74. https://escholarship.org/content/qt7n4322vm/qt7n4322vm.pdf?t=nl5jht (24.7.2019)

Schedl, M., Hauger, D., Schnitzer, D. (2012). A model for serendipitous music retrieval. // Proceedings of the 2nd International Workshop on Context-awareness in Retrieval and Recommendation (CaRR 2012). Lisbon: ACM, 10-13

Schedl, M., Hauger, D. (2015). Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 947-950

Swanson, K. (2013). A case study on Spotify: Exploring perceptions of the music streaming service. // Journal of the Music & Entertainment Industry Educators Association 13, 1, 207-230. http://www.meiea.org/resources/Journal/Vol.13/Swanson-MEIEA_Journal_vol_13_no_1_2013-p207.pdf (24.7.2019)

Weigl, D. M., Guastavino, C. (2011). User studies in the music information retrieval literature. // Proceedings of the 12th International Society for Music Information Retrieval Conference. Miami, Florida: ISMIR, 335-340

Wilson, T. D. (1997). Information behaviour: an interdisciplinary perspective. // Information Processing & Management 33, 4, 551-572

Wilson, T. D. (1981). On user studies and information needs. // Journal of Documentation 37, 1, 3-15

Zhang, Y. C., Seaghdha, D. O., Quercia, D., Jambor, T. (2012). Auralist: Introducing serendipity into music recommendation. // Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM). Seattle, WA: ACM, 13-22

https://doi.org/10.17234/INFUTURE.2019.16

Original scientific paper

Automated Phonetic Transcription of Croatian Folklore Genres Using Supervised Machine Learning

Nikola Bakarić
University of Applied Sciences, Velika Gorica, Croatia
nbakaric@gmail.com

Davor Nikolić
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
dnikoli@ffzg.hr

Summary

This paper aims to detect the possibilities of automatic text transcription for the purpose of preparing a corpus for further natural language processing analysis. The corpus contains various Croatian folklore genres. The transcription goal is to have one character represent one phoneme and remove spaces between accentuated and non-accentuated words. This knowledge independent system is trained using supervised learning methods and applied to the rest of the corpus using classifiers such as the naïve Bayes, k-nearest neighbour, support vector machine and others. The results are compared to a human-annotated sample to determine accuracy.

Key words: text transcription, automation, natural language processing, supervised learning, Croatian folklore genres

Introduction

This paper is a part of a larger research effort which deals with automated classification of Croatian oral literature. In order to approach the problem, the examined corpus of oral literature needed to be prepared, normalised and transcribed to a certain degree. One of the preparatory steps is the syllabification of the corpus of Croatian oral literature. The transcription is an important condition for correct syllabification with regards to pronunciation. The goal is to have one character represent one phoneme. Although Croatian language spelling is mostly phonetic (Pravopis, 2019), there are instances where pronunciation differs significantly. A good portion of such instances can be solved using simple transcription rules, the simplest being the transcription of digraphs lj, nj and dž to ḷ, ń and ǯ respectively. There are phenomena, like the yat reflex, which are not so straightforward (“jat,” 2019) from a computational perspective and require a more complex approach. Apart from digraphs and the yat reflex, there is phonetic assimilation which occurs in pronunciation when two phonemes form a new sound when spoken together (Yule, 2002). However, the most numerous differences in pronunciation and spelling is the removal of pauses/spaces between accentuated and non-accentuated words, enclitics and proclitics. This is topic of the research presented in this paper.

The problem

As mentioned before, Croatian spelling is phonetic to a very high degree, however there are situations where assimilation and other pronunciation phenomena occur. One of the most common phenomena is the fusion between non-accentuated words (enclitics and proclitics) and their accentuated counterparts in pronunciation. This phenomenon could be described by a simpler rule-based model using their definitions. Enclitics in Croatian language are non-accentuated, present tense and aorist forms of the verbs biti and htjeti, non-accentuated forms of pronouns and the word li (Enklitika, 2019). Proclitics can be monosyllabic, some disyllabic and trisyllabic propositions, conjunctions and particles (Proklitika, 2019). However, both types are ambiguous and can be mistaken for other word types with different functions in pronunciation and spelling. Therefore, supervised machine learning was selected as a more robust and flexible approach to the problem.

Methodology

In order to conduct this preliminary research, a small corpus was prepared. The corpus consisted of 69 blessings and 75 tongue twisters, altogether 1167 words with 1026 occurrences of the space character. It is a part of a larger corpus of Croatian folklore genres collected in the manuscript archives of the Chair of Croatian oral literature at the Department of Croatian Language and Literature, Faculty of Humanities and Social Sciences at the University of Zagreb. A copy of the corpus was further prepared by an expert human annotator who manually marked the instances of the space character which are omitted in pronunciation. At this point, the two copies differed only in the deleted space characters. Table 1 shows several examples of the original and annotated corpus.

Table 1. Examples of annotation

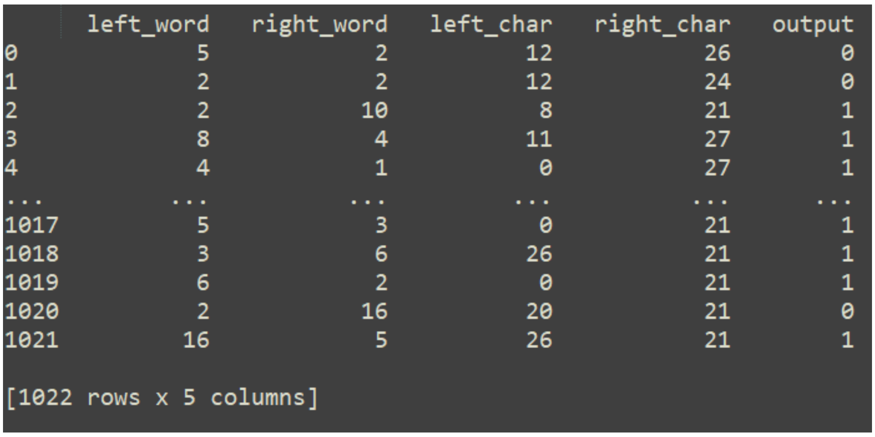
Original text	**Annotated**
Više ti Bog dao nego što ima zvjezda na nebu.	Višeti Bog dao negoštoima zvjezda nanebu.
Na štriku se suši škotski šosić.	Naštrikuse suši škotski šosić.
Sobzirom na obzir da je moj obzir obzirniji otvog, tvoj obzir kao obzir ne dolazi u obzir.	Sobzirom naobzir dajemoj obzir obzirniji otvog, tvoj obzir kaoobzir nedolazi uobzir.
Prst u pitu, prst u tikvu.	Prst upitu, prst utikvu.
Moja fajfa, stara fajfa, moja fajfa, dobra fajfa. Moja fajfa tak dobre fajfa da ni jedna fajfa ne fajfa tak dobre kak moja fajfa fajfa.	Moja fajfa, stara fajfa, moja fajfa, dobra fajfa. Moja fajfa tak dobre fajfa danijedna fajfa nefajfa tak dobre kakmoja fajfa fajfa.

Source: Authors

The feature selection was based on experience and observation, and two groups of two features were selected. As the non-accentuated words are usually shorter in character length than accentuated words, we selected word length as the first feature group. It consisted of two features, length of the word to the immediate left of the space character and length of the word to the immediate right of the space character. The dataset for word length showed a positively skewed normal distribution of length for both left and right words. The left word length variable set consisted of 14 categories with 4.4 average word length. The right word length variable set consisted of 16 categories with 4.8 average word length.

The second group observed the characters in the immediate left and right of the space. The characters were numerically encoded replacing the letter ‘a’ with 0, ‘b’ with 1, ‘c’ with 2 and so on. Left character variable set consisted of 29 categories with more than half (524) occurrences belonging to the vowels ‘a’, ‘e’ and ‘i’, which is to be expected for Croatian language. The right character variable set was comprised of 25 categories. Here more than half (520) occurrences belonged to consonants such as ‘p’, ‘b’, ‘s’, ‘d’ and ‘k’. The character variable set does not seem to follow normal distribution.

The features were assembled into a sequence of lists (vectors) to which a final value was added, a 0 or 1, depending on the existence of the space character in the annotated corpus. All features and the target value were extracted using custom Python scripts and organised as seen in Figure 1 using the Pandas module for Pyhton (McKinney, 2010).



```
      left_word  right_word  left_char  right_char  output
0             5           2         12          26       0
1             2           2         12          24       0
2             2          10          8          21       1
3             8           4         11          27       1
4             4           1          0          27       1
...         ...         ...        ...         ...     ...
1017          5           3          0          21       1
1018          3           6         26          21       1
1019          6           2          0          21       1
1020          2          16         20          21       0
1021         16           5         26          21       1

[1022 rows x 5 columns]
```

Figure 1. Dataset structure, source: Authors.

The data containing feature vectors and target values was processed using several classification algorithms using the scikit-learn module for Python (Pedregosa et al., 2011). The classification problem was binary as the algorithms had to place each instance of the space character into one of two groups, either deleted or not deleted. The classification algorithms used were:

- Naive Bayes (Gaussian/normal, Multinomial and Complement)
- Support vector machines (Support vector classifier)
- K-nearest neighbour (Nearest centroid classifier)
- Neural network (Multi-layer Perceptron)

The naïve Bayes classifier is one of the simplest, yet most effective classifiers in machine learning tasks (Zhang, 2004), especially in natural language processing. Main feature of this classifier is that it ignores any conditional dependence between observed features thus making it simple yet robust. It has proved to be very successful in many machine learning applications. Several variations of the classifier were tested as its performance depends on the distribution of the input variables.

The scikit-learn tutorial (“Support Vector Machines,” 2019) describes the Support vector machines as a set of supervised learning methods used in classification and regression. The support vector classifier module was used. In short, it maps the feature vectors into a model and then finds the margin between two classes. In our case, it used the training set to create a model and calculate a margin. It then mapped the test set vectors to either side of the margin, thus classifying it to the delete space or do-not-delete space group.

The k-nearest neighbour classifier is another simple yet effective method (Goldberger, Roweis, Hinton, & Salakhutdinov, 2005) which has several variations. The one used here is the nearest centroid classifier. The method uses the training set to evaluate the nearest neighbours of the test set vector thus determining its class.

Classification using neural networks is slightly more complex than the previous methods. The application of a Multi-layer Perceptron requires additional preparation and fine adjustment of the classifier parameters (LeCun, Bottou, Orr, Müller, 1998). MLP is a non-linear supervised learning algorithm described as a deep neural network with several (at least 3) layers. It learns a function using the training dataset and the provided dimension parameters which passes values along the network nodes (Scikit-Learn Developers, 2018).

All classifiers were applied alongside k-fold cross validation, a method which prevents overfitting in a supervised machine learning environment by separating the dataset into k sections which are then alternated as the training and test sets (Hastie, Tibshirani, Friedman, 2009). The dataset was separated into 10 sections for cross-validation. Each set was used as a test set while the remaining nine sets were used for training. The results presented in the following chapter are the averages of these 10 classification iterations.

Results

One of the aims of this research was to establish the best features for this particular classification experiment. Therefore, the classification algorithms in combination with three different feature sets were applied. The first set of results, presented in Table 2, includes only two features, length of the word to the immediate left of the space character and length of the word to the immediate right of the space character.

Table 2. World lengths as features

Classifier	Accuracy and standard deviation
Multi-layer Perceptron	0.85 (+/- 0.08)
naïve Bayes (Gaussian)	0.84 (+/- 0.07)
Support vector classifier	0.83 (+/- 0.09)
k-NN (Nearest centroid)	0.78 (+/- 0.15)
naïve Bayes (multinomial)	0.75 (+/- 0.09)
naïve Bayes (complement)	0.64 (+/- 0.15)

Source: Authors.

The second set of results, presented in Table 3, includes only the numerically encoded characters left and right of the space character. A general drop in accuracy when compared to the results to Table 2 could be connected to the fact that the set does not seem to follow normal distribution.

Table 3. Characters as features

Classifier	Accuracy and standard deviation
Support vector classifier	0.75 (+/- 0.00)
naïve Bayes (Gaussian)	0.75 (+/- 0.00)
Multi-layer Perceptron	0.75 (+/- 0.01)
naïve Bayes (multinomial)	0.61 (+/- 0.17)
naïve Bayes (complement)	0.56 (+/- 0.18)
k-NN (Nearest centroid)	0.55 (+/- 0.17)

Source: Authors.

Table 4 presents the third set of results which contains a combination of all four features. Here the results are similar to Table 2 where we observed only word length. It could be argued that characters as features do not contribute to, and in some cases decrease the classification accuracy. This indicates that their information value regarding the phenomenon is lower when compared to word length. However, the feature could be simplified by reducing the number of categories, perhaps only to vowels, consonants and punctuation. Another possibility is to scale and weight the features and try to improve accuracy.

Table 4. Word length and characters as features

Classifier	Accuracy and standard deviation
Multi-layer Perceptron	0.85 (+/- 0.09)
Support vector classifier	0.81 (+/- 0.07)
naïve Bayes (Gaussian)	0.80 (+/- 0.09)
naïve Bayes (multinomial)	0.73 (+/- 0.08)
naïve Bayes (complement)	0.65 (+/- 0.14)
k-NN (Nearest centroid)	0.58 (+/- 0.18)

Source: Authors.

Regarding the classifiers, the results show that some handle more features better than others, while some prefer certain types of feature value distributions. The Multi-layer perceptron neural network classifier has proven the most accurate when observing word lengths (Table 2) and the combination of word length and character quality. It is worth noting that initially the naïve Bayes (Gaussian) was the top scoring classifier in the word length environment until we increased the size of hidden layers in the Multi-layer perceptron from (5, 3) to (6, 4). The preparation of the dataset with regards to feature scaling seems to be very important and this leaves room for improvement with the adjustment of parameters for all tested classifiers

Conclusion

Apart from preparing corpuses for academic investigation, which is the main motive of the authors, this preliminary research effort has shown that there are a lot of interesting topics in text to speech translation. This is especially true for small languages which do not have the vast amounts of available lexical data which is the basis of most TTS systems (Mana, Massimino, & Pacchiotti, 2001). While the results of certain classifier models are promising, there is room for improvement in dataset preparation, feature selection and tweaking classifier parameters. In order to design a universal model, the corpus should be increased and expanded to include general language.

However, the models relatively high accuracy while using certain classifiers shows promise. The authors plan to develop it further and use it in preparing oral literature corpuses for further analysis. Perhaps a derivation of it will someday be included into a Croatian text-to-speech system.

References

Enklitika. (2019). Hrvatska enciklopedija, mrežno izdanje. Retrieved from http://www.enciklopedija.hr/Natuknica.aspx?ID=17990

Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R. (2005). Neighbourhood Components Analysis. // Advances in Neural Information Processing Systems 17, 513-520. Retrieved from https://cs.nyu.edu/~roweis/papers/ncanips.pdf

Hastie, T., Tibshirani, R., Friedman, J. (2009). Model Assessment and Selection. // The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 219-257

Jat. (2019). Hrvatska enciklopedija, mrežno izdanje. http://www.enciklopedija.hr/natuknica.aspx?ID=28821

LeCun, Y., Bottou, L., Orr, G., Müller, K. (1998). Efficient BackProp. // Neural Networks: Tricks of the Trade 1998. https://doi.org/10.1192/bjp.112.483.211-a

Mana, F., Massimino, P., Pacchiotti, A. (2001). Using machine learning techniques for grapheme to phoneme transcription. // EUROSPEECH 2001 - SCANDINAVIA - 7th European Conference on Speech Communication and Technology. https://pdfs.semanticscholar.org/ce0e/7ca7c745c2a65b6f3ac7be5df5a8e72065fe.pdf

McKinney, W. (2010). Data Structures for Statistical Computing in Python. // Proceedings of the 9th Python in Science Conference, 51-56. http://conference.scipy.org/proceedings/scipy2010/mckinney.html

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R.,… Duchesnay, Fré. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830. http://scikit-learn.sourceforge.net

Pravopis. (2019). Hrvatska enciklopedija, mrežno izdanje. http://www.enciklopedija.hr/natuknica.aspx?id=50013

Proklitika. (2019). Hrvatska enciklopedija, mrežno izdanje. http://www.enciklopedija.hr/natuknica.aspx?id=50588

Scikit-Learn Developers. (2018). Neural network models (supervised). Retrieved September 10, 2019. https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Support Vector Machines. (2019). https://scikit-learn.org/stable/modules/svm.html (10.9.2019)

Yule, G. (2002). The study of languge. Cambridge: Cambridge University Press.

Zhang, H. (2004). The optimality of naive Bayes. In AAAI-04. https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf

https://doi.org/10.17234/INFUTURE.2019.17

Preliminary communication

Arabic Speakers as Croatian Language Learners
Electronic Educational Games as a Support for Learning

Maja Matijević
Institute of Croatian Language and Linguistics, Zagreb, Croatia
mmatijevic@ihjj.hr

Josip Mihaljević
Institute of Croatian Language and Linguistics, Zagreb, Croatia
jmihalj@ihjj.hr

Summary
In the last decade, Croatia has been the transit or receiving country for many refugees who come from the Middle East and Africa and who mostly speak Arabic as their mother tongue. Native speakers of Arabic are faced with many difficulties in the process of learning Croatian. Croatian is written from left to right, and Arabic from right to left. Phonological inventories of Croatian and Arabic are very different, vowels are generally not written in Arabic (or more precisely they are not written with graphemes), grammar is very different, etc. All this demotivates students at the initial stage of learning. Based on the experience of one of the authors with teaching Croatian to asylees and asylum seekers who mostly came to Croatia from the war-ridden Middle East and African countries in the last few years and who are mainly speakers of Arabic, the authors started developing educational games which could facilitate their initial steps in learning Croatian. Games focus on the acquisition of the Latin script and the Croatian phonological inventory (games in which the player writes the pronounced sound, hangman with or without the picture of the required word, etc.). As the difference between capital and small letters does not exist in Arabic, games that tackle certain orthographic issues are also developed. All games have explanations in Arabic that prevent ambiguity and show differences between Croatian and Arabic. The learning material also emphasizes the difference between the phonemes p and b and introduces as many vowel games as possible. The goal is to reduce the beginner's fear of language learning and motivate the learners.

Key words: Arabic language, Croatian as a foreign language, refugees, educational games, game development, language learning, SLA

Introduction

For the great majority of people who live in Croatia, Croatian is their mother tongue, the common language, and the language which is being studied in schools. However, many people who come to Croatia do not speak Croatian and they want to stay in Croatia for many reasons. One of the reasons for immigration is war in the refugee's homeland and their fear of persecution because of their race, religion, nationality, etc.[1] The situation in Europe has significantly changed in the last decade. Because of the war in the Middle East and African countries, Europe has become a safe place for refugees. The migration crisis also affected Croatia, primarily as a part of the 'Balkan route' which has opened in 2015 when many refugees passed through Balkan countries on their way to Western European countries. However, Croatia was not only a transit but also a receiving country for many refugees who decided to stay in it (especially after 2013 when Croatia became a European Union member). According to the latest official data (MUP, 2019) in March 2019, there were 767 asylees in Croatia, the number of asylum seekers is not known.

[1] Based on the definition of refugee in *Convention Relating to the Status of Refugees* from 1951 (UNHCR, 1951), all migrants who have 'well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion' are called *refugees*. In this work, term *refugee* will be used as a hyperonym to *asylum seeker* and *asylee,* which will be distinguished to avoid terminological confusion. The term *asylum seeker* will refer to a refugee requesting asylum in another country and it will differ from term *asylee* which will be used for a person who officially has asylum.

When a person gets asylum in Croatia, he has the right to stay in the country, move and express his religion freely, have medical insurance, attend school, learn the language and have access to information. To have all these rights, he has to respect the Constitution of the Republic of Croatia, cooperate with institutions and try to integrate into the Croatian society (Zakon o međunarodnoj i privremenoj zaštiti, 2015). One of the first steps in the integration process is learning the language, which is both the right and the obligation of an asylee. Users on popular language learning platforms such as Memrise[2] and DuoLingo[3] have created separate quizzes for Arabic students learning English. Qian and Clark (2016) conducted a research using the Academic Search Complete Database. They analyzed 137 studies on game-based learning material. Their research shows that the influence of games on the success of learning is mostly dependent on the way they combine educational content with certain game mechanics which are successful in the entertainment game industry as well. They think that specifically designed games for learning different subjects and aiming at different groups defined by age and gender mostly work better than typical commercial and educational games. The problem is how to develop specific games for specific groups of students (e.g. native speakers of Arabic learning Croatian) because that requires programming and finding or creating appropriate graphical resources. The hypothesis of this paper is that efficient interactive content can and should be made for the native speakers of Arabic learning Croatian. The main goal is to present electronic educational games for learning Croatian specially developed to meet the needs of Arabic refugees. Games like this might be a good stimulus for all Arabic speakers in their initial stages of learning Croatian.

Croatian language courses and asylees

Refugees who come to Croatia are mostly exposed to Croatian for the first time in collective reception centers in Zagreb and Kutina where they listen to people who speak Croatian. Taking into account that language courses are expensive and generally not available to refugees, it is very common that their first steps usually depend on civil society organizations and volunteers. These volunteers are rarely educated teachers of Croatian as a foreign language, and even if they are, they are mostly not educated to meet the specific needs of a particularly vulnerable group of students such as refugees (Đurđević, Podboj, 2016: 246). Refugees could also start learning Croatian using the Croatian language Moodle course[4] or mobile applications such as Learn Croatian. Speak Croatian[5], Learn Croatian Language with Master Ling[6], Jednostavno naučite hrvatski[7], etc. Unfortunately, some Moodle courses do not use the capabilities of the system in its full extent. Some lecturers only upload Word, PDF or PowerPoint files as learning materials, which are not much different from the books as they mostly contain static texts and images. Moodle does have integrated options for creating quizzes, crosswords, hangman games, class scenarios and other interactive contents (Birkić et al., 2019: 49-54). There is even an option to implement interactive content from the H5P platform which includes creating web responsive quizzes, timelines, virtual tours, word pronunciations, tests, essays, etc. (Ibid.: 28). Organized obligatory language courses for asylees are not held regularly and they are available only to asylees and not to asylum seekers (Đurđević, Podboj, 2016: 248). The first free Croatian language course led by experienced teachers, performed by student volunteers and held not only for asylees but also for asylum seekers started in 2017 and was performed at the Faculty of Humanities and Social Sciences in Zagreb[8]. The main condition for entering the course was knowing the Latin script and basics of Croatian. After the course, students would get a B1.1. certificate if they

[2] English for Arabic speakers: www.memrise.com/course/276681/english-for-arabic-speakers-2/ (20.10.2019)

[3] Free Language Courses for Arabic Speakers: www.duolingo.com/courses/ar (20.10.2019)

[4] Moodle courses *Free Online Croatian Courses – A1.HR and A2.HR* are created by Croaticum – Centre for Croatian as a Second and Foreign Language and Centre State Office for Croats Abroad and can be accessed at page https://croaticum.ffzg.unizg.hr/?page_id=5024 since December 2018.

[5] Learn Croatian. Speak Croatian: https://play.google.com/store/apps/details?id=com.atistudios.mondly.hr&hl=en_US (26.10.2019)

[6] Learn Croatian Language with Master Ling:
https://play.google.com/store/apps/details?id=com.simyasolutions.ling.hr&hl=hr (26.10.2019)

[7] Jednostavno naučite hrvatski: https://play.google.com/store/apps/details?id=simply.learn.croatian&hl=hr (26.10.2019)

[8] Until now, three B1.1. courses for asylees and asylum seekers were held at Croaticum – Centre for Croatian as Foreign and Second Language. Teachers Jelena Cvitanušić Tvico and Ranka Đurđević mentored students Ivana Đerke, Ivana Bauk, and Maja Matijević who performed classes for 45 students (15 per a one-semester course).

passed the final test with the score of at least 61% correct answers. Even if they knew the Latin script, there were a lot of students who had problems with distinguishing some Croatian phonemes, such as u and o, p and b, etc. and with Croatian orthography (Matijević, 2018: 5-6). In 2018 the same faculty launched a course Croatian as a foreign language and service-learning for graduate students. Within that course free classes Latin literacy for asylees and asylum seekers[9] were held. The classes were designed for asylees and asylum seekers who are illiterate in the Latin script and the main goals were that after the class students recognize and distinguish Croatian sounds (heard or written), can pronounce and write Croatian sounds, know the Croatian alphabet and can hold a basic conversation in Croatian (Ocvirk, Radošević, Sammartino, 2019: 29). Teacher experiences from these courses were a starting point for creating electronic games for that group of users. Both courses also have students who had traumatic experiences which affect their learning motivation and sometimes cognitive abilities. The stressful and traumatic experience can be manifested as fatigue, lack of concentration, difficulties in remembering, somatic disorders, missing classes, etc. Moreover, the teacher should take into account the level of student's previous education and possible cultural differences that affect the learning process, like not understanding the atmosphere in the class typical for western societies (Đurđević, Podboj, 2016: 250). In addition to traumatic and stressful experience, asylees and asylum seekers might, as all students of foreign languages, have the fear of language. The fear of language is defined as a fear we feel in situations that require the use of a non-mother tongue in which we feel incompetent (Mihaljević Djigunović, 1998: 52). The fear of language might be even greater for asylum seekers because their possibility of permanent stay in Croatia depends on having a B1.1. certificate. If in addition to all this, the difference between Croatian and Arabic is taken into account, it is understandable that learning Croatian might be very challenging for Arabic native speaking refugees.

Croatian vs. Arabic

Croatian is a Slavic flective language written in the Latin script which is very different from Arabic and also from English or French, the languages that might be more familiar to some Arabic speakers[10]. The Arabic script is written from right to left, and the Latin script from left to right. The Arabic script does not differentiate between capital and small letters and uses punctuation differently from most Indo-European languages. In addition to that, Arabic speakers who write in a language that uses the Latin script tend to use a comma instead of a period (Pučko otvoreno učilište, 2018: 18). Arabic script also does not record vowels as graphemes, but adds small marks above, under or around consonants which are the root of the word (Matijević, 2018: 6). On the grammatical level, the Arabic language does not have the verb to be ('She is happy' or 'He is a teacher' would be 'She happy' or 'He teacher'). A descriptive adjective usually comes after the noun it describes ('nice girl' would be 'girl nice') and there is no neutral gender in Arabic. Also, some nouns do not have the same gender as in Croatian, so the word for table (stol) which is of male gender in Croatian, is of female gender in Arabic (Pučko otvoreno učilište, 2018: 18).

Designing learner's material

Thinking about all factors that make learning Croatian hard (experienced traumas, fear of language, unavailability of professional courses, and the difference between Arabic and Croatian), electronic educational games seem to be a good means to facilitate learning Croatian on the first level.

Freely available online games would be educational material that allows a student to have privacy, make mistakes without fear, and repeat lessons as many times as necessary. The first idea was to represent the Croatian phonological system using the pronunciation of sounds and using animation for writing letters (that would be especially useful because looking at animation would bring to mind the

[9] The head of that faculty course was prof. Zrinka Jelaska, mentoring teachers were Jelena Cvitanušić Tvico and Ranka Đurđević, and students were Đurđica Ocvirk, Lovro Radošević, and Francesca Sammartino who performed classes for 12 students.

[10] To some students English is known as a *lingua franca* and French is known because it is used in Arabic countries such as Lebanon, Algeria, Tunis, and Morocco. The Latin script might be easier for students who know English or French, but on the other hand, students might have problems in learning Croatian if they rely only on their knowledge of English or French because their phonological inventory differs from Croatian.

different direction of writing). This interactive content is inspired by content present on site Russian For Everyone[11] where there are animations for handwriting of each letter and audio files for the pronunciation. There are also images and audio pronunciations for words and quizzes for each lesson. Another idea was that a completely new writing system might be intimidating for learners, so games will have explanations in Croatian and in Arabic[12]. Translation to the learner's mother tongue should reduce the fear of language and make learning quicker. In the initial stage (animation of writing, pronunciation) there are Croatian sounds divided into groups based on similar sound formation or similar shape in the Latin script. Sounds are divided into these groups: (1) A, E, I; (2) O, U; (3) K, G, H, (4) T, D; (5) L, J, Lj; (6) M, N, Nj; (7) B, P; (8) V, F, R; (9) S, Š; (10) Z, Ž; (11) C, Č, Ć; (12) Đ, Dž. Each sound in each group can be listened to and the animation can be seen. After that, every sound is presented in all positions, in the beginning, in the middle, and in the end of the word. Frequent Croatian words are chosen as well as internationalisms or names of Arabic countries or cities. Every word can also be pronounced by clicking on it and every word has a simple illustration or picture which represents its meaning. With 5 examples of words that have a certain sound in all positions, every group of sounds has few simple sentences which are useful for basic communication. Looking at a picture, the written word under it and clicking on it has the goal to increase phonological awareness. Phonological awareness is the learner's ability to divide words into sounds and to integrate sounds into words. It is also the ability to associate the sound with the letter and to understand the written word (Kolić-Vehovec, 2011: 17). After lessons on sounds, the learner should be able to establish a relation between phonemes and graphemes and, for example, be able to divide nebo (sky) into n-e-b-o and understand the question What is n-e-b-o? After these lessons, the learner will have games that implicitly increase phonological awareness: games in which the player recognizes the pronounced sound in the written word, memory games (connect the Croatian word with the picture), games in which the player recognizes the pronounced word, anagram games, hangman, bingo, etc. Words used in these games consist of sounds which appear in the lessons. In addition to this, words that have a semantic relation with those words occur in the games. So if in the lesson in which the learner learns the letter d the word dječak (boy) appears, in games after the lesson the antonym djevojčica (girl) would also appear; if in the lesson the learner learns star (old), mlad (young) would be added; kava (coffee) would be associated with co-hiponyms sok (juice), čaj (tea), vino (wine), pivo (beer), etc. After playing these games, learners should achieve these two goals:

1. Gain phonological awareness of Croatian phonemes and graphemes;
2. Acquire vocabulary which is necessary for learning Croatian on a higher language level.

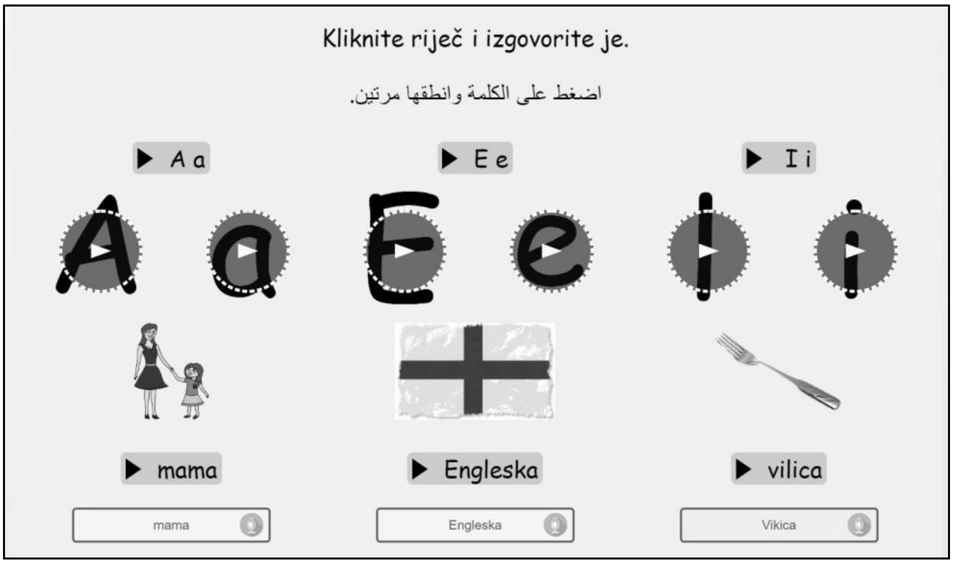
Games are also adjusted to Arabic speakers so lessons in which the student learns the difference between o and u, p and b or f, v, and r (which are problematic for Arabic speakers) would have more examples. After this level, games dealing with orthography (practicing capital and small letters) could be implemented as well as many other games that tackle specific grammatical problems and are connected to certain lessons (verb to be, Croatian declension and verbal system, etc.).

Technology used for creating interactive educational content

The interactive educational content the authors created is still in its demo version. Most of the content still needs to be expended, tested on students and then modified for final publication. However, most of the initial ideas for interactive content are implemented on a site at the moment of writing of this paper. The site is currently published as a GitLab page[13]. GitLab service allows the creation of private online repositories which can be privately accessed, updated, and shared using Git language or Git GUI clients like GitKraken[14]. Gitlab also has an option to publish a website on their server free of charge and hide the URL for web indexing by search bars (GitLab, 2018). When the content is finished it will be published within the module for foreigners of the Croatian Web Dictionary – Mrežnik. Mrežnik is a project conducted at the Institute of Croatian Language and Linguistics that aims at creating a free, monolingual, easily searchable hypertext online dictionary of the Croatian standard language with 10,000 entries (Hudeček, Mihaljević, 2017: 172). In addition to entries for

[11] Russian For Everyone: www.russianforeveryone.com/ (21.10.2019)
[12] Translator to Arabic was asylee Waddah Almasri who attended the B1.1. course at Croaticum in 2019.
[13] Learn Croatian alphabet: https://borna12.gitlab.io/igre-mreznik/sadrzaji_za_strance/ (26.10.2019)
[14] GitKraken: https://www.gitkraken.com/ (26.10.2019)

adult Croatian speakers, the dictionary also includes entries for children (3,000) and entries for non-native speakers (1,000) (Hudeček, Mihaljević, 2017: 175). Interactive content for learning the Croatian alphabet was divided into letters using drop-down menus with which users can directly open and close content on the site. Each letter has animations which demonstrate the correct order of line movements with the pen to get a certain letter. Font used for these animations was Comic Sans because it is recommended that fonts which are harder to read are better for studying letter writing because it is easier to see lines of different letters at different angles, and there is little to no uniformity in the Comic Sans design. The entire point of Comic Sans is that each letter is distinct from all others. People who have dyslexia usually use this font because the irregularly-shaped letters make it easier to break words down into their components and interpret them correctly (Newton, 2018). The order of line movement was taken from the book Initial writing in the Croatian language which has a detailed study and explanation of how to learn students the Croatian handwriting (Bežen, Reberski, 2014: 124-258). This was created to visually help learners in learning the Latin alphabet. Animations were created as .gif images by using image layers in the open-source free software GIMP 2.10[15]. GIMP is usually not used for animations but by having an option to have multiple layers in an image, it was possible to select the outline for each letter in a font and through each layer precisely draw a part of the letter by filling the outline. Layers could then be used as frames in an animation. On the site, the user can start a .gif animation by clicking on it. For playing .gif files like videos gifsee.js[16] JavaScript library was used. Audio files that were used for pronouncing letters, words or sentences in Croatian were played using Soundcite.js which allows inline blending of the text and audio. With Soundcite.js the audio is not isolated from the text and can be placed inside any part of a paragraph without breaking the text (Knightlab, 2013). The user only clicks on a word, letter or sentence displayed as a text to hear the pronunciation. Demo audio files were currently recorded by a laptop microphone, using a male and a female voice, but in the future, professional recording of new audio files is planned. Above certain words and sentences, there is a picture that represents that word or sentence. Below the link for audio pronunciation, an input form for users was added where they can pronounce words themselves in the microphone and those words will be written inside the input form. If the users pronounce words or sentences correctly the input form will turn green, otherwise, it will turn red.


Kliknite riječ i izgovorite je.

اضغط على الكلمة وانطقها مرتين.

Figure 1. Users can hear an audio recording of a word and try to pronounce it themselves

This web speech to text API was introduced in 2012 by the W3C Community. The goal was to enable speech recognition and synthesis in modern browsers. Currently, in 2019 this API is still a working draft only supported by Chrome browsers (W3C, 2012). The API currently supports 120 different

[15] GNU Image Manipulation Program (GIMP): https://www.gimp.org/ (26.10.2019)
[16] gifsee.js: https://klombomb.github.io/gifsee.js/ (26.10.2019)

languages and language dialects (Google Cloud, 2019). One of these languages is Croatian. This speech to text input forms ignore uppercase letters and punctuation, since there is no way to recognize them through voice (e.g. Croatian word oprosti or sorry can be treated as one word in a multi-word sentence connected to it or as a one-word sentence and it is hard for the computer to differentiate between these and similar cases). This API works well for short sentences and single words, so it mostly translates speech to text correctly. However, one Croatian sentence Mi imamo ljubičastu ljuljačku. (We have a purple swing.) had to be removed because API would always turn word ljuljačku (swing) into ljujašku. Sometimes it also turns incorrectly pronounced word into a correct word, e.g. word krumpir (potato) can be pronounced krompir and the API will still, probably because of the autocorrect options, register it as the word krumpir. Explanations on the site are currently written in Croatian and Arabic. For Arabic, web code had to be modified since Arabic is written and read from right to left. Explanations in other foreign languages could be added in the future.

Educational games

Games for learning the Croatian alphabet and words created so far include two quizzes, a hangman game, a memory game, and a game in which letters are rearranged to get a word that corresponds to the object shown on the picture. Game types for learning were chosen based on their global popularity and familiarity. Games such as quizzes, word puzzles, and memory are very easy to play and most people know how to play them. Also, these types of games are usually present on dictionary and encyclopedia websites, e.g. word games on Merriam-Webster[17] and quizzes on encyclopedia Britannica[18]. The games are still being developed (except for memory games), so new assignments will be added and some functionalities and technical improvements will be implemented later. Some of these games use and mix popular gamification elements such as scoring, time limits, difficulty adjustments, leaderboards, and virtual awards such as medals (Table 1). The games are web-based, responsive and support touch controls so they can be played on many computer devices that have one of the modern web browsers.

Table 1. Gamification elements in each game

game	gamification elements
quiz for learning Croatian letters	• difficulty adjustments • leaderboards • scoring • time limit • virtual medals
quiz for recognizing correctly spelled words	• leaderboards • scoring • time limit • virtual medals
the hangman game	• leaderboards • scoring • time limit • virtual medals
game for rearranging letters	• leaderboards • level selection (based on content) • scoring • time limit • virtual medals
memory game for foreigners learning Croatian	• level selection (based on content) • scoring

The first quiz is for learning Croatian letters[19]. In this game, a player gets a random word represented with a picture and an audio file on which a randomly selected letter of the given word is pronounced. An audio file is played by clicking on the speaker icon and after the player hears the letter he must

[17] Merriam-Webster games: https://www.merriam-webster.com/word-games (20.10.2019)

[18] Encyclopedia Britannica Quizzes: https://www.britannica.com/quiz/browse (20.10.2019)

[19] Quiz is for learning Croatian letters: https://borna12.gitlab.io/igre-mreznik//sadrzaji_za_strance/kivz-prepoznaj-slovo/index.html (20.10.2019)

select all occurrences of that letter in a given word. After the selection, the player can confirm his answer and will immediately get feedback for his answer. This game is effective because players can learn how to recognize letters and see their positions in relation to other letters that form the word. The pictures were added to help players recognize the word meaning.

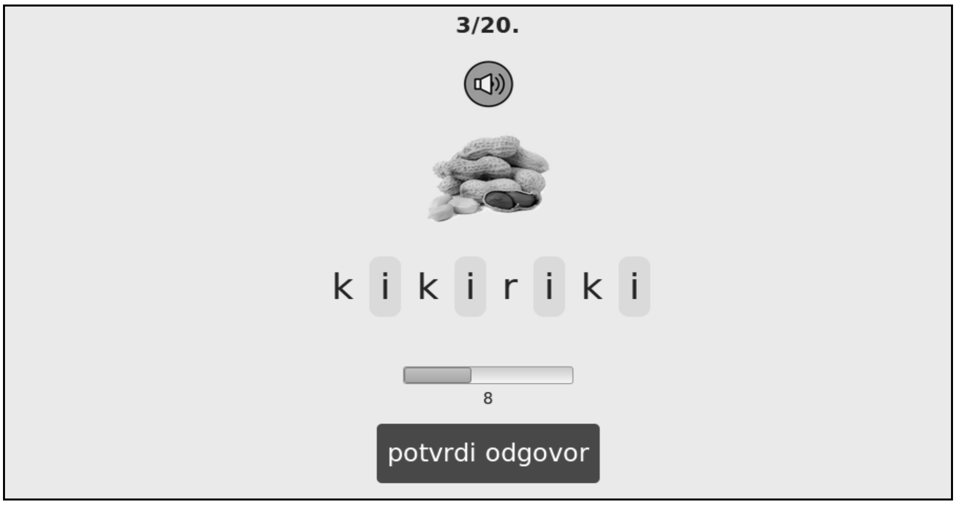


Figure 2. The game for recognizing letters in a pronounced word

In case of an incorrectly selected letter the player will get the correct answer, and in case of the correctly selected letter the player will get a picture of a happy face and the points earned for that question. There is also a different audio file played when submitting the correct and the wrong answers. A player can choose to play the game with the time limit switched on or off. With the time limit off there is a table of results but no virtual awards for the first, second or third place since the scores of players with the same number of points for correct answers cannot be diferentiated. With the time limit on it is easier to rank players because the time required for the response (in seconds or milliseconds) to each question can be taken into account and therefore more diverse results for each game can be obtained. The time limit can be set to 10 or 20 seconds depending on the player's choice. The leaderboards for these time limits are different. The player gets the overall score when he finishes the quiz and can submit the score under the custom username and emoticon that is displayed on the left-hand side of the name.

Most native speakers of Arabic who learn Croatian usually make mistakes in pronouncing and writing certain words. They sometimes make mistakes with only one phoneme. This is why the second quiz aims at recognizing correctly spelled words based on their pronunciation[20]. The quiz functions similarly to the previous one. The player gets a picture and presses the speaker icon to hear the pronunciation of the word. After that, the player gets four possible answers that are all similar to each other in spelling and only differentiated by one or two letters. The player has 20 seconds to choose the correct answer.

[20] Quiz for recognizing correctly spelled words: https://borna12.gitlab.io/igre-mreznik//sadrzaji_za_strance/kivz-prepoznaj-rijec/index.html (20.10.2019)

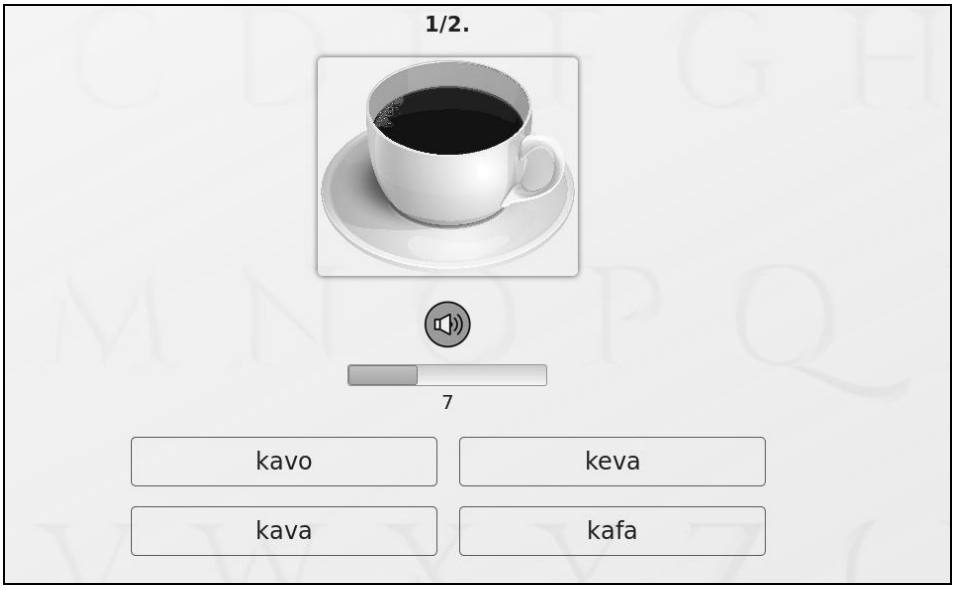


Figure 3. Game for choosing correct spelling for the heard word

The hangman game[21] shows pictures and empathy fields for letters that form the word that corresponds to the picture. The player has 40 seconds to choose the right letters and can only make four mistakes. Letters can be typed with the keyboard but buttons for letters were made for touchscreens as well. Wrongly selected letters are displayed below and drawings for the part of the hangman are generated below the letter buttons. The game has the same element as quizzes with the feedback for answers, the time limit for questions and the leaderboards and virtual medals for best players. The hangman can help students practice Croatian alphabet and vocabulary. By playing the hangman game, students will think about the letter that is in the word. They will also think about what the right word is. If the hangman game is used by teachers during class, it can be one of the ways to help students practice their confidence and to express their thought (Mandasari Manan, 2016: 141).

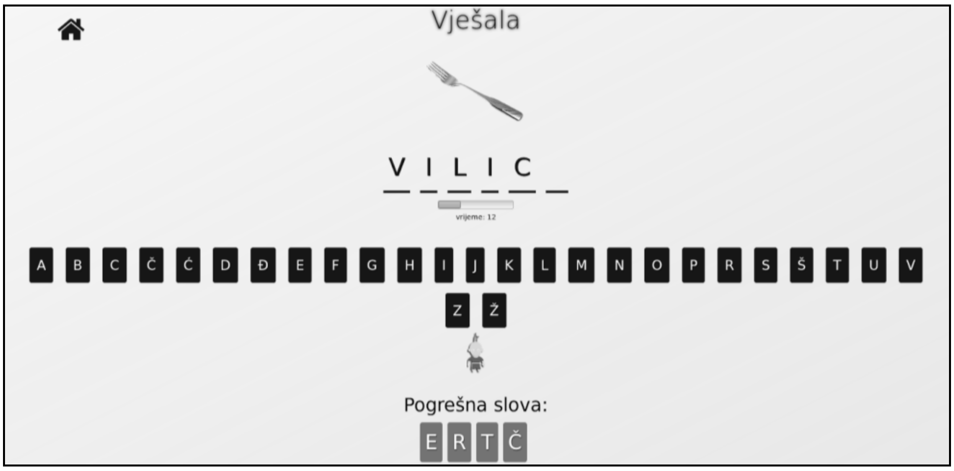


Figure 4. The hangman game for recognizing words represented by pictures

The game of rearranging letters[22] to get a word has content divided into categories. Player can choose a category, each category has its own words, e.g. colors or animals. Like in the hangman game, the player gets a picture but also randomly shuffled letters that form the word corresponding to the picture. The player has to use the letters and rearrange them to spell the word correctly.

[21] The hangman game: https://borna12.gitlab.io/igre-mreznik//sadrzaji_za_strance/vjesala/index.html (20.10.2019)

[22] Game for rearranging letters. https://borna12.gitlab.io/igre-mreznik//sadrzaji_za_strance/premetaljka/index.html (20.10.2019)

Memory games have proven to be adaptable for learning many different languages, which we can see through the examples of different memory games present on the site Jezične igre (Language games[23]). The site has memory for learning Latin, German and Croatian words as well as the Glagolitic alphabet. The memory game for learning the Glagolitic alphabet which was published in February 2019 currently has 758 submitted results and 559 likes on the Facebook (187 on the original post, 372 on the shared post)[24]. The memory game[25] for foreigners unlike some other games is finished and is currently publicly available online on the previously mentioned site Jezične igre. The advantage of memory games is that they can be used for matching many different elements like matching words for the same concept in different languages, matching words with other semantically related words (synonyms, antonyms, etc.), matching words with pictures, etc. In this game, the player has to pair the picture with the word. The player can choose the category for words he needs to pair, e.g. vehicles, fruit, vegetables, food, clothes, jobs, and animals. For each category, there is a different number of cards that need to be paired. There is no time limit but there is a timer which checks how long it took a player to finish the game and what was his best result. Although the game is publicly available, it can be updated with new words and pictures. Audio recordings for the pronunciation of words can also be implemented inside the game so the player can hear the pronunciation when he opens the card.

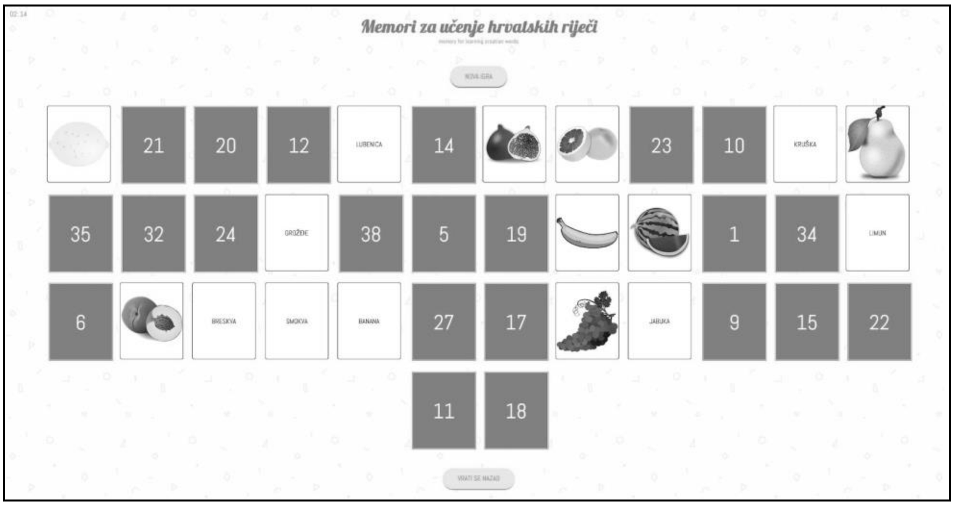


Figure 5. The memory game for matching words with pictures

Conclusion

Electronic educational games for learning Croatian tailor-made for native speakers of Arabic could be useful for all learners, but especially for asylees and asylum seekers who have many difficulties at the beginning of the learning process. Some of the presented games are still in the process of development and will be improved and modified in the future, e.g. some of them such as memory games lack leaderboards, and the hangman games lack the option to choose the level based on the content (food, animals, etc.). Also, more words can be inserted into the game. The data obtained from these games will also be used for analyzing results of players who are learning Croatian as a foreign, second or inherited language since submitted player results are automatically stored on the Google tables. From these tables and data automatically collected from Google Analytics there is an option to check how many times a certain game has been played, how good are the results based on the score, and how many recurring players there are compared to new players. Such data analysis will enable to check the interest of learners for using games in the learning progress. By comparing the results in games with the written practical exam we can also check their effectiveness as learning material. The data of this analysis can also be used in creating new or improving the existing games and interactive material for language learning.

[23] Language games: https://jezicneigre.com/ (20.10.2019)

[24] Games for learning Glagolitic script: https://www.facebook.com/ihjj.hr/photos/a.687321037952455/2715935941757611/?type=3&theater (20.10.2019)

[25] Memory game for foreigners: http://jezicneigre.com/hr/memori-nazivi/ (20.10.2019)

Acknowledgments

This paper is written within the research project Croatian Web Dictionary – MREŽNIK (IP-2016-06-2141), financed by the Croatian Science Foundation.

References

ATi Studios. (2019). Learn Croatian. Speak Croatian. https://play.google.com/store/apps/details?id=com.atistudios.mondly.hr&hl=en_US (26.10.2019)

Axosoft (2016). GitKraken: Free Git GUI Client. https://www.gitkraken.com/ (26.10.2019)

Bežen, A., Reberski, S. (2014). Početno pisanje na hrvatskome jeziku. Zagreb: Institut za hrvatski jezik i jezikoslovlje

Birkić, T., Golem, K., Kučina Softić, S., Martinović, Z., Radobolja, T., Zemljak Pećina, A. (2019). Sustav za e-učenje Merlin: priručnik za studente. Zagreb: Sveučilišni računski centar

Croaticum. Free Online Croatian Courses – A1.HR and A2.HR. 19.12.2018. https://croaticum.ffzg.unizg.hr/?page_id=5024 (26.10.2019)

Duolingo. (2014). Free Language Courses for Arabic Speakers. 7.3.2014. https://www.duolingo.com/courses/ar (20.10.2019)

Đurđević, R., Podboj, M. (2016). Izbjeglice kao posebna kategorija učenika inog jezika. // Strani jezici: časopis za primijenjenu lingvistiku 45, 3-4, 245-261

Encyclopedia Britannica. Encyclopedia Britannica Quizzes. (2019). https://www.britannica.com/quiz/browse (20.10.2019)

GIMP (2012). GNU Image Manipulation Program (GIMP). 10. 2. 2012. https://www.gimp.org/ (26.10.2019)

GitLab (2017). The DevOps Lifecycle with GitLab. 28. 04. 2017. https://about.gitlab.com/stages-devops-lifecycle/ (26.8.2019)

Google Cloud (2019). Cloud Speech-to-Text: Language support. 18. 6. 2019. https://cloud.google.com/speech-to-text/docs/languages (26.8.2019)

Hudeček, L., Mihaljević, M. (2017). The Croatian Web Dictionary Project – Mrežnik. // Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference / Kosem, I. et al. (eds.). Brno – Leiden: Lexical Computing CZ s.r.o., 172-192

Institut za hrvatski jezik i jezikoslovlje. (2019). Games for learning Glagolitic script. 21. 2. 2019. https://www.facebook.com/ihjj.hr/photos/a.687321037952455/2715935941757611/?type=3&theater (20.10.2019)

Klombomb, N. (2017). gifsee.js. 13.2.2017. https://klombomb.github.io/gifsee.js/ (26.10.2019)

Knightlab. (2013). SoundCite - Knight Lab Projects. 14. 5. 2013. https://projects.knightlab.com/projects/soundcite (26.8.2019)

Kolić-Vehovec, S. (2018). Razvoj fonološke svjesnosti i učenje čitanja: trogodišnje praćenje. // Hrvatska revija za rehabilitacijska istraživanja 39, 1, 17-32. https://hrcak.srce.hr/11618 (25.10.2019)

Mandasari Manan, R. (2016). The Use of Hangman Game in Motivating Students in Learning English. // ELT Perspective. 4, 2

Matijević, M. (2018). O sretnome vikendu i novim pogledima – nastavnička iskustva s tečaja hrvatskoga jezika za azilante i tražitelje azila. // Hrvatski jezik 5, 4, 5-8. https://hrcak.srce.hr/217536 (25.8.2019)

Matijević, M., Mihaljević, J. (2019a). Learn Croatian alphabet. 21. 10. 2019. https://borna12.gitlab.io/igre-mreznik/sadrzaji_za_strance/ (26.10.2019)

Matijević, M., Mihaljević, J. (2019b). Quiz for recognizing correctly spelled words. 21. 10. 2019. https://borna12.gitlab.io/igre-mreznik//sadrzaji_za_strance/kivz-prepoznaj-rijec/index.html (20.10.2019)

Matijević, M., Mihaljević, J. (2019c). Game for rearranging letters. 21. 10. 2019. https://borna12.gitlab.io/igre-mreznik//sadrzaji_za_strance/premetaljka/index.html (20.10.2019)

Matijević, M., Mihaljević, J. (2019d) Memory game for foreigners. 21. 10. 2019. http://jezicneigre.com/hr/memori-nazivi/ (20.10.2019)

Matijević, M., Mihaljević, J. (2019e). Quiz for learning Croatian letters. 21. 10. 2019. https://borna12.gitlab.io/igre-mreznik//sadrzaji_za_strance/kivz-prepoznaj-slovo/index.html (20.10.2019)

Matijević, M., Mihaljević, J. (2019f). The hangman game. 21. 10. 2019. https://borna12.gitlab.io/igre-mreznik//sadrzaji_za_strance/vjesala/index.html (20.10.2019)

Memrise. (2014). English for Arabic speakers. 21. 5. 2014. https://www.memrise.com/course/276681/english-for-arabic-speakers-2/ (20.10.2019)

Merriam-Webster. (2019). Merriam-Webster games. 2019. https://www.merriam-webster.com/word-games (20.10.2019)

Mihaljević Djigunović, J. (1998). Uloga afektivnih faktora u učenju stranog jezika. Zagreb: Filozofski fakultet

Mihaljević, J. (2019). Language games. 15. 10. 2019. https://jezicneigre.com/ (20.10.2019)

MUP. (2019). Statistički pokazatelji tražitelja međunarodne zaštite do 31. 3. 2019. https://mup.gov.hr/UserDocsImages/statistika/2019/Tra%C5%BEitelji%20me%C4%91unarodne%20za%C5%A1tite%20u%202019%20godini/29-04-statistika-trazitelji-1-3-2019.pdf (25.8.2019)

Newton, A. A. (2018). Get Over Yourself and Start Writing in Comic Sans. 18. 12. 2018. https://lifehacker.com/get-over-yourself-and-start-writing-in-comic-sans-1831177236 (26.8.2019)

Ocvirk, Đ., Radošević, L., Sammartino, F. (2019). Opismenjavanje na hrvatskome kao inome jeziku tražitelja azila i azilanata. Manuscript. Zagreb: Filozofski fakultet

Pučko otvoreno učilište (2018). Poučavanje djece kojoj hrvatski nije prvi jezik. Zagreb: Pučko otvoreno učilište Korak po korak

Quian, M., Clark, K. R. (2006). Game-based Learning and 21st century skills: A review of recent research. // Computers in Human Behavior 63, 50-58

Rochtchina, J. (2006). Russian For Everyone. 23.11.2006. http://www.russianforeveryone.com/ (21.10.2019)

Simya Solutions Ltd. (2019a). Jednostavno naučite hrvatski. 25. 10. 2019. https://play.google.com/store/apps/details?id=simply.learn.croatian&hl=hr (26.10.2019)

Simya Solutions Ltd. (2019b) Learn Croatian Language with Master Ling. 25. 10. 2019. https://play.google.com/store/apps/details?id=com.simyasolutions.ling.hr&hl=hr (26.10.2019)

UNHCR (1951). Convention Relating to the Status of Refugees. https://www.unhcr.org/3b66c2aa10 (25.8.2019)

W3C. (2012). Web Speech API Specification. 19.10.2012. https://w3c.github.io/speech-api/speechapi.html (26.8.2019)

Zakon o međunarodnoj i privremenoj zaštiti (2015). 17. 11. 2015. // Narodne novine 70/2015. https://mup.gov.hr/gradjani-281562/moji-dokumenti-281563/stranci-333/zakon-o-medjunarodnoj-i-privremenoj-zastiti/653 (25.8.2019)

https://doi.org/10.17234/INFUTURE.2019.18

Original scientific paper

Learning Japanese Script through Storytelling and Multimedia

Marija Bilić
University of Regensburg Germany
marija.bilic@stud.uni-regensburg.de

Tomislava Lauc
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
tlauc@ffzg.hr

Sanja Kišiček
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
smatic@ffzg.hr

Summary

This paper offers a rationale for storytelling and multimedia in the context of learning the Japanese script Hiragana. There is a growing interest in the field of computer-assisted language learning (CALL) of Japanese and the use of multimedia to achieve learning goals. The complexity of the Japanese writing system raises a lot of questions on how to accelerate and improve the process of learning Japanese scripts. In the domain of CALL and multimedia, this paper considers the methodological approach to language learning and presents a design of a multimedia tutorial for learning Japanese script Hiragana. The paper synthesizes multimedia and storytelling with its implications to learning and instructional design to inform creation of such a framework.

Key words: multimedia, storytelling, CALL, learning Japanese

Introduction

The development of computers and information and communication technologies (ICT) has impacted how approaches to teaching and learning have changed through history. Computers have been used in language learning since the 1960s (Seljan et al. 2004) and from that time to the present we are talking about computer-assisted language learning (CALL). With multimedia, opportunities to improve language learning and teaching have significantly increased (Kišiček et al., 2010). However, multimedia alone is not enough to achieve the desired learning outcomes. Multimedia needs to be integrated in the learning environment through an instructional design that purposefully leads the student to desired learning outcomes. Thus, the question that is asked in the context of language didactics is how to conceptualize foreign language teaching to achieve the learning goals more effectively in the context of multimedia didactics (Matasić, Dumić, 2012).

CALL is an interdisciplinary field that integrates applied linguistics and computing. According to Beatty (2003), CALL is "any process in which a student uses a computer and, as a result, improves his or her language." Specifically, CALL encompasses several issues such as material design, technologies, pedagogical theories, and teaching methods. Beatty (2003) states that CALL is still a highly unstructured discipline as it is constantly evolving with breakthroughs in pedagogy and with advances in hardware and software development. The development of CALL is also affected by the increasing level of computer literacy of teachers and students using CALL applications. Since the 1990s, a large number of CALL systems has been developed. During this period, there was an increasing emphasis on multimedia CALL, that is, CALL systems in a multimedia environment.

Simultaneous presentation of different multimedia elements creates an authentic environment for language learning including all the necessary aspects for learning a language: reading, writing, listening and speaking (Lauc et al., 2007). This paper describes the elements of instructional design such as multimedia and storytelling, on the example of the Japanese script Hiragana.

Multimedia in learning Japanese

In terms of how to learn and teach Hiragana script, tasks are traditionally based on reading and writing exercises, and various mnemonics are often used to improve memorizing of particular characters. The same principle applies to computer-aided language learning tools.

The opportunities offered by the multimedia environment have changed the concepts of learning and teaching. The new scientific discipline - multimedia didactics - is taking a central stage. Multimedia didactics deals with the learning goals and the design of educational content, as well as with the evaluating the effectiveness of multimedia projects and applications (Matasić, Dumić, 2012). The basis of this is, of course, the development of learning and teaching strategies, didactic and media design, and communication via the Internet. One of the most complex tasks in this context is to prepare a multimedia learning project. This job is extremely complex because it involves a large number of professionals in various fields. Professors, teachers, and lecturers are responsible for the material itself; instructional designers are responsible for presenting the material adapted to the new media, and the implementation itself is done by development experts such as graphic designers and developers. Also, multimedia language learning project is specific to language that is being learned.

The use of multimedia in language learning will not be the same for every language. In the example of computer-aided and multimedia learning in Japanese, it is evident that there is a large number of programs and applications that aim to teach Japanese script. The Japanese writing system consists of three letters, a hiragana, a katakana (kana) and a kanji (Yamaguchi, 2007). Numerous studies have been carried out to determine how and to what extent CALL and multimedia applications help to learn Japanese letters, and they also list various advantages and disadvantages of existing applications.

Hiragana and katakana mnemonics are often used by native speakers in Japanese language teaching and learning (Gilhooly, 2003). The usefulness of such a method of learning and remembering is recognized more widely, as indicated by the existence of material for English speakers. Since mnemonics for learning Japanese script are mainly available only for major world languages, as mentioned by Librenjak (2018) in Teaching Kana to Croatian Students, mnemonics were created for Croatian students of Japanese. The main goal was to design an illustration that connects the kana letter with the Croatian word beginning with the same syllable, for example, the hiragana letter あ / a / is shown as an aquarium in which a fish swims. These mnemonics were implemented in an Anki application that is based on a time-lapse learning method - how often a student answers the wrong question so often that question is repeated. Anki exists as a computer and mobile version which allows the student not to be attached to the classroom or computer, but to learn and repeat anytime, anywhere. Besides, each replay card contains multimedia, mostly in text and image (with Anki supporting both audio and video, but these elements are not included in the canoe mnemonics listed). It should be noted that Librenjak (2018) intended the application as supplementary material for learning scripture, in combination with teaching and traditional teaching methods (such as writing and reading exercises). By itself, the application cannot transfer all the skills needed to master the letter completely.

There is a great deal of research that analyses letter learning apps. Thus Moroz (2013) compares the advantages and disadvantages of two kanji learning applications. One of them (KanjiBox) is based on repetition cards (similar to Anki in the previous example) but also includes various quizzes and other types of exercise tasks. It also offers the option of ranking according to student performance and records various statistics. Another application (Kotoba!) is like a dictionary whose entries include kanji, reading it, translating it into English, and example sentences.

Mayer (2017) brings research/based principles for how to design computer-based multimedia instructional materials to promote learning. Arima (2009) considers aspects of instructional design when designing CALL and multimedia learning website and states that the high-quality instructional design includes interface consistency, simplicity, ease of navigation, and an adequate combination of colors and fonts. The site navigation needs to be designed in a visible and easily accessible location without interfering with learning. Websites can offer different multimedia elements such as videos for learning the correct way to write characters and audio files with authentic Japanese pronunciation.

Storytelling

Storytelling is the communication between the storyteller and the audience (Diaz, Fields, 2007). Storytelling today is no longer a method limited solely to the field of social sciences and humanities. Stories are increasingly used in both business and marketing, areas where large amounts of data are managed. The story turns hard data into meaningful information that can be more easily conveyed to others, which also contributes to better decision making. In addition to giving seemingly impersonal data and charts the ability to capture audience attention, stories can be action-packed in many ways. This applies not only to the audience to whom the story is being told, but also to the authors of the story themselves. In their research, Halpern and Lepore (2015) attempted to evoke "authorial identity" in students to involve them more in their research work. They believe that storytelling is a more personal way of accessing scholarly work and that the desire to tell a personal story could encourage students to become more engaged in their work and less plagiarized. This should also contribute to a better quality of student work. The research showed that the students were extremely motivated and put a lot of effort into their work because they wanted to convey their story in the best possible way. They liked the freedom of creative expression that they did not normally feel when writing scientific papers. In doing so, Halpern and Lepore (2015) have proven that storytelling can also have a positive effect on the scientific field.

Storytelling has found its application in multimedia learning as well. Using multimedia - digital images, videos, voice recordings, and music - it is possible to create a digital story. The concept of digital storytelling originated in the 1990s when Dana Atchley, Joe Lambert, and Nina Mullen founded the Center for Digital Storytelling (CDS). Although the original application of digital stories was the telling of personal life stories, today they are increasingly used for educational purposes. On the one hand, they offer teachers the opportunity to attract students' attention and to bring rich multimedia content to students with complex material. Also, students who are given the task of making a digital story on a topic have the freedom to express themselves creatively, which fosters their interest in the task and in the detailed exploration of the topic they are addressing.

In his literature review The Role of Storytelling in Language Learning, Lucarevschi (2016) brings on storytelling:

> Storytelling is one of the oldest forms of human communication, and much has been said in the literature about its effectiveness as a pedagogical tool in the development of language skills in first (L1) language, and also in a foreign or second language (L2), regardless of learners' age or background (e.g. Isbell, Sobol, Lindauer, Lowrance, 2004; Cameron, 2001). Furthermore, storytelling is even claimed to be more effective in language teaching than traditional teaching materials, such as textbooks. Indeed, studies generally believe that effectiveness of storytelling relies on the fact that it is fun, engaging and highly memorable, raising learners' interest in listening to stories, as well as in speaking, writing and reading about them (e.g. Atta-Alla, 2012, Kim, 2010; Wajnryb, 2003).

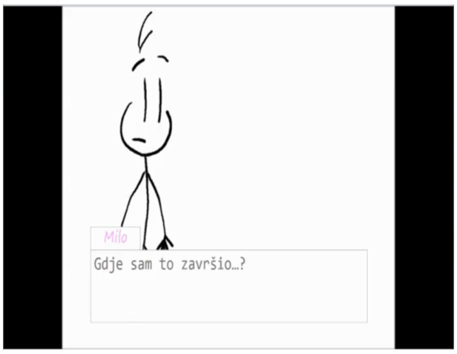
A story can get people's attention and evoke an emotional response, thus encouraging action to achieve a goal. Adding storytelling as a method of transferring knowledge to the benefits of multimedia learning is one way to achieve greater student motivation and interest in the material, and thus attract their attention to make learning effective and successful.

An example of learning Japanese script thorough multimedia and storytelling

The tutorial[1] that is presented in this paper is a short version of a potentially larger more comprehensive tutorial completely based on a story and interactive play in a virtual environment. The presented version of the tutorial for learning Japanese Hiragana script covers fifteen of the existing forty-six Hiragana characters. The tutorial was created using the HTML graphic element Canvas. The goal of the tutorial presentation is to describe how this tutorial uses storytelling and multimedia to enhance learning and achieve language learning goals. The tutorial is originally intended for younger learners due to its playful nature, and it is appropriate for other audiences as well.

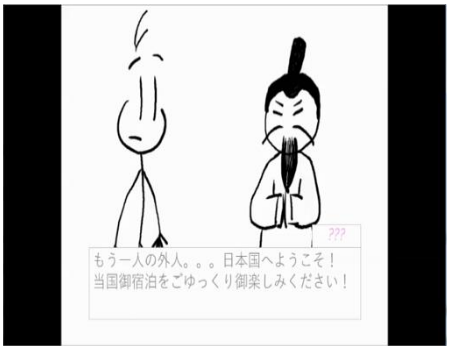
[1] The tutorial is created by Marija Bilić within the course Multimedia Knowledge Presentation and presented as a part of her BA thesis.

The story that introduces the user to the tutorial is very simple, but it is designed to fit as many aspects as possible (learning Japanese script). Although not complex, it can be analyzed by Aristotle's three-member structure, which is also described in Nussbaumer Knaflic (2015). The introduction of the story clarifies the context. The learner gets to know the main character Milo who wanders around an unknown land (Figure 1). In this aspect, there is already a problem that will lead to conflict and tension. Milo meets an Asian-looking character who speaks to him in an unusual language. It is a moment of conflict. The protagonist faces a problem of misunderstanding, due to language (see Figure 2). Then Milu meets another character, a winged cat named Muki, who explains the situation to both the protagonist and the listener (Figure 3). The Asian-looking gentleman is the ruler of the strange land of Nihon and he alone has the power to get Milo out of that country so he can return home. However, in order for Milo to ask him for help, he must learn the language the ruler speaks. This part can be understood as an appeal for learning not only to the protagonist, but primarily to the user who identifies with the main character (Figure 4).



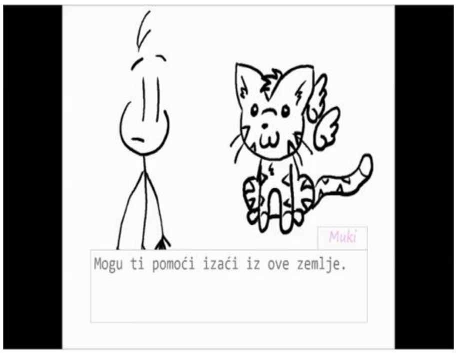
Gdje sam to završio…?

Figure 1. Milu in unknown land



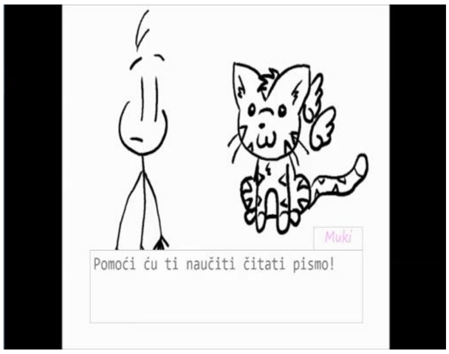
もう一人の外人。。。日本国へようこそ！
当国御宿泊をごゆっくり御楽しみください！

Figure 2. Conflict-unknown language



Mogu ti pomoći izaći iz ove zemlje.

Figure 3. Meeting the cat



Pomoći ću ti naučiti čitati pismo!

Figure 4. Getting help from the cat

The central part of the story is the main learning content itself (Figures 5-9). The student, together with the protagonist Milo, learns about the hiragana script and solves quizzes to test his knowledge. Each step in learning leads to culmination of the story, which culminates in a ruler's quiz. The ruler's quiz is the ultimate test of knowledge learned, and in the context of the story represents the protagonist's confrontation with a problem that led to conflict and tension. The story ends when the student/protagonist answers all the ruler's questions correctly. The protagonist then resolves the conflict and brings the story to an end. The learner needs to persevere until successful as this is the only way to continue the story.

In line with Arima (2009), the design was sought to include interface consistency, simplicity, ease of navigation, and an adequate combination of colors and fonts as shown in Figures 1-4 above, and Figures 5-9 below. Also, Segoe Print and Segoe Script fonts that resemble handwriting were used, with association to calligraphy. Considering multimedia, the tutorial uses text, image, sound, video and animation. A combination of text and image is dominant, which explains and illustrates the material. The main purpose of combination of text and image is to promote learning (Mayer 2017). Another purpose is to tell a story that increases the motivation to learn. Video as a combination of dynamic image and sound is used at the very beginning of the tutorial and serves to introduce the user to the story and explain its task (Figures 1-4). For this reason, the purpose of the video is obvious in the narrative aspect of the tutorials. Animations that represent only a dynamic image without a sound aspect are applied as feedback in quizzes (Figures 8 and 9). A certain atmosphere is reflected also through interactivity. When the learner answers the question correctly, an animation of a flowering Japanese cherry blossom is displayed. If the answer is incorrect, the flower turns black as if it faded.

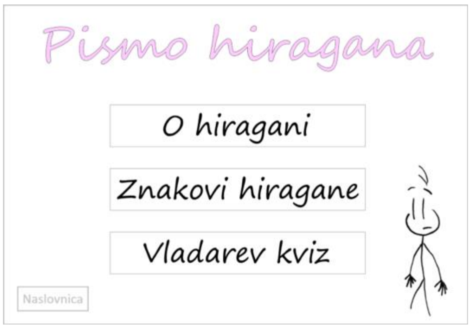
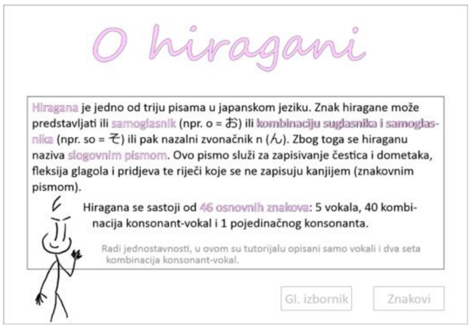


Figure 5. Initial tutorial screen



Hiragana je jedno od triju pisama u japanskom jeziku. Znak hiragane može predstavljati ili samoglasnik (npr. o = お) ili kombinaciju suglasnika i samoglasnika (npr. so = そ) ili pak nazalni zvonačnik n (ん). Zbog toga se hiraganu naziva slogovnim pismom. Ovo pismo služi za zapisivanje čestica i dometaka, fleksija glagola i pridjeva te riječi koje se ne zapisuju kanjijem (znakovnim pismom).

Hiragana se sastoji od 46 osnovnih znakova: 5 vokala, 40 kombinacija konsonant-vokal i 1 pojedinačnog konsonanta.

Radi jednostavnosti, u ovom su tutorijalu opisani samo vokali i dva seta kombinacija konsonant-vokal.

Figure 6. About Hiragana

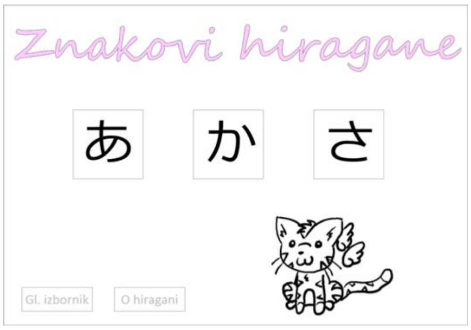


Figure 7. Hiragana Characters

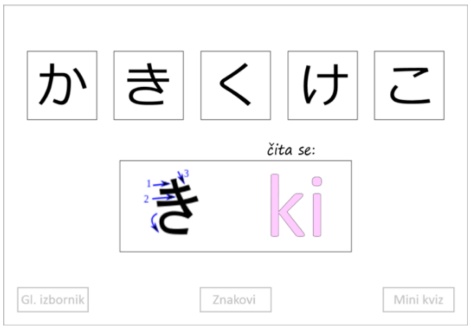


Figure 8. Reading and Writing Hiragana

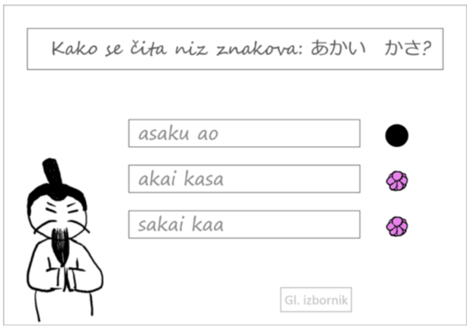

Kako se čita niz znakova: あかい　かさ?

Figure 9. Pronouncing characters (quiz)

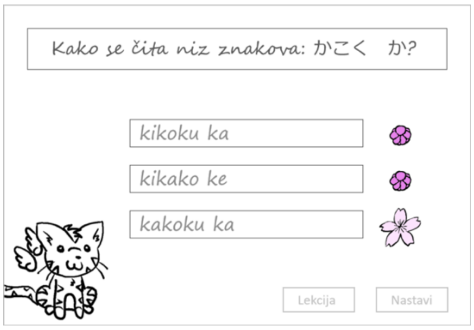

Kako se čita niz znakova: かこく　か?

Figure 10. Pronouncing characters

Stepping in front of the ruler and answering his questions is a crucial moment of the story because depending on the student’s, i.e. protagonist's performance, the story will have a positive ending. This fact serves as a motivation for the student to learn the material as much as possible and successfully solve the ruler's quiz. If the student fails to answer all the questions from the first, he or she must go a step back and repeat the preparatory quizzes. However, this is not failure for the learner, as there is an unlimited number of attempts for the learner to take on the ruler’s quiz.

Milo happily leaves Nihon and goes home enriched with new knowledge and experiences. The very end of the story is no surprise as it was already announced in the introduction. For this reason, the end of the story may also be considered an incentive for the student to learn and repeat until the solution to the problem presented at the very beginning of the story is reached.

The story itself is simple and it is presented through a simple interface. The main goal is learning Japanese script, and due to the simplicity of the story, the learner is not distracted from the purpose of the tutorial. Nevertheless, the story itself and the multimedia elements that flow through the story, remove the sense of 'learning', and the students' motivation becomes organic. Furthermore, identifying with the main character instils empathy in the student and gives him a goal that, while seemingly different, actually leads to the learning goal, which is to master the hiragana letter. With storytelling techniques, the intention is to provide additional motivation for language learning and contribute to faster and better acquisition of new knowledge.

What is worth noting is the lack of animations to show proper character writing. These animations would greatly contribute to the learning of hiragana writing. They would clarify possible concerns that arise when writing character is shown solely by the image with accompanying arrows (Figure 7). Therefore, it would be advisable to introduce such animations in a future version of the tutorial. The same is true for listening and proper pronunciation exercises that would also contribute to better mastering characters. Recordings of the correct pronunciation could be implemented in conjunction with a textual representation of the reading of each character, which would result in the learner learning both visually and audibly. This could speed up the learning process and make learning faster. Pronunciation exercises could be performed so that the student repeats what he or she is hearing on the sound recordings. Both multimedia elements and storytelling, with varying degrees of realization, should contribute to better and faster learning.

Conclusion

Computer-assisted language learning (CALL) has nowadays emerged as a methodology of language learning that shows positive results concerning the speed and ease of language acquisition, standalone or combined with classical teaching. Along with CALL applications, there is often talk of multimedia. By using different types of media, multimedia environment can enable a better understanding of different aspects of a language. Images can create mnemonics for learning a letter or word, recordings of authentic pronunciation help practice listening and acquire the proper accent of language, and animations or videos can aid learning through interactivity, that mere text cannot. Multimedia offers a variety of opportunities to improve language learning and teaching, and there is a growing number of available multimedia applications and tutorials for language learning. Although the interest in learning Japanese has only recently emerged, a large number of existing multimedia applications and research on the subject point to the opportunities offered in the field of teaching the Japanese language. The complexity of the Japanese writing system raises many questions about how to speed up and improve the process of adopting letters. As outlined in this paper, a combination of storytelling and multimedia is promoted as a way to prompt interest for learning, maintain learning motivation and achieve learning goals.

References

Arima, Y. (2009). Importance of Aesthetics in Language Learning Websites: Students’ Preferences Regarding Kana Learning Websites. Master thesis. University of Colorado

Beatty, K. (2003). Teaching and Researching Computer-assisted Language Learning. London [etc.]: Longman, 1-11, 16-51

Diaz, K., Fields, A. M. (2007). Digital Storytelling, Libraries, and Community. U: N. Courtney, ur., Library 2.0 and beyond: innovative technologies and tomorrow's user. 1st ed. Westport: Libraries Unlimited, 129-139

Gilhooly, H. (2003). Beginner's Japanese Script. 2nd ed. London: Hodder & Stoughton, vii-x, 60-72

Halpern, R., Lepore, L. (2015). Scholarly Storytelling: Using Stories as a Roadmap to Authentic and Creative Library Research. / Swanson, T. A, Jagmann, H. (eds.). Not just where to click: teaching students how to think about information, 1st ed. Chicago: Association of College and Reearch Libraries, A Division of the American Library Association, 349-365

Kišiček, S., Boras, D., Bago, P. (2010). Designing Educational Contents in and for the Electronic Environment. // ITI 2010 (32nd International Conference on Information Technology Interfaces); Proceedings of the ITI 2010 32nd International Conference on Information Technology Interfaces, Cavtat, 403-408

Lauc, T., Matić, S., Mikelić Preradović, N. (2007). Project of developing the multimedia software supporting teaching and learning of English vocabulary. // The Future of Information Sciences: INFuture2007 – Digital Information and Heritage. Zagreb, 493-499

Librenjak, S. (2018). Teaching Kana to Croatian Students through Native Language Mnemonics and Spaced Repetition. // Ueyama, M., Srdanović, I. (eds.) // Digital Resources for Learning Japanese. Bologna: Bononia University Press, 55-71

Lucarevschi, C. R. (2016). The Role of Storytelling on Language Learning: A Literature Review. // Working Papers of the Linguistics Circle of the University of Victoria 26, 1, 24-44

Matasić, I. and Dumić, S. (2012). Multimedia Technologies in Education. Media research: scientific and professional journal for journalism and media 18, 1, 143-151

Mayer, R .E. (2017) Using multimedia for e-learning. // Journal of Computer Assisted Learning 33, 5, October, 403-423

Moroz, A. (2013). App Assisted Language Learning: How Students Perceive Japanese Smartphone Apps. Master thesis. University of Alberta

Nussbaumer Knaflic, C. (2015). Storytelling with data: a data visualization guide for business professionals. Hoboken: John Wiley & Sons, 2, 165-185

Seljan, S., Berger, N., Dovedan, Z. (2004). Computer-Assisted Language Learning (CALL). // MIPRO 2004; Proceedings of the 27th International Convention MIPRO 2004: MEET + HGS. Rijeka: Liniavera, 262-266

Yamaguchi, T. (2007). Japanese Linguistics. An Introduction. London, New York: Continuum

https://doi.org/10.17234/INFUTURE.2019.19 Preliminary communication

Gamification in E-Lexicography

Josip Mihaljević
Institute of Croatian Language and Linguistics, Zagreb, Croatia
jmihalj@ihjj.hr

Summary

Gamification has become very popular in recent years. Many industries and scientific areas are trying to gamify their activities to make learning and work easier and more fun. Gamification can be achieved through educational games or by using certain game elements. The success of an educational game depends on how it incorporates game mechanics and educational content with gamification elements such as score, competition, ranking, and giving rewards. E-lexicography has great potential for using gamification to improve user's experience online. Encyclopaedias have a lot of interesting content that can easily be gamified through quizzes, crossword puzzles, jigsaw puzzles, and simulations. Dictionaries can also be gamified to help those who are learning words, structures or definitions in their mother tongue or in a foreign language. The focus of the paper will be on the analysis of the existing educational games and their gamification elements presented on available e-lexicographic sites. It will be determined which types of games are used for learning different content. The initial sample consists of 181 online dictionaries and 71 online encyclopaedias. The results of the analysis are displayed through tables show that more than 85% of e-lexicographic publications still don't have any game elements. Only 26 dictionaries and 10 encyclopaedias have some type of gamified content. The game type which occurs most often are quizzes, and the most common gamification element is scoring followed by levels divided by content. Badges and leaderboards are not used by many e-lexicographic publications that have games although they are considered to be an important factor in successful gamification of content as they raise students' interest and motivation. The purpose of this research is to find useful data and examples for the future creation of the gamification conceptual framework for Croatian online dictionaries. Based on the results of the analysis the author will present his demo version of educational games that are being made for the project The Croatian Web Dictionary – Mrežnik.

Key words: educational games, e-lexicography, gamification, language learning, Mrežnik

Introduction

Electronic lexicography has started its development in the middle of the 20th century under the name computer/computational lexicography (Granger, 2012: 1). Some of the first publicly available e-dictionary contents were stored on optical media (CD-ROM or DVD-ROM) which were published in addition to printed books. Some of these electronic dictionaries on optical media even had extra multimedia content such as audio and video files and some children lexicographic publications such as The First Croatian School Dictionary (Čilaš Šimpraga, Jojić, Lewis, 2008) also had games on a DVD. Internet technology up to the appearance of HTML5 in 2008 was still not developed for displaying multimedia content and advanced animation through web browsers. That is why early e-lexicography publications that wanted to have extra multimedia content had to be developed as executable programs for certain operating systems or as Flash applications. HTML5 changed that, so most of the multimedia content could be displayed and run on modern browsers without the need to pre-install certain programs. Also a huge rise in popularity and usage of many different types of content management systems such as WordPress and MediaWiki and dictionary writing systems such as TLex and iLex have led to a workflow where multiple users can easily work through specially designed graphical user interfaces for writing and editing lexicographic entries which are stored in online databases. These types of software also allow the users to graphically design the appearance of their dictionaries or encyclopaedias for the web. Another major advantage of e-lexicography publication on the internet is that there is no limit of the amount of content that can be displayed. Printed dictionaries and encyclopaedias are limited by paper size. Dictionaries and encyclopaedias which are stored on optical and disk media are also limited by storage size of the media. This is much less of a problem when publishing a dictionary or an encyclopaedia as a website since storage on a computer server or a cloud is usually larger than that of an optical and disk media and it can always be dynamically increased based on the needs of the user. Web published lexicography works are also always available for the user through the internet and their lexicographic entries and information can always be updated. Entries can be interlinked and external links can be added. Development and advantages of e-lexicography have led to its complete alignment with the basic concept of lexicography and numerous predictions about the complete disappearance of paper-based lexicographic publications (Granger, 2012: 2). Hill and Lauffer (2000: 73) discovered that web dictionaries have a positive effect on learning words and word meanings because they contain extra information that usually doesn't appear in paper dictionaries such as collocations, translation and, links to similar or connected words. Kraus et al. (2017: 177-179) based on their research of the content in encyclopaedias for children and adults on Britannica Kids, Q-files and KidzSearch Encyclopedia have identified the main elements that should be considered when developing good educational lexicographic websites such as well-organized content, effective search engine, good web page names, simple design of the homepage, well placed internal and external links on the site, supplementary multimedia for certain contents, sharing content on social networks and the ability to have multiple editors that can work on website content at the same time. Currently, there is no exhaustive research about how to gamify lexicographic content or analysis about games and game types e-lexicography sites usually have.

On gamification

Using electronic games for learning purposes can be the subject of study of many scientific disciplines, E-learning, gamification, and game-based learning are popular terms associated with educational games. Gamification has been more and more the subject of research in recent years due to its implementation into various mobile applications and educational systems. Gamification doesn't have a unique definition. One of the most quoted works From Game Design Elements to Gamefulness: Defining 'Gamification (Deterding et al., 2011: 10) defines gamification as the usage of existing game elements in situations which are not considered as a game or don't have game like characteristics. Game elements are derived from computer games and the most popular ones being used are game avatars, scores, leaderboards, levels, difficulty adjustments, virtual awards such as badges, etc. All these elements are interconnected, e.g. you cannot have leaderboards without scores, and should be well implemented inside learning applications. Rangaswami mentions that applications which have well designed awarding systems for working on certain assignments and reaching certain goals, like the games, are a key to success because they motivate users to continue work. Most research about using gamification for educational purposes has been positive, e.g. see Ortiz et al. (2016), Sitzmann (2011), Jagušt et al. (2018), Ružić and Dumančić (2015), Qian and Clark (2016). However, there are some doubts about whether gamification has a positive practical effect on long-time learning. Two studies (Fitz-Walter, Tjondronegoro, Wyeth, 2011; Montola et al., 2009) mention that there are a lot of problems in designing gamified content because some applications and games automatically give users scores for repeating certain mechanic actions to get points and don't change their contents or difficulty adjustment based on the player's results, which sometimes doesn't lead to cognitive development of brain or long term knowledge. Dominguez et al. (2013: 386) have mentioned in their resource that students are motivated when they receive badges for certain activities which they do in class, but that their results on written exams are mostly the same as those of students who didn't use any gamification elements during class. Markopoulos et al. (2015: 130-131) have mentioned that badly designed game content usually doesn't motivate students because they are not fun or interesting, and students don't feel like they have learned anything. He also mentioned how hard it is for a teacher to create specifically designed gamification content because they require certain technical skills and time spent on developing them. When designing gamelike educational content it is important to know the main learning goal that should be achieved with player interactivity. That way certain game elements, mechanics and designs can be well mixed with the educational content. Gamified content also must be appropriated for certain users based on their age, knowledge, interest, and motoric functions. To keep players happy sometimes existing game mechanics and elements have to be adjusted or upgraded and many different types of games have to be combined or a new type of game with unique rules has to be invented.

Games and game elements in lexicography

Games in lexicography can be diverse because of many different topics that can be covered from language learning, spelling, alphabet, grammar, history and culture, etc. There are also many different types of educational games such as quizzes, crossword puzzles, jigsaw puzzles, and unique games that have their rules and playing mechanics. Some online tools on lexicographic sites such as Scrabble Word Finder, while not being considered as a game, can help players when they are playing board game Scrabble where the goal is to match letter tiles to get words that exist in the dictionary[1]. Some sites like Macmillan Dictionary, in addition to games, have written assignments and tasks which are stored as PDF documents[2]. Some of these documents may even ask a player to use a dictionary while solving assignments. However, gamification elements cannot be present interactively on their own on paper since there is no visual feedback or sometimes there is nobody scoring and making sure that the player follows the rules of the game. This is not a problem when using computer games or applications since they automatically score your activities based on their algorithm and can restrict user interaction with educational content based on their design.

Methodology

The analysis conducted here will only focus on games which are present on dictionary websites. Games which are present on optical drives would be hard to analyse for foreign lexicography publications and some of them are made using old technologies, e.g. The First Croatian School Dictionary uses the old version of Macromedia Flash player, so it is hard to run the DVD on modern systems. The analysis will also not cover language learning and educational platforms such as Memrise and Duolingo since those are not websites for dictionaries or encyclopaedias even if they contain certain contents and assignments from published lexicography works. Some dictionaries and encyclopaedias such as TheFreeDictionay.com[3] and Ancient History Encyclopaedia[4] have mobile applications but these will be analysed in another paper. Encyclopaedias and dictionaries were analysed in separate tables but the elements of analysis were the same. The sample includes 71 online encyclopaedias and 182 online dictionaries. It includes many different general and technical lexicographic publications and even some terminological databases. Small glossaries on portals such as Time and Climate of Croatian Adriatic[5] where not included in the analysis because of the small scope of their content. Electronic lexicographic work published as PDF files on a website was also not included in the analysis, because they are not real websites and don't have enough interactive elements. Most popular and well know world lexicographic publication such as Dictionary by Merriam-Webster[6] and Encyclopaedia Britannica[7] were chosen upfront, along with the Croatian lexicographic publications such as Croatian Orthographic Manual[8] and Croatan Encyclopedia[9] from The Miroslav Krleža Institute of Lexicography. Other popular e-lexicographic publications were found using academic search programs such as RefSeek[10] and iSEEK[11] and Wikipedia has a list of all collected online dictionaries[12] and encyclopaedias[13]. Dictionaries mentioned on The European

[1] Collins Dictionary. Scrabble Word Finder. 20.12.2012. https://www.collinsdictionary.com/scrabble/scrabble-word-finder/ (13.6.2019)

[2] Macmillan Dictionary. Language puzzles. 02.05.2013. https://www.macmillandictionary.com/language-games/puzzles (14.6.2019)

[3] Google play. Dictionary. https://play.google.com/store/apps/details?id=com.tfd.mobile.TfdSearch (17.6.2019)

[4] Google play. Ancient History Encyclopedia. https://play.google.com/store/apps/details?id=com.ah.ahe (17.6.2019)

[5] Vrijeme i klima hrvatskog jadrana. Pojmovnik. http://jadran.gfz.hr/pojmovnik.html (17.6.2019)

[6] Merriam-Webster. Dictionary by Merriam-Webster. https://www.merriam-webster.com/ (17.6.2019.)

[7] Britannica.com. Encyclopedia Britannica. https://www.britannica.com/ (17.6.2019)

[8] Hrvatski pravopis. Rječnik. http://pravopis.hr/ (17.6.2019)

[9] Leksikografski zavod Miroslav Krleža. Hrvatska enciklopedija. http://www.enciklopedija.hr/ (17.6.2019)

[10] RefSeek. 30 Best Online Dictionaries and Thesauri. https://www.refseek.com/directory/dictionaries.html (17.6.2019)

[11] iSEEK.com. iSEEK - Education. http://www.iseek.com/iseek/home.page (17.6.2019)

[12] Wikipedia. List of online dictionaries. 24.07.2019 https://en.wikipedia.org/wiki/List_of_online_dictionaries (17.6.2019)

[13] Wikipedia. List of online encyclopedias. 20.07.2019. https://en.wikipedia.org/wiki/List_of_online_encyclopedias (17.6.2019)

Dictionary Portal[14] were also analysed. The sites that were not in English or Croatian were automatically translated on the site through Google Chrome translation plugin that is built inside the browser. The first step of the analysis was to check if there are any games present on the dictionaries or encyclopaedias websites. Those sites with games would later be additionally analysed to check present game types and present gamification elements. Additional data such as dictionary and encyclopaedia types were also collected.

Dictionary analysis

From the sample of 181 dictionaries, only 26 dictionaries have any type of games. The present game contents which were identified on dictionaries websites can be categorized as: quizzes, hangman, drag and drop or connecting games, memory, crosswords, filling in the blanks, typing the correct word based on hearing, puzzles, offline game material, fast typing or dactylography games, and games for finding words.

Table 1. Game contents in dictionaries

	quiz	hangman	connect-ing games	memory	cross-word	fill in the blank	game for typing heard words	puzzle	offline material for games	fast typing game	word finding game
Bildwörterbuch	no	no	yes	no	no	no	yes	no	no	no	yes
Merriam-Webster	yes	no	yes	no	yes	yes	yes	yes	no	no	no
Oxford Dictionaries	yes	no	no	no	no	no	no	no	no	no	no
Macmillan English Dictionary	yes	no	no	no	no	no	no	no	yes	no	no
Vocabulary.com	yes	no	no	no	no	yes	yes	no	no	no	no
Collins Online Dictionary	no	no	no	no	no	no	no	no	yes	no	no
Dictionary.com	yes	no	no	no	yes	no	no	no	yes	no	no
Longman Dictionary of Contemporary English	yes	no	yes	no	no	yes	no	no	no	no	no
LEO	yes	no	no	no	no	yes	no	no	no	no	no
Your dictionary	no	no	no	no	no	no	no	no	yes	no	no
Englesko-hrvatski kemijski rječnik & glosar	no	no	no	yes	no	no	no	no	no	no	no
The Free Dictionary	yes	yes	yes	no	no	no	yes	no	no	no	no
Merriam Webster Visual Dictionary	no	no	yes	no	no	no	yes	no	no	no	no
Van Dale	yes	no	no	no	no	no	no	no	no	no	no
Arhivistički rječnik	no	no	no	no	no	no	no	no	no	yes	no

[14] dictionaryportal.eu. European Dictionary Portal. http://www.dictionaryportal.eu/en/ (17.6.2019)

BSL Signbank	yes	no	no	no	no	no	no	no	no	no	no
Dictionary of the dialects of Jutland	yes	no	no	no	no	no	no	no	no	no	no
Sprotin Online dictionaries	yes	no	no	no	no	no	no	no	no	no	no
NGT Signbank	yes	no	no	no	no	no	no	yes	no	no	no
Online English Turkish and Multilingual Dictionary	yes	yes	no	yes	yes	no	no	no	no	no	yes
Diccionario Clave	no	no	yes	no	no	no	no	no	no	no	no
Diccionario visual	no	no	yes	no	no	no	no	no	no	no	no
Romanian Language Dictionaries	yes	yes	no	no	no	no	no	yes	no	no	no
Infopédia Dicionários Porto Editora	yes	no	no	no	no	no	no	no	no	no	no
Duden Online-Wörterbuch	yes	no	no	no	no	no	no	no	no	no	no
Dictionnaire visuel	no	no	no	no	no	no	no	no	no	no	no

Table 2. Number of game contents presented in dictionaries

quizzes	connecting games	game for typing heard word	offline material for games	fill in the blanks games	puzzles	hangman	memory games	crosswords	word-finding games	fast typing games
17	7	5	4	4	3	3	2	3	2	1

From the results we can see that most dictionaries that have games usually have quizzes (65% of dictionaries with games have quizzes). Most of these quizzes are multiple choice where the user clicks on the correct answer. Three dictionaries BSL Signbank, NGT Signbank, and Sprotin Online dictionaries have picture-based quizzes for learning the sign language. The second most present game type are games for connecting word with other words or words with images and games for typing words they hear. Connecting games are present in picture dictionaries such as Bildwörterbuch, Merriam Webster Visual Dictionary and Dictionnaire visuel. Games for typing words that they hear are usually meant for foreigners who are learning a foreign language.

Table 3. Number of gamification elements presented in dictionary games

scoring	levels or difficulty selection	time limits	badges	leaderboards
16	10	8	3	4

From the results we can see that scoring is the element which is most often present in all games. Dictionary sites Vocabulary.com and The Free Dictionary are even built as a system in which registered users get scores on their profiles for using dictionary content. They are also the only dictionaries that have implemented badges that come as rewards to players for finishing certain activities in the dictionary or for completing certain games. Time limit is only present in quizzes and in the case of quizzes on Merriam-Webster site can be turned off. Most quizzes also have leaderboards but they are usually made for registered users of dictionary websites and in the cases of Vocabulary.com and The Free Dictionary are connected with badges. Levels or difficulty selections are usually present as different lectures similar to Longman Dictionary of Contemporary English[15] or games on Merriam-Webster dictionary and Romanian Language Dictionaries[16] allow players to adjust difficulty by selecting easier sets of questions, turning of time limit or reducing the number of possible answers for selection. Arhivistički rječnik (engl. Croatian Archival Dictionary) has a dactylography game[17] that, similar to arcade games, gets harder as the player progress through game levels and the goal of the game through each play through is to get to the highest level starting from the first level.

Encyclopaedia analysis

From 71 encyclopaedias only 10 encyclopaedias have any type of games. The game types identified are mostly the same as dictionary ones. The only difference is that there are no fast typing games or games for typing words the player hears. Also, there are some unique educational games that have their set of rules and gameplay mechanics so they are categorized as unique educational games. Gamified content of Columbia Encyclopedia is displayed through Fact monster portal[18] and not the official site.

Table 4. Game contents in encyclopaedias

	quiz	hangman	connectin g games	memory	crossword	fill in the blank	puzzle	offline material for games	word-finding game	unique games
Britannica	yes	no	no	no	no	no	no	no	no	no
Columbia Encyclopedia	yes	yes	no	no	no	no	no	no	no	no
Encyclopedia Smithsonian	yes	no	yes	no	no	no	no	yes	no	no
Medline Medical Encyclopedia	yes	no	yes	yes	yes	yes	no	yes	yes	no
Wikipedia	no	no	no	no	no	no	no	no	no	no
Baidu Baike	no	no	no	no	no	no	no	no	no	no
Krugosvet	yes	no	no	no	no	no	no	no	no	no
Nationalen-cyklopedin	yes	no	no	no	no	no	no	no	no	no
Encyclopédie La-rousse en ligne	yes	no	no	no	no	no	no	no	no	no
World Book Encyclopedia	yes	no	yes	no	yes	no	no	yes	no	no

Table 5. Number of game types present in encyclopaedias

quizzes	unique games	connecting games	puzzles	crosswords	fill in the blanks games	number memory games	offline material for games
8	3	3	3	2	1	1	1

[15] Longman Dictionary of Contemporary English. Free English exercises. https://www.ldoceonline.com/exercise/ (18.6.2019)

[16] dexonline. Moara cuvintelor. https://dexonline.ro/moara (18.6.2019)

[17] Hrvatsko arhivističko društvo. Arhivistički rječnik - tipkalica. 26.10.2016. https://www.had-info.hr/arhivisticke-igre/arhivisticki-rjecnik-tipkalica (18.6.2019)

[18] Fact monster. Fact monster - Homework Help, Dictionary, Encyclopedia, and Online Almanac. https://www.factmonster.com/ (18.6.2019)

Similar to dictionaries, quizzes are also the most frequently present game type. Encyclopedia Britannica is an exception here with quizzes made for many different categories in which registered users on the site can compete one against another and share their score through social media. Encyclopedia Smithsonian has a lot of unique educational games such as simulation and strategy game Aquation: The Freshwater Access Game[19] where a player must build water factories in certain countries using available resources, then spread water through pipes in certain countries so that everyone can get an equal supply of water. A player has to be careful of the amount of money he spends for building pipes, factories, and doing scientific research for improving production of quality water. This game teaches the players how to manage resources and the importance of water distribution. Wikipedia is the only crowdsourced encyclopaedias that has its games that are based on using the Wikipedia. The Wikipedia Adventure[20] guides players through a few levels on how to edit Wikipedia articles, add references, and how to distinguish the bad source of information from the good one. Another game The Wiki game[21] gives a player a random article and a time limit in which he must navigate by using hyperlinks on Wikipedia, to a certain article that is in some way connected to the first article. Gamification elements that were studied in these games are the same as for dictionaries.

Table 6. Number of gamification elements in encyclopaedias

scoring	levels or difficulty selection	time limits	badges	leaderboards
10	5	5	2	3

All of the encyclopaedias with games have scoring systems integrated into them. Half of the encyclopaedias with games have game content separated by levels based on subject and difficulty. A lot of quizzes have time limits. Wikipedia and Baidu Baike are the only encyclopaedias that have badge systems for rewarding users for learning encyclopaedias content. These two sites along with Encyclopedia Britannica also have leaderboards for registered users which can track their scores and states for different games.

Interpreting the overall results

From the sample of 252 dictionaries and encyclopaedias only 26 dictionaries and 10 encyclopaedias have some type of gamified content. The overall results of the analysis show that more than 85% of e-lexicographic publications still don't have any game elements. A quiz with selectable multiple answers is a game type which occurs most often (17 dictionary and 8 encyclopaedia sites with quizzes). The most common gamification element is scoring (16 dictionaries and 10 encyclopaedias) followed by levels divided by content (10 dictionaries and 5 encyclopaedias). The reason why quizzes are the most popular educational content is probably because they are conceptually and technically easier to make than other games. Also, they have simple rules, they are easy to play and can educate the player directly by going straight to questions. They are much easier to program than other mentioned game types and there is a lot of software for creating them, e.g. quiz creation GUI in Moodle and H5P platforms. Quizzes are also more flexible for gamifying any content because you can make fun and interesting questions for any subject. Quizzes can cover learning trivia, grammar, spelling, pronunciation, history, culture, etc. However, by analysing the quizzes on lexicographic sites it is important to mention that quizzes that have instant feedback seem more effective since the player gets a direct response to his actions. The player gets feedback why something is the correct answer and points and results are displayed, e.g. quizzes on the Merriam-Webster site. Other game types in addition to quizzes should also be used, e.g. word-finding games or Scrabble, because they can offer more dynamic gameplay or in case of crosswords and puzzles develop certain cognitive functions because a player has to actively think to solve the problem. Although badges and leaderboards are considered to be an important factor in successful gamification of the content, as they can raise students' interest and motivation, they are not used by many e-lexicographic publications that have

[19] Smithsonian Science Education Center. Aquation: The Freshwater Access Game. https://ssec.si.edu/aquation (18.8.2019)

[20] Wikipedia. The Wikipedia Adventure. https://en.wikipedia.org/wiki/Wikipedia:The_Wikipedia_Adventure (18.8.2019)

[21] The Wiki Game. Wikipedia Game - Explore Wikipedia!. https://www.thewikigame.com/group (18.8.2019)

games. There is a small number of unique educational game types similar to some mentioned for encyclopaedias. These custom-made unique educational games with their gameplay mechanics, storylines that can smoothly connect gameplay, story and education is something that should be developed and used more often even if these types of games are more complex to make from the technical and conceptual aspect.

Creating games for dictionaries

The purpose of this research is to find useful data and examples for the future creation of gamification conceptual framework for the currently developed Croatian online dictionary – Mrežnik. Mrežnik is a project from the Institute of Croatian Language and Linguistics that aims at creating a free, monolingual, easily searchable hypertext online dictionary of the Croatian standard language with 10,000 entries (Hudeček, Mihaljević, 2017: 172). In addition to basic definitions, the dictionary also includes definitions for children (3000) and definitions for non-native speakers (1000) (Hudeček, Mihaljević, 2017: 175). For all three modules, the author is developing lexicographic games similar to the mentioned ones. These games will be implemented into the dictionary structure of Mrežnik. They will be divided into games for learning spelling, grammar, words, word relations, and word meanings and even games for learning special or old alphabets. Some games have already been developed and published on other Institute sites. On the site Croatian in school (hrvatski.hr) there is a dactylography game where you have to type the correct words fast before it falls to the bottom of the screen[22]. Words are taken from Mrežnik and they are specifically selected because they create a spelling problem even for native speakers of Croatian (e.g. č, ć, dž, đ, ije, je). There is also a crossword game for elementary school children with questions that require correct spelling and knowing certain word meanings[23], memory for matching Croatian words with anglicism[24] and there are a lot of quizzes. One of the quizzes is for recognizing the ancient Croatian alphabet Glagoljica (engl. Glagolitic script)[25]. This quiz has a time limit of 10 seconds for recognizing each letter, gives feedback for every answer, and has leaderbords for every player who submits his results by typing his username, without any need to register. Top three players get medals: gold, silver, bronze. Similar to this game, a game for learning the Braille alphabet by clicking on empty circles to get certain symbols[26] and a quiz for learning the alphabet of the sign language[27] are being developed. A game for learning the Croatian grammar[28] is being developed where a player has to write the correct form of a given verb and, in some cases,, after the correct answer, he can get an extra multiple choice question, such as which language rule was used to get the correct form. The site for learning Croatian words and alphabet for non-native learners of Croatian is also in an early stage of development and contains input forms where users can pronounce certain words using a microphone and their pronunciation will be checked automatically[29]. This site will also have quizzes, memory, and hangman games for learning words and alphabet. Other games are being developed as well, such as drag and drop games for categorizing animals, plants, planets, and trees[30]. There is a plan to develop more games and possibly more game types. All of these games will use gamification element of scoring and most of them will have leaderboards and badges for best players. Some of the games will have difficulty adjustments and level selections based on language content.

Acknowledgments

This paper is written within the research project Croatian Web Dictionary – MREŽNIK (IP-2016-06-2141) financed by the Croatian Science Foundation.

[22] Hrvatski u školi. Utipkaj riječ. http://hrvatski.hr/igra/4/ (22.7.2019)

[23] Hrvatski u školi. Prvi školski pravopis - križaljka. http://hrvatski.hr/igra/5/ (22.7.2019)

[24] Hrvatski u školi. Bolje je hrvatski! - pamtilica. http://hrvatski.hr/igra/3/ (22.7.2019)

[25] Hrvatski u školi. Zna glagoljicu. http://hrvatski.hr/games/kviz-glagoljica/ (22.7.2019)

[26] GitLab. Prepoznaj brajicu. https://borna12.gitlab.io/igre-mreznik/brajica/ (22.7.2019)

[27] GitLab. Slova znakovnog jezika. https://borna12.gitlab.io/igre-mreznik/kviz-znakovi/ (22.7.2019)

[28] GitLab. Upisivanje nastavaka za glagole. https://borna12.gitlab.io/igre-mreznik/kivz-upisi-naziv/ (22.7.2019)

[29] GitLab. Learn Croatian words and alphabet. https://borna12.gitlab.io/igre-mreznik/sadrzaji_za_strance/ (18.8.2019)

[30] GitLab. Planeti sunčeva sustava. https://borna12.gitlab.io/igre-mreznik/planeti/ (22.7.2019)

References

Botički, I., Jagušt, T., So, H. J. (2018). Examining Competitive, Collaborative and Adaptive Gamification in Young Learners' Math Learning. Computers and education 125, 444-457

Britannica.com. Encyclopedia Britannica. https://www.britannica.com/ (17.6.2019)

Čilaš Šimpraga, A., Jojić, L., Lewis, K. (2008). Prvi školski rječnik. Zagreb: Institut za hrvatski jezik i jezikoslovlje

Collins Dictionary. Scrabble Word Finder. 20.12.2012. https://www.collinsdictionary.com/scrabble/scrabble-word-finder/ (13. 6.2019)

Deterding, S., Dixon, D., Khaled, R., Nacke, L. (2011). From game design elements to gamefulness: defining gamification. // Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments / Lugmayr, Artur (ed.). New York: ACM, 9-15

dexonline. Moara cuvintelor. https://dexonline.ro/moara (18.6.2019)

dictionaryportal.eu. European Dictionary Portal. http://www.dictionaryportal.eu/en/ (17.6.2019)

Dominguez, A., Saenz-de-Navarrete, J., de-Marcos, L., Fernandez-Sanz, L., Pages, C., Martinez-Herraiz, J. J. (2013). Gamifying learning experiences: Practical implications and outcomes. // Computers & Education. 63, 386–392

Dumančić, M., Medica Ružić, I. (2015). Gamification in education. // Informatologia 48, 3-4, 198-204

Fact monster. Fact monster - Homework Help, Dictionary, Encyclopedia, and Online Almanac. https://www.factmonster.com/ (18.6.2019)

Fitz-Walter, Z., Tjondronegoro, D., Wyeth, P. (2011). Orientation passport: Using gamification to engage university students. // Proceedings of the 23rd Australian computer-human interaction conference / Shen, H. et al. (eds.). Canberra : ACM, 122-125

GitLab. Learn Croatian words and alphabet. https://borna12.gitlab.io/igre-mreznik/sadrzaji_za_strance/ (18.8.2019)

GitLab. Planeti sunčeva sustava. https://borna12.gitlab.io/igre-mreznik/planeti/ (22.7.2019)

GitLab. Prepoznaj brajicu. https://borna12.gitlab.io/igre-mreznik/brajica/ (22.7.2019)

GitLab. Slova znakovnog jezika. https://borna12.gitlab.io/igre-mreznik/kviz-znakovi/ (22.7.2019)

GitLab. Upisivanje nastavaka za glagole. https://borna12.gitlab.io/igre-mreznik/kivz-upisi-naziv/ (22.07.2019)

Google play. Ancient History Encyclopedia. https://play.google.com/store/apps/details?id=com.ah.ahe (17.6.2019)

Google play. Dictionary. https://play.google.com/store/apps/details?id=com.tfd.mobile.TfdSearch (17.6.2019)

Granger, S. (2012). Electronic lexicography: From challenge to opportunity. // Electronic Lexicography / Granger, S., Paquot, M. (eds.). Oxford: Oxford University Press, 1-11

Hill, M., Laufer, B. (2000). What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? // Language Learning and Technology 3, 2, 58-76

Hrvatski pravopis. Rječnik. http://pravopis.hr/ (17.6.2019)

Hrvatski u školi. Bolje je hrvatski! - pamtilica. http://hrvatski.hr/igra/3/ (22.7.2019)

Hrvatski u školi. Prvi školski pravopis - križaljka. http://hrvatski.hr/igra/5/ (22.7.2019)

Hrvatski u školi. Utipkaj riječ. http://hrvatski.hr/igra/4/ (22.7.2019)

Hrvatski u školi. Zna glagoljicu. http://hrvatski.hr/games/kviz-glagoljica/ (22.7.2019)

Hrvatsko arhivističko društvo. Arhivistički rječnik - tipkalica. 26.10.2016. https://www.had-info.hr/arhivisticke-igre/arhivisticki-rjecnik-tipkalica (18.6.2019)

Hudeček, L., Mihaljević, M. (2017). The Croatian Web Dictionary Project–Mrežnik. // Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference / Kosem, I. et al. (eds.). Brno–Leiden : Lexical Computing CZ s.r.o., 172–192

iSEEK.com. iSEEK - Education. http://www.iseek.com/iseek/home.page (17.6.2019)

Kraus, C., Jermen, N., Jecić, Z. (2017). An insight into online encyclopedias for children and young adults. // INFuture2017: Integrating ICT in Society / Atanassova, I. et al. (eds.). Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, 167–180

Leksikografski zavod Miroslav Krleža. Hrvatska enciklopedija. http://www.enciklopedija.hr/ (17.6.2019)

Longman Dictionary of Contemporary English. Free English exercises. https://www.ldoceonline.com/exercise/ (18.6.2019)

Macmillan Dictionary. Language puzzles. 02.05.2013. https://www.macmillandictionary.com/language-games/puzzles (14.6.2019)

Markopoulos, A. P., Fragkou, A., Kasidiaris, P., Davim, J. P. (2015). Gamification in engineering education and professional training. // International Journal of Mechanical Engineering Education 43, 2, 118-131

Merriam-Webster. Dictionary by Merriam-Webster. https://www.merriam-webster.com/ (17.6.2019)

Montola, M., Nummenmaa, T., Lucero, A., Boberg, M., Korhonen, H. (2009). Applying game achievement systems to enhance user experience in a photo sharing service. // Proceedings of the 13th international academic mindtrek conference: Everyday life in the Ubiquitous Era / Lugmayr, A. et al. (eds.). New York, NY: ACM, 94-97

Ortiz, M., Chiluiza, K., Valcke, M. (2016). Gamification in Higher Education and STEM: A Systematic Review of Literature. // 8th Annual International Conference on Education and New Learning Technologies–Edulearn16 / Gomez Chova, L. et co. (ed.). Barcelona: IATED Academy, 6548-6558

Quian, M., Clark, K. R. (2006). Game-based Learning and 21st century skills: A review of recent research. // Computers in Human Behavior 63, 50-58

RefSeek. 30 Best Online Dictionaries and Thesauri. https://www.refseek.com/directory/dictionaries.html (17.6.2019)

Sitzmann, T. (2011). A Meta-analytic Examination of The Instructional Effectiveness of Computer-Based Simulation Games. // Personal Psychology / Kraimer, M. L. (ed.). New York: Wiley Periodicals, 489-528

Smithsonian Science Education Center. Aquation: The Freshwater Access Game. 26.9.2017. https://ssec.si.edu/aquation (18.8.2019)

The Wiki Game. Wikipedia Game - Explore Wikipedia! 10.2.2018. https://www.thewikigame.com/group (18.8.2019)

Vrijeme i klima hrvatskog jadrana. Pojmovnik. http://jadran.gfz.hr/pojmovnik.html (17.6.2019)

Wikipedia. List of online dictionaries. 24.7.2019 https://en.wikipedia.org/wiki/List_of_online_dictionaries (17.6.2019)

Wikipedia. List of online encyclopedias. 20.7.2019. https://en.wikipedia.org/wiki/List_of_online_encyclopedias (17.6.2019)

Wikipedia. The Wikipedia Adventure. 21.1.2014. https://en.wikipedia.org/wiki/Wikipedia:The_Wikipedia_Adventure (18.8.2019)

https://doi.org/10.17234/INFUTURE.2019.20 Preliminary communication

Entrepreneurship and Service Learning of Students of Information Sciences and Informatics

Hana Josić
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
hjosic@ffzg.hr

Nives Mikelić Preradović
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
nmikelic@ffzg.hr

Summary:
The first aim of this paper was to research the attitudes of students of information sciences and informatics towards acquiring entrepreneurial skills during their study. The second aim was to determine the level and frequency of students' civic (community) engagement. The sample included 211 (mostly) undergraduate students from several public universities in Croatia. The analysis of responses revealed that students perceive most of the entrepreneurial skills as very important to master. It also revealed that they lack opportunities for service-learning and community engagement during their study. These results will be used to design an academic course that links service-learning and entrepreneurship and enables students to acquire skills that they perceive as relevant for their future careers.

Key words: economic and social challenges, civic engagement, service-learning education, skills development, entrepreneurial skills

Introduction

Faced with the economic, environmental and social challenges, education today is in a very unenviable position. To achieve their full potential as adults, students need to develop a range of skills, such as problem solving, critical thinking, communication, teamwork and self-management. A successful career development depends on the aforementioned set of skills that employees should acquire during their studies through learning that is expanded and enhanced by work-based learning, learning-by-doing or any other innovative teaching and learning practice which builds work-related experience that can help students to become more successful in the modern job market (Pizika, 2014). Universities can provide students with the entrepreneurial learning styles needed to develop higher levels of transferable skills and teach them methods to start and grow their own businesses and thus contribute to the society. One way of incorporating entrepreneurship education into study programs is through academic service-learning, a method that enables the application of what has been learned through an academic course to a project aimed at satisfying a community need (Mikelic Preradovic, 2009). Identifying skills that can and should be acquired through such forms of teaching and students' attitude towards them is of great importance in creation of a curriculum that develops student's entrepreneurship skills through service-learning.

Entrepreneurship as a skill is applicable to all walks of life, which is why it can be categorized as transversal. "It enables citizens to nurture their personal development, to actively contribute to social development, to enter the job market as employee or as self-employed, and to start-up or scale-up ventures which may have a cultural, social or commercial motive." (Bacigalupo et al., 2016: 10).

The entrepreneurial skills cannot be completely separated from the personal characteristics, aspirations and motivations of an individual who possesses them. If students are provided with opportunities for continuous reconfiguration or revision of these skills through innovative forms of entrepreneurial education, apart from their personal development, the foundation for a more stable economy is being created simultaneously. The economic challenges that Europe is currently facing have prompted, among other things, the creation of an Entrepreneurship Competence Framework known as EntreComp. EntreComp was developed by the Joint Research Centre (JRC) of the European Commission to improve the entrepreneurial capacity of European citizens as well as their competences. The survey administered in this study (that aimed to research the attitudes of students of information sciences and informatics towards acquiring entrepreneurial skills) was mostly based on the EntreComp framework.

Entrepreneurship education

Entrepreneurial education is a long-life process, which involves putting ideas into practice, innovation, creativity, risk taking, as well as the ability to plan and manage projects to achieve objectives (Mihalache, 2012). It encompasses opportunity recognition capabilities, as well as skills of commercializing a concept, marshalling resources in the face of risk and initiating a business venture (Jones & English, 2004). It is incumbent upon universities to enable students to develop higher levels of transversal skills. "Essentially, a teaching style that is action-oriented, encourages experiential learning, problem-solving, project-based learning, creativity, and is supportive of peer evaluation. It is thought that such a process best provides the mix of enterprising skills and behaviors akin to those required to create and manage a small business" (Jones, English, 2004: 2).

Lackeus (2015) in his paper on entrepreneurship education assumes the development of a culture that moves "about", "for" and "through" entrepreneurship. Teaching “about” entrepreneurship gives a general theoretical understanding of phenomenon, while teaching "for" entrepreneurship provides a requisite knowledge and skills. Finally, teaching "through" entrepreneurship requires students to experience real entrepreneurial learning (Kyro, 2005, as cited in Lackeus, 2015). According to Bacigalupo (2016: 14), "the progression in entrepreneurial learning is made up of two aspects: a) developing increasing autonomy and responsibility in acting upon ideas and opportunities to create value; b) developing the capacity to generate value from simple and predictable contexts up to complex, constantly changing environments."

The goal is to enable a student to achieve his or her own abilities through the development of the ability to identify, discover and evaluate opportunities, obtain required resources, elaborate business plans, establish a business and provide the management. In non-Economics academic disciplines (such as Information sciences), creating a curriculum with elements of entrepreneurship represents a challenge for teachers due to the lack of required teaching skills in entrepreneurship (Mihalache, 2012). While technical and professional skills are offered as a part of Information Sciences and Informatics curricula in Croatia, entrepreneurial skills are still rarely taught, especially at the introductory-level study, although students need to develop such skills well before they start planning their future careers.

Academic service-learning

Academic service-learning, an innovative approach to teaching and learning that brings students, academics, and community to jointly develop solutions for challenging issues (Mikelic Preradovic, 2015) is increasingly used in higher education in many parts of Europe (Millican et al., 2019), but so far to a much lesser extent in post-communist countries. For instance, service-learning education is still underrepresented in academic curricula in Croatia (Mikelić Preradović, Mažeikienė, 2019: 180). Higher education, emphasizing cooperation, democratic citizenship and moral responsibility, connects the wider community through service-learning and prepares students to meet the primary needs of society (Astin et al., 2000). Studies show that academic service-learning helps students gain higher levels of problem-solving skills, critical and creative thinking, communication skills, teamwork, interpersonal and intercultural skills, leadership as well as academic skills and personal and civic values (Astin et al., 2000 , Carrington, Selva 2010; Harris, Jones, Coutts 2010; Milne, Gabb, Leihy 2008; Prentice, Robinson 2010; Rochford, 2014).

It is nowadays used in all academic disciplines, not only social sciences and the emphasis is put on the SL in engineering education programs in many EU countries, since it fits well with the descriptors outlined in the requirements of these programs, such as learning from real life situations and heightening social awareness (Mikelić Preradović, Stark, 2019).

The latest study that measured the impact of SL on students' perceptions of their respective learning processes reports significantly more changes connected to a concept of learning that included contents, behaviors, and personal changes (Macías-Gomez-Estern et al., 2019).

The primary goals of service-learning are structured extensive student reflection, application of learning in real-life settings, and relevant service. It connects students with the community in which they live, encouraging active and purposeful participation in the local community, adopting learning outcomes related to the content of the subject, and finally developing skills and knowledge that help students understand the community needs, which ultimately leads to more active engagement in the community. Students who are working in and for the community enhance their sense of civic responsibility and develop values that underlie action, while improving the quality of life of the entire community.

Methodology

The general research objective of this study was to determine the need for the education combining academic service-learning and social entrepreneurship in the field of Information Sciences and Informatics.

The specific research objectives were:

- to identify students' attitudes and perceptions towards acquisition of entrepreneurial skills,
- to identify students' attitudes and perceptions towards civic (community) engagement,
- to identify the relevant set of skills required to increase student employability and implement the change in the local community,
- to find out whether a university course that combines academic service-learning and social entrepreneurship is needed to help address entrepreneurial skills shortage.

A survey was created to address the research objectives. A volunteer sample consisting of 211 respondents that chose to respond to the online questionnaire was recruited from the beginning of April to the end of May 2019.

As many as 81.74% of the volunteer respondents belonged to the age group of 18-21 years, while the oldest respondent was 26. The volunteer sample included 211 (mostly) undergraduate students of Information Sciences from the Faculties of Humanities and Social Sciences (University of Zagreb, University of Osijek and University of Zadar) as well as students of Informatics from the Faculty of Organization and Informatics (University of Zagreb) and Department of Informatics at the University of Rijeka.

The survey consisted of three parts. The first part examined students' attitudes towards acquiring certain entrepreneurial skills during their studies (through a course on entrepreneurship) and was based on the EntreComp - Entrepreneurship Competence Framework (Bacigalupo et al., 2016), which encompasses 3 competence areas – "Ideas and opportunities", "Resources" and "Into action" – each area including 5 competences which together form "building blocks of entrepreneurship" (Bacigalupo et al., 2016: 1).

The first set of five entrepreneurial skills in EntreComp's framework is represented as "Ideas and opportunities", and covers competences of identifying, harnessing and creating opportunities, as well as consistency in following them: spotting opportunities, creativity, vision, valuing ideas, as well as ethical and sustainable thinking. The ideas include creativity, innovation, risk-taking and the ability to identify positive entrepreneurial patterns as well as usable opportunities.

The other competence area of the framework is named "Resources". These resources represent the entrepreneurial 'know how', financial and economic literacy, self-awareness and self-efficacy, motivation and perseverance, mobilizing resources and others.

The third competence area, "Into action", involves the ability to mobilize and motivate others, taking the initiative, planning and management, coping with the ambiguity, uncertainty and risk, teamwork, collaboration and learning through experience. The order of elements does not imply a sequence in the acquisition process or a hierarchy, i.e. none is more important than the other (Bacigalupo et al., 2016).

After carefully reading each of 44 statements (named "Descriptors" in the EntreComp framework), students expressed a degree of agreement with the statement on a five-point Likert scale ranging from 1 to 5, in which 1 indicates 'not important' and 5 'very important'. The second part of the survey examined students' experiences with community engagement - volunteering and other forms of informal community engagement, aiming to determine the time invested and the most common area of interest. In addition, the general familiarity with the model of service-learning and their views on the strengths and weaknesses of service-learning courses were examined. The last part of the survey collected demographic data.

Results

The statistical analysis of the results is based on descriptive statistics because the volunteer sample is considered biased (the students who were most likely to participate in the survey voluntarily do not necessarily reflect the whole student population).

The first part of the survey examined respondents' views on entrepreneurial skills through three sets of statements (descriptors), each of which relates to one of three areas within EntreComp's model of entrepreneurial skills. The first group of statements refers to the area "Ideas and possibilities". The largest number of respondents (69.67%) considered the descriptor Act responsibly as very important (Figure 1), making it the highest ranked descriptor in this group of statements.

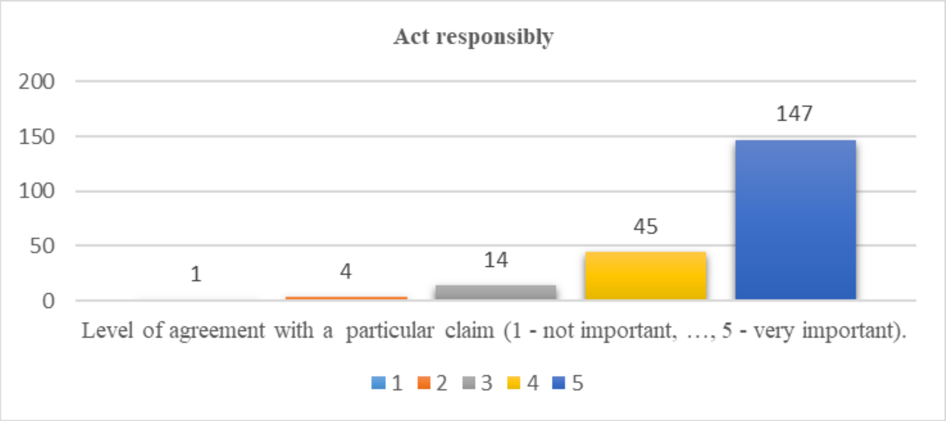


Figure 1. Respondents' opinion on the importance of acting responsibly

The lowest ranked statement in this group was Judge what value is in social, cultural and economic terms. Nobody considered this competence as 'not important', 6.63% of students opted for 'slightly important', 29.38% of respondents positioned themselves as 'neutral', while 36.5% students considered this descriptor as 'important' and 27.5% of respondents rated the development of this ability in the course on academic service-learning as 'very important' (Figure 2).

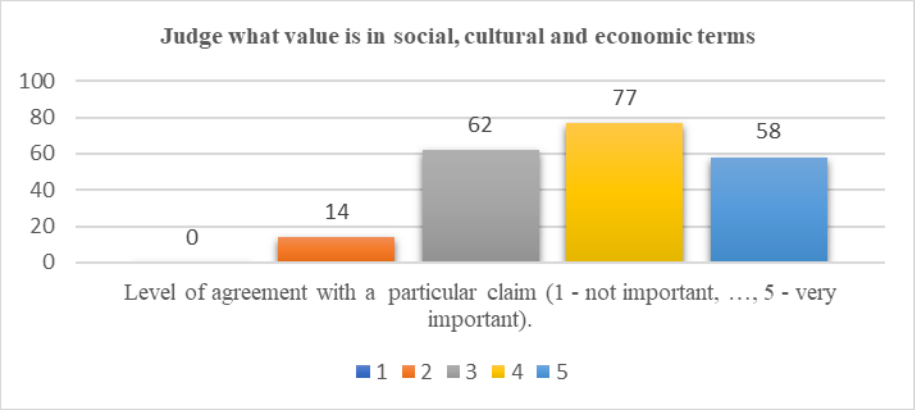


Figure 2. Respondents' opinion on the importance of judging the values in different terms

The second group of statements (descriptors) belongs to the competence area named "Resources", where personal, material and non-material resources are distinguished. The highest ranked statement in this group is Be resilient under pressure, adversity and temporary failure (68.25% of participants rated this competence as 'very important' for future entrepreneurs), while the descriptor Inspire and enthuse relevant stakeholders has the lowest rank in the group, as presented in Figures 3 and 4.

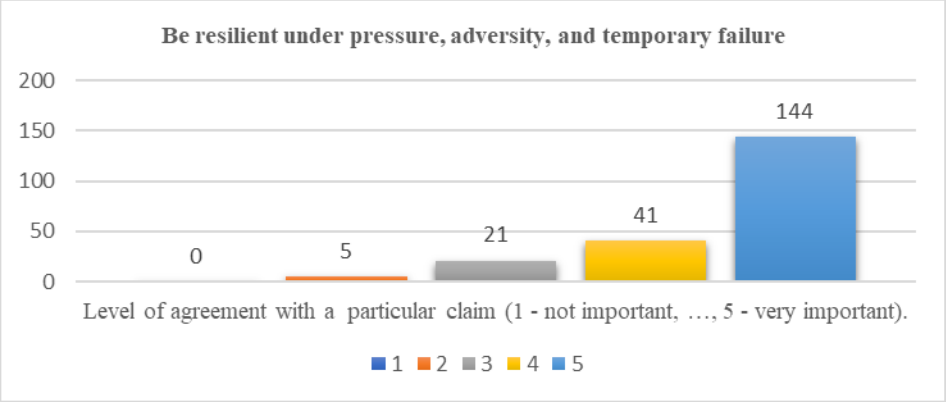


Figure 3. Respondents' opinion on the importance of being resilient

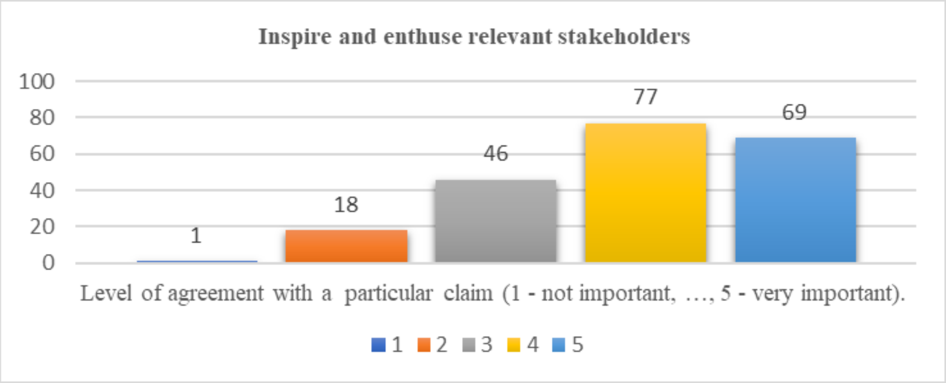


Figure 4. Respondents' opinion on the importance of inspiring and enthusing relevant stakeholders

The last (third) group of statements in this section of the survey relates to the individual's actions. The highest ranked statement in this group is Reflect and learn from both success and failure (your own and other people’s), since 121 students (57.9%), rated this competence as 'very important' and 31.1% of students rated it as 'important', as shown in Figure 5.

By defining this competence as the most important in this group of descriptors, students acknowledged that entrepreneurship education through academic service-learning should enable them to experience real entrepreneurial learning, connect theory with practice and reflect upon it through structured reflection.

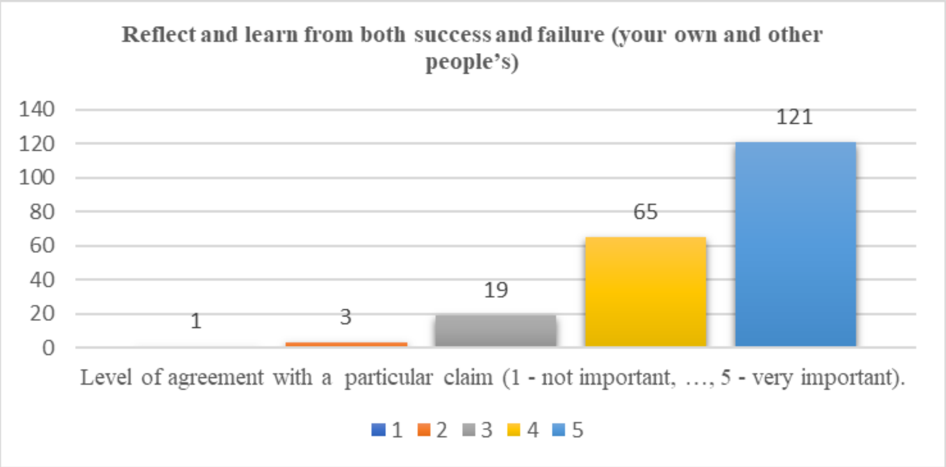


Figure 5. Respondents' opinion on the importance of reflecting and learning from success and failure

As presented in Fig.6, the descriptor with the lowest rank in the group is related to testing ideas and prototypes from the early stages in order to avoid failure. Just over a quarter, 27.14% of students positioned themselves as 'neutral', while rating the development of this skill in the course on academic service-learning.

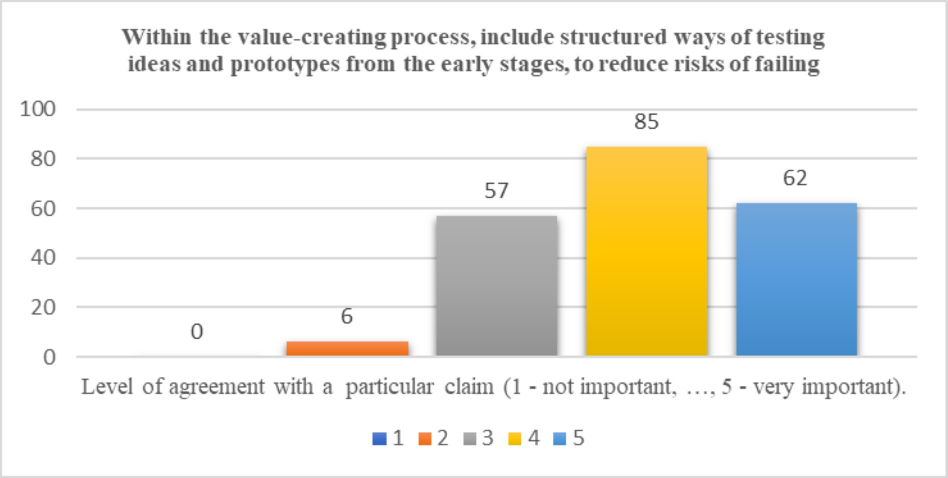


Figure 6. Respondents' opinion on the importance of testing ideas and prototypes to reduce the risk of failure

The results show that the average rating of all three sets of statements in the first part of the survey is approximately equal and very high, 4.2. It is noticeable that students recognize most of the skills and competencies listed in this survey as important or very important for implementation in a course that combines service-learning and entrepreneurship education. These results reveal that students recognize the importance of developing the full range of skills needed to succeed in their careers upon graduation.

The second part of the survey reveals that less than a quarter of respondents, 23% of them, have volunteered within an organization or informally in their community within the past year from completing the survey. In a comparison with the results reported by the National Council for Voluntary Organizations in England in 2018, according to which as many as 39% of young people 16 to 24 years have volunteered at least once in the past year and where those who volunteer make up as much as 28% of the youth population once a month (Department for Digital, Culture, Media and Sport, 2018), this is a small number. However, the number of students in Croatia who made some form of donation during the period of one year before completing our survey is slightly higher. Thus, a total of 80 participants (38%) donated money or goods to an organization. Most often donations were made to organizations dedicated to health, animal welfare, youth and national / local development. However, compared to the statistics for England in 2018 (where 57% of young people aged 16-24 made donations in the four weeks prior to completing their survey), our result indicates that young people in our society are still lagging behind in terms of social awareness, civic and community engagement.

Finally, only 14% respondents attended a service-learning course, showing that majority of students never had an opportunity to take the course on service-learning. These 14% of students have recognized the applicability of skills in numerous areas of life and the identification of community needs as the strong points in the course, while at the same time stating that there were no weak points of the course.

Conclusion

Introducing entrepreneurship education through an academic service-learning course in information sciences and informatics would enables students to acquire skills that they perceive as relevant for their future careers and apply theoretical knowledge or newly acquired skills engaging in meaningful and personally relevant community service while developing capabilities such as problem solving, critical and creative thinking, communication skills, teamwork, interpersonal and intercultural skills, leadership, as well as academic skills and personal and civic values (Astin et al., 2000; Carrington, Selva, 2010; Harris, Jones, Coutts, 2010; Milne, Gabb, Leihy, 2008; Prentice, Robinson 2010; Rochford, 2014).

After analysing the results of a survey conducted in five Croatian public universities, we can conclude that students of information sciences and informatics perceive the following entrepreneurial skills as the most important to master: (a) combining knowledge and resources to achieve valuable effects, (b) developing a vision to turn ideas into action, (c) acting responsibly, (d) being resilient under pressure and temporary failure, (e) making the most of limited resources, (f) demonstrating effective communication, persuasion, negotiation and leadership, (g) defining priorities and adapting to unforeseen changes, (h) networking and (i) reflecting and learning from both success and failure.

Likewise, the results revealed that students lack opportunities for service-learning and community engagement during their study, especially at the undergraduate level. Since service-learning education in Croatia is still in the initial phase of institutionalization, it is implemented top-down, starting from the graduate studies, which is a cause for a small percentage of students participating in a service-learning course.

The results show that there is a need for an undergraduate course that will link service-learning and entrepreneurship and enable students to acquire skills that they perceive as relevant for their future careers, as well as community engagement.

Acknowledgments

This research is supported by the Erasmus+ Knowledge Alliances Programme of the European Union and is part of the project Rural 3.0: Service Learning for the Rural Development. We wish to thank in particular Philine van Overbeeke from the Rotterdam School of Management of the Erasmus University for the design of the survey.

References

Astin, A. W., Vogelgesang, L. J., Ikeda, E. K., Yee, J. A. (2000). How service-learning affects students: Executive summary. Los Angeles: Higher Education Research Institute

Bacigalupo, M., Kampylis, P., Punie, Y., Van den Brande, G. (2016). EntreComp: The Entrepreneurship Competence Framework. Luxembourg: Publication Office of the European Union

Carrington, S., Selva, G. (2010). Critical social theory and transformative learning: evidence in pre-service teachers' service-learning reflection logs. Higher Education Research & Development 29, 1, 45-57

Department for Digital, Culture, Media and Sport. (2018). Community Life Survey 2017/2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/734726/Community_Life_Survey_2017-18_statistical_bulletin.pdf (15.8.2019)

Harris, L., Jones, M., Coutts, S. (2010). Partnerships and learning communities in work- integrated learning: designing a community services student placement program. Higher Education Research & Development 29, 5, 547-559

Jones, C., English, J. (2004). A contemporary approach to entrepreneurship education. Education & Training 46, 8-9, 416-423

Lackeus, M. (2015). Entrepreneurship in education, what, why, when, how. Entrepreneurship 360 background paper. Paris: OECD

Macías-Gomez-Estern, B., Arias-Sánchez, S., José, M., Macarro, M., Regla Cabillas Romero, M., Martínez Lozano, V. (2019). Does service-learning make a difference? comparing students' valuations in service-learning and non-service-learning teaching of psychology, Studies in Higher Education. doi: 10.1080/03075079.2019.1675622

Mihalache, M. (2012). Teaching methods for acquiring entrepreneurial skills in higher education in Romania. 2012. https://www.researchgate.net/publication/267626259_teaching_methods_for_acquiring_entrepreneurial_skills_in_higher_education_in_romania (16.8.2019.)

Mikelić Preradović, N., Mažeikienė, N. (2019). Service-learning in post-communist countries. Embedding Service-learning in European Higher Education / Aramburuzabala, P., McIlrath, L., Opazo, H. (eds.), Routledge, London, 180-195

Mikelić Preradović, N., Stark, W. (2019). Identified service-learning practices in European higher education. Embedding Service-learning in European Higher Education. / Aramburuzabala, P., McIlrath, L., Opazo, H. (eds.), Routledge, London, 109-131

Millican, J., Pollack, S., Zani, B., Stark, W., Mikelić Preradović, N., Aramburuzabala, P. (2019) The changing face of higher education. Embedding Service-learning in European Higher Education. / Aramburuzabala, P., McIlrath, L., Opazo, H. (eds.), Routledge, London, 36-50

Mikelić Preradović, N. (2009). Učenjem do društva znanja: teorija i praksa društveno korisnog učenja. Zagreb: Zavod za informacijske studije

Mikelić Preradović, N. (2015). Service-Learning. Encyclopedia of Educational Philosophy and Theory. Peters, M. (ed.), Springer Singapore, 1-6

Milne, L., Gabb, R., Leihy, P. (2008). Good Practice in Service-learning. Melbourne: Post compulsory Education Centre, Victoria University

Pizika, N. (2014). Embedding Employability Skills in Computer and Information Science Program Curriculum. // World Academy of Science, Engineering and Technology International Journal of Humanities and Social Sciences 8, 2, 381-385

Prentice, M., Robinson, G. (2010). Improving Student Learning Outcomes with Service-learning. American Association of Community Colleges

Rochford, R. A. (2014). Service-Learning: A Vehicle for Enhancing Academic Performance and Retention among Community College Developmental Reading and Writing Students. Service-Learning at the American Community College. Community Engagement in Higher Education. / Traver, A. E.; Katz Z. P. (eds). New York: Palgrave Macmillan

https://doi.org/10.17234/INFUTURE.2019.21

Original scientific paper

Information and Communication Technology in the Rehabilitation of Hearing-Impaired Children

Mihaela Konjevod
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
mihaela.konjevod94@gmail.com

Vesna Mildner
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
vmildner@ffzg.hr

Tomislava Lauc
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
tlauc@ffzg.hr

Summary

This paper deals with the use of information and communication technology in the development of phonological awareness in children with hearing impairment. The research was conducted at Polyclinic for the Rehabilitation of Listening and Speech (SUVAG). The Glaskalica app on a tablet device has been used to compare the phonological awareness test results that are obtained by using the paper-pencil method and through information and communication technology. There were 10 participants aged 6 to 8 years (four girls and six boys). The results show that ICT is interesting and motivating to children and has the potential to improve phonological awareness skills.

Key words: children with hearing impairment, information and communication technology, rehabilitation, education, phonological awareness

Introduction

Today, computer literacy and information literacy are as important as general literacy. That is why children should become familiar with information and communication technology (ICT) from early preschool age. ICT, especially in the rehabilitation and education of hearing-impaired children, plays a significant role in providing visual feedback, keeping attention and motivation. The technology effectively addresses the difficulties faced by the hearing-impaired child during speech development. Nikolić Margan and Bunčić (2016) have demonstrated that use of technology contributes to the prolongation of concentration during rehabilitation, for 15-20 minutes.

The first computer system for speech rehabilitation was developed by Nickerson and Stevens (1973). To date, many systems, programs, and applications for hearing therapist have been developed. The SpeechViewer II program, together with the therapist, has enabled the prelingually deaf boy to establish a proper voice height (Öster, 1995). With the AURIS program, three out of five hearing-impaired children between two and six years old who had not previously verbally communicated learn at least one new term. Children did not show boredom during the teaching process. With the program, the children successfully and correctly pronounce the terms and differentiate the meaning of different sentences with the same term (Sarmaşik et al., 2009).

The ARTUR program increased interest in children with hearing impairment to practice pronunciation. The main advantage of the program is providing feedback in the form of clear instructions on how to improve articulation (Engwall et al., 2006). Nasiri et al. (2017) have developed the game by which children can learn words that they are expected to know by the age of seven. The game consists of an avatar controlled by the child's voice commands. During the game, an object appears in the scene, and avatar collides with that object and object's sound is played and repeated by the child. After the child completes this teaching phase, s/he enters the test phase and system show us her/his improvement by giving validation recognition percentage" (Nasiri et al., 2017: 6).

Research on the impact of information and communication technology on education and rehabilitation of children with a hearing impairment has shown positive results. More attention is being given to applications that can be used on a tablet and mobile devices because it is proved that these devices are more appropriate and easier for pre-school and early school-age children (Geist, 2014). Besides, it is proved that there is a positive correlation between the use of ICT and academic achievement of pupils with hearing impairment (Egaga, Aderibigbe, 2015) and a positive correlation between playing on tablet and self-esteem (Bahatheg, 2014).

Phonological awareness

The development of reading skills is one of the most important tasks of early education. This is also one of the most challenging tasks faced by children with severe and profound hearing impairment (Harris, Beech, 1998). "Phonological awareness refers to the recognition, creation and handling of smaller parts of the word, such as recognizing the words that are being rhymed, counting the syllables, separating the beginning of the word from the end, and separating the letters in the words" (Pavliša, Lenček, 2011: 2). Rakhshanfadaee and Salehi (2016) researched first-class pupils who had moderately and profound hearing impairment. They proved that the pupils had acquired phonological awareness skills and could make phonological decisions. Children with profound hearing impairment had greater difficulties than those with severe impairment, but both groups had weaker results compared to hearing children and those with moderate hearing impairment. It has been shown that the degree of hearing damage also plays a major role in the development of phonological awareness. Cognitive abilities, short-term verbal memory, and language comprehension are also important. This means that a child needs to understand a particular speech part, keep it long enough in memory, and then realize it verbally (Sindik, Pavić, 2009). Also, the beginning of the rehabilitation process and use of a cochlear implant or other hearing aids also plays a major role. It should be mentioned that these and many other factors influence the results obtained by examining phonological awareness in this study.

Problem statement and research goal

The influence of ICT on the development of phonological awareness in children with impaired hearing is insufficiently explored. Besides, in Croatia, there is hardly any research on ICT in the rehabilitation and education of children with hearing impairment. Also, it is necessary to develop the appropriate information and communication technologies, programs and applications so that the health and pedagogical system can follow the changes in the modern era. The purpose of this study was to raise awareness of the importance of ICT use in working with hearing-impaired children. In Croatia, this is a group for which the tools and applications have recently been developed. The goal of this study was to investigate the effectiveness and importance of ICT in working with hearing-impaired children. The research objectives were:

- To find out whether the results of phonological awareness are better if the information and communication technology is used.
- To find out whether therapists and hearing-impaired children have a positive attitude towards information and communication technology.

Methodology

The research was conducted in two meetings, individually with children and their therapists. In the first meeting, the children recognized missing phonemes in words on paper (the first, the last, and all phonemes in the word). The words were taken from the Glaskalica app. In the next meeting, the children recognized missing phonemes in words in the Glaskalica app. Participants filled out the questionnaires regarding phonological awareness and motivation.

Participants

Ten participants aged six to eight who have impaired hearing were involved in this study. They all go to the kindergarten or first grade of primary school in Polyclinic SUVAG. There were six boys and four girls, which is a total of two participants at the age of eight, six at the age of seven and two at the age of six. The participants where A, B, C, D, E, F, G, H, I and J. There were also four therapists: 1, 2, 3 and 4. This is shown in Table 1.

Table 1. Therapists, work experience in rehabilitation and child participants

Therapists	Experience in rehabilitation	Child participants
1	21 years	G
2	9 years	I, J
3	25 years	C, D
4	25 years	A, B, E, F, H

Materials

In this research, ICT-AAC Glaskalica app was used. Glaskalica has been developed within the ICT-AAC project "Competence network based on information and communication technologies for innovative services aimed for people with complex communication needs". The established competence network enables collaboration, knowledge transfer and technology for people with complex communication needs. The ICT-AAC Glaskalica app helps in improving phonological awareness. Phonological awareness is important for reading comprehension. The application involves recognizing the first, the last, and all phonemes in words. In addition, the application distinguishes tasks according to their complexity. There are over 200 words in the application that have been selected by the experts. Each word is presented to the user by an image representing the term corresponding to the given the word. The words can be pronounced so that users, besides the visual template, can also have a sound template. That is very useful and important for hearing-impaired children. On the home screen, the user can choose one of the three games. The games differ according to the position and the number of phonemes that must be chosen. Before starting the game, the user selects the complexity of the words that will be offered during the game. Easier or difficult tasks can be selected. After selecting the desired complexity, the image appears on the screen. The user has to guess the first, the last or all phonemes, depending on the selected game. Children with normal speech development first recognize the words that rhyme. Then they can separate the word into syllables. Five-year-olds can recognize the first and the last phoneme in the word. The ICT-AAC Glaskalica app is the first application in the Croatian language that encourages the development of phonological awareness. The app is intended for use on portable devices such as tablet and mobile devices. The previously developed technology for stimulating these aspects is based on computer programs, which is maybe not the best for pre-school children.[1]

Questionnaires

The research methodology employed in this study was based on self-made questionnaires. There were five questionnaires. The questionnaire for testing phonological awareness was made using the words from the ICT-AAC Glaskalica app. There were ten different words in which the first phoneme needed to be guessed, ten words in which the last phoneme needed to be guessed and ten words where all phonemes in the word needed to be guessed. This questionnaire was used for testing the phonological awareness skills by the paper-pencil method. The questionnaire for monitoring the child's response while using the app was created to record correct and incorrect answers of children. If they did not know the right answer from the first attempt, the answer was considered incorrect. Because three phonemes are given to them and if they don't know the right answer from the first attempt, they are finding the answer throughout the method trial and error, and that is not representing their true phonological awareness. The questionnaire for examining how much the children enjoyed using the Glaskalica app consisted of eight questions. A 5-point Likert scale was used in this questionnaire. The questions were: "Did you like this game?" "Did you get bored while playing this game?", "How hard was it to play this game?", "Would you like to play this game again?", "Would you like to play this game together with your friends or family?", "Was this game too long?", "Was this game too short?", "Have you already played a similar game, if you did what is the name of the game?" The last two questionnaires were made for therapists. In the first one, they present their own opinions and impressions on how the child behaves in a rehabilitation class where information technology was used. The questions were: "How much did a child like this game?", "Was the child bored when playing this game", "How hard was it for the child to play this game?" and the space for comments if

[1] http://www.ict-aac.hr

there is any. Consideration was also given to the therapists' opinion on how much a child was satisfied because they are experts who are working with the child constantly and they can recognize the child's satisfaction or dissatisfaction. In the second questionnaire, they provided answers to the questions: "How much did you like the app?", "How useful would this application be for rehabilitation?" To address these questions, a 5-point Likert scale was used in the questionnaires for therapists.

Testing phonological awareness

The research was conducted in two meetings, individually with each participant for 15 to 25 minutes, depending on the abilities of the individual participant. It started at the end of September 2018 and was completed in December 2018. The research was carried out in the room for individual rehabilitation in kindergarten and room for individual rehabilitation in elementary school of the Polyclinic for the Rehabilitation of Listening and Speech, SUVAG (with the consent of the parent and the participants). During the first meeting, each participant independently solved the tasks of recognizing the first, the last, and all phonemes in words on paper. There were three groups of tasks, consisting of ten tasks each (ten for the first, ten for the last and ten for all phonemes in the word). Under each word, three phonemes were offered, of which the participant should decide for the one (s)he heard. Firstly, the word was read to the participant, if necessary, a few times, and then (s)he independently selected a single phoneme using the pencil. Each next word was read, and then the participant chose the phoneme (s)he heard from the offered phonemes and so on to the end. During the second meeting, participants solved the tasks of recognizing missing phonemes using the Glaskalica app. It was possible to distinguish tasks in the application according to their complexity. Thus, more manageable tasks were used. The questions from the questionnaires regarding motivation were asked after the end of the testing session.

Results and discussion

Measurement of phonological awareness was carried out in two different ways. The aim was to compare the results obtained by using the paper-pencil method and the results obtained with the ICT. The average test score in recognition of the first phoneme by using the paper-pencil method was 94%, with results ranging from 80% to 100%. In recognition of the first phoneme using the app, the average score was 86% with results ranging from 60% to 100%. One participant scored 60%, and three scored 70%. Also, they achieved the lowest results in the overall survey. Children recognized the first phoneme in the word better if the words were on paper. In this case, the use of ICT has not given better results (Figure 1).

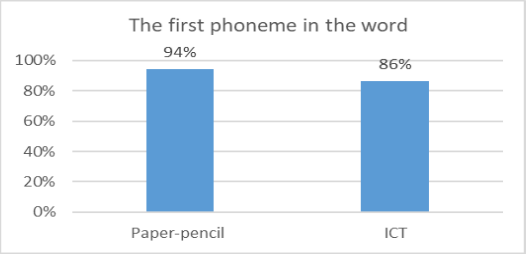


Figure 2. The average test score of recognizing the first phoneme

In recognition of the last phoneme using a paper-pencil method, the average test score was 80%, with results ranging from 40% to 100%. In recognition of the first phoneme using the app, the average test score was 89%, with results ranging from 50% to 100%. When using ICT as learning support, children recognized the last phoneme slightly better compared to paper as a medium (Figure 2).

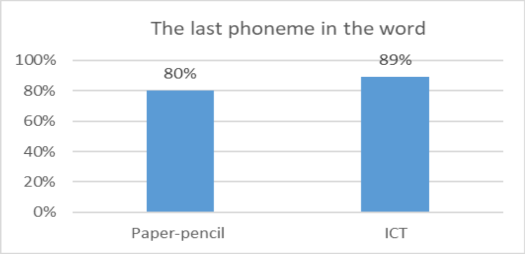


Figure 2. The average test score of recognizing the last phoneme

The average test score of the recognition of all phonemes in the word with the paper-pencil method was 52%, with results ranging from 10% to 100%. In recognition of the first phoneme through the information and communication device, the average test score was 56%, with results ranging from 10% to 100%. In this case, children were slightly better at recognizing the order of all phonemes in the word using the app (Figure 3).

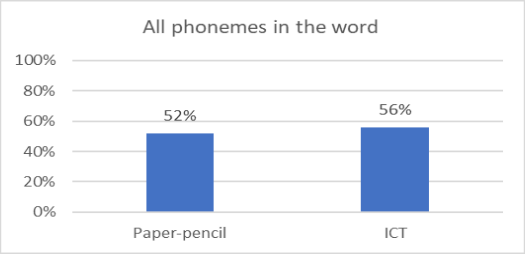


Figure 3. The average test score of recognizing all phonemes

Although children recognized the first phoneme in the word slightly better if the word was on paper, out of the total number of participants, 50% had the same maximum score or even better score in recognizing the first phoneme using the app. Out of the total number of participants, 90% had the same maximum score or better score in recognizing the last phoneme using the app. Out of the total number of participants, 50% had the same maximum score or better score in recognition of the order of all phonemes in the word using the app. After the end of the testing session, the questions from a pre-prepared questionnaire were asked to identify participants' interest in the application. Response from the therapist is presented in parallel with the child's answers, as it is shown in Figure 4. The question was: "How much did you like this game?" The child and the therapist were able to choose a number from 1 to 5 (1 meaning non-liking at all, 5 meaning liking very much). There was a slight difference regarding participants D, F, and G and therapists.

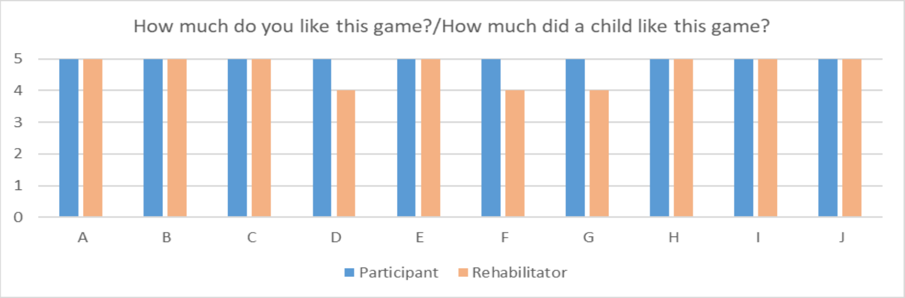


Figure 4. Comparison of children's and therapists' answers.

As shown in Figure 5, the question was "Did you get bored while playing this game?" The child participants were able to choose a number from 1 to 5, where number 1 represents the absence of boredom, and number 5 represents a great presence of boredom. Nine out of 10 participants (90%) responded with number 1, such indicating that it was not boring at all while playing this game. Only one participant responded with number 3. Because of the difficulty of establishing meaningful communication with participant F, the therapist's response indicated to us that the child liked the application.

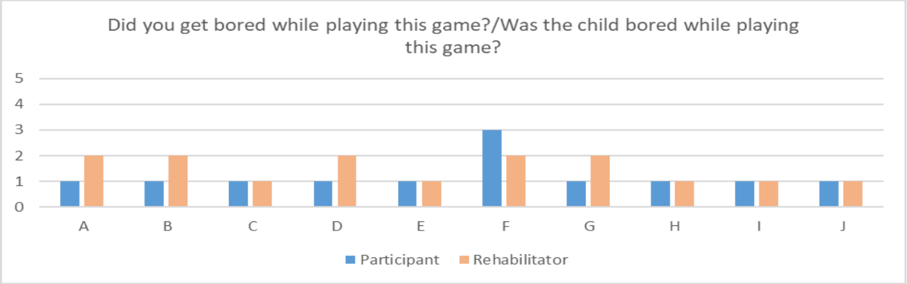


Figure 5. Comparison of children's and therapists' answers

As shown in Figure 6, the question was "How difficult was it to play this game?" The participants were able to choose a number from 1 to 5. Number 1 indicated that the game was not at all difficult, and the number 5 indicated that it was very difficult. Seven out of the 10 participants (70%) have chosen number 1, such indicating that it was not difficult at all during the game. Three out of the 10 participants have chosen number 3, such indicating that this game was neither difficult nor easy. For two participants, this game was difficult or very difficult. The answers of the child participants approximately match the answers of the therapists. A large difference between participants and therapist responses is found regarding participant B since the therapist saw how much time a child needed to choose the right answer.

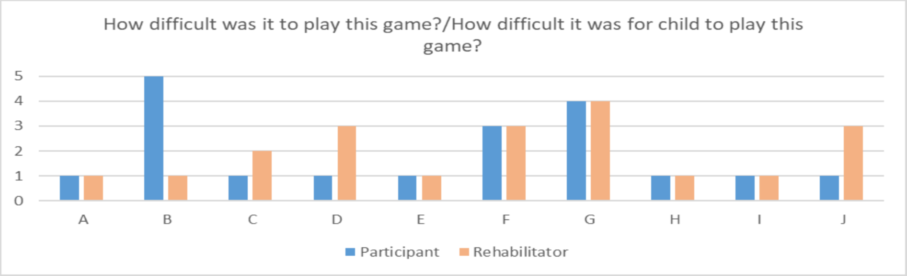


Figure 6. Comparison of children's and therapist's answers

For children participants, there were 5 questions with yes/no responses. All participants answered that they would love to play this game again. Nine out of 10 participants answered that they would like to play this game together with friends or family, and that game was not too long or too short. Five out of 10 participants answered that they have already played a similar game, although, they did not know the name of the game. The answers from the questionnaire for therapists showed that they all like the app and find it useful or very useful. Also, all of them think that some form of ICT should be used in the rehabilitation of hearing impairment.

Limitations of the research

The words on paper were the same for all participants. But on the tablet, the words were automatically generated by the application. Therefore, for results that were better or worse by using ICT, we can't say with certainty that these are solely due to the use of ICT. To use such game-designed apps, in the rehabilitation of children with hearing impairment, it should be possible to choose the words that are going to be used at a rehabilitation class. Also, the research was carried out on a small number of subjects and a suitable sample. It should be mentioned that the group of children was heterogeneous with respect to their level of speech development.

Conclusion

The study shows that children with hearing impairment have a positive attitude toward information and communication technology and that they are happy to use it. This is also the opinion of their therapists. All therapists have realized the importance and the need to use the Glaskalica application as well as other educational applications. The results indicate that ICT has the potential to improve phonological awareness skills. What should be the main goal of each rehabilitation method is to enable the child with hearing impairment to achieve speech communication and to develop language and their abilities. If information and communication technology can help in that process, it should be included in the pedagogical and health care system, thus increasing the chances of most children with impaired hearing to develop their skills for communication.

References

Bahatheg, R. O. (2014). Deaf children and iPad technology: Improving the self-concept of deaf and hard of hearing children. // Canadian International Journal of Social Science and Education 1, 107-120

Egaga, P., Aderibigbe, I., Akinwumi, S. (2015). Efficacy of information and Communication technology in enhancing learning outcomes of students with hearing impairment in Ibadan. // Journal of Education and Practice. 6, 30, 202-205

Engwall, O., Bälter, O., Öster, A. M., Kjellström, H. (2006)- Designing the user interface of the computer-based speech training system ARTUR based on early user tests. // Behaviour and Information Technology 25, 4, 353-365

Geist, E. (2014). Toddlers through preschool: Using tablet computers with toddlers and young preschool. // YC Young children 69, 1, 58-63

Harris, M., Beech, R. J. (1998). Implicit phonological awareness and early reading development in prelingually deaf children. // Journal of Deaf Studies and Deaf Education. 3, 3, 206-216

Nasiri, N., Shirmohammadi, S., Rashed, A. (2017). A serious game for children with speech disorders and hearing problems. // IEE 5th International Conference on Serious Games and Applications for Health, 1-7

Nickerson, R., Stevens, N., K. (1973). Teaching speech to the deaf: can a computer help? // Audio and Electroacoustics, IEEE Transactions. 21, 445-455

Nikolić Margan, A., Bunčić, A. (2016). Razvoj aplikacije za pomoć u rehabilitaciji djece sa govorno-jezičnim poremećajima. // Sveučilište u Zagrebu, 1-63

Öster, A. M. (1995). Teaching speech skills to deaf children by computer-based speech training. //Proceedings of 18th International Congress on Education of the Deaf, Tel Aviv, Israel

Pavliša Ivšac, J., Lenček, M. (2011). Fonološke vještine i fonološko pamćenje: neke razlike između djece urednog jezičnog razvoja, djece s perinatalnim oštećenjem mozka i djece s posebnim jezičnim teškoćama kao temeljni prediktor čitanja. // Hrvatska revija za rehabilitacijska istraživanja 47, 1, 1-16

Rakhshanfadaee, A., Salehi, M. (2016). Phonological awareness in children with hearing loss. // The Hearing Journal 69, 9, 32-35

Sarmasik, G., Serbetcioglu, B., Kut, A. (2009). Computer aided education and training tool for hearing impaired children: AURIS. // ICL Proceedings, 427-433.

Sindik, J., Pavić, M. (2009). Povezanost općih kompetencija i fonološke svjesnosti kod predškolske djece. // Život i škola 22, 2, 62-77

https://doi.org/10.17234/INFUTURE.2019.22

Preliminary communication

Media Freedom and Regulation in the Context of Reporting on National Security Issues

Dora Gelo
University of Zagreb, Croatia
doragelo@gmail.com

Summary

The conflict in focus – the relationship between the concept of secrecy, which the national security system is entitled to, and the request of the public, which is a norm of a democratic society – is only one of the elements appearing in a specific relationship between the system and the media. On the other hand, as stated by authors Peter Gill and Mark Phythian in their book Intelligence in an Insecure World (2018), a part of their relationship can also be marked by a combination of dependence, manipulation, support and praise, which can lead to conflicts of other relevant categories from the group which are related to independence and the quality of information. A similar situation involving the relationship with the source is also conspicuous here – the relationship between the journalist and the source, the system and the source or, also according to the authors mentioned above, the possibility of fraudulent, including illegal, behaviour, in order to obtain the information inaccessible to the public.

Key words: media regulation, secrecy, national security, public interest, self-censorship, democracy, oversight

Introduction

This paper will look into the notion of media regulation, with special attention to self-regulation in relation to national security system in a broader sense, defined by Grizold (1994), as well as in relation to the public interest, since the basic security of the state and its citizens is unquestionably in the public interest as well as media freedom. By national security system, according to Grizold (1994), we consider mechanisms that ensure "capacity of the states to protect their basic social values against internal or external threats (i.e. to maintain peace and guarantee freedom), to prevent danger and fear – but also their ability to ensure social development as well as well-being of their population". Furthermore, Grizold (1994) holds the security as an immanent structural element of society in a way that "it involves a state in which the balanced physical, spiritual, psychical and material existence of an individual and the community as a whole is ensured in relation to other individuals, communities, as well as to natural environment". The national security system in modern states has started to embrace and discuss this broadest definition but still, when it comes to the national security system we still stand more with the traditional concept including a few non-military innovations like ensuring quality of life (food, environment protection). Starting from the basic principle that media freedom cannot be absolute, controversial points should be carefully examined in the framework of current circumstances which are very different from, for example, the circumstances in the late 19th-century France, when the Declaration of the Rights of Man and of the Citizen (which covered the freedom of expression and the freedom of the press) was adopted. Different circumstances mostly mean the development of the media, as well as the multilateral aspect of political communication, the speed of information, the frequency of the change of relationships on all levels included in the transfer of socially relevant information. In its essence, media regulation has existed for several centuries. Between the 16th and the 19th century, in West Europe and North America a battle was fought in the name of political freedom and human rights against publishing limitations and for the industry, including copyright. Over time, when new media appeared, development has been noticed in terms of discussions, framework and modalities. Today, when we talk about media regulation, we also talk about the reasons which legitimise such a procedure, so that the achieved level of freedom of thought, expression and right to information is not lost or significantly reduced by such procedures.

The main aim of the paper is to research the concept of media regulation in the context of national security in the broad sense, especially media self-regulation as the possible tool for maintaining optimal use of the media freedom in achieving goals related to the public interest in defined context.

Research will be done by analysing one representative case study (dismissal of the SOA director Dragan Lozančić, 2016). In addition to the case study, qualitative content analysis and narrative analysis are applied as well as analysis of the appropriate theories, analytical jurisprudence and a background analysis of included professions.

Case study

On 5 February 2016, the website Dnevnik.hr (2016a) published a communication from the Office of the President of the Republic of Croatia:

“The President of the Republic of Croatia, Kolinda Grabar-Kitarović, has signed the Decision on the dismissal of the Director of Security and Intelligence Agency, Dragan Lozančić, due to a breach of the Security and Intelligence System Act.

The President personally informed Director Lozančić that she had lost her confidence in him.

The President of the Republic of Croatia, Kolinda Grabar-Kitarović, has also signed the Decision on the dismissal of the Head of Office of the National Security Council (UVNS), Ivica Panenić, due to the failure to ensure the implementation of activities within the scope of the UVNS.

The decisions on the dismissals have been delivered to the Government of the Republic of Croatia.”

It was later reported (Vlašić, 2016) that the Government had confirmed that they had received the Decision of the President of the Republic of Croatia (referred to in the text below, except in quotes, as the President) on the dismissal of the Director of the Security and Intelligence Agency (referred to in the text below, except in quotes, as Director) due to a breach of the Security and Intelligence System Act: “We have received an initiative, which the President started in accordance with the Act on the Security and Intelligence System of the Republic of Croatia.”

Relevant provisions of the Act on Security and Intelligence System of the Republic of Croatia

Article 66 of the Act on Security and Intelligence System of the Republic of Croatia defines elements of the procedure. The appointment or dismissal is co-signed by the President of the Republic and the Prime Minister. Pursuant to the said article, Director may be relieved of his/her duty if he/she: requests it personally; is incapacitated for the performance of his/her duties; does not implement the decisions of the President of the Republic and the Prime Minister which direct the work of the security and intelligence agency or does not implement their measures related to the oversight of work; violates the Constitution, laws or other rules and regulations; exceeds or abuses his/her authority; violates the confidentiality of classified data; and is convicted for a crime which renders him/her unworthy of the position. The procedure for the dismissal of the Director may be initiated by the President of the Republic, the Prime Minister and the Croatian Parliament. When the dismissal procedure is initiated by the President of the Republic or the Prime Minister, the Croatian Parliament may be asked for an opinion before a decision on the dismissal is issued. When the dismissal procedure is initiated by the Croatian Parliament because of illegality of the work of the agency or its employees discovered in the oversight procedure, the President of the Republic and the Prime Minister issue a decision on the dismissal.

Comparative analysis of the selected articles

Units for analysis are articles from the few national newspapers (online editions) chosen by convenience sampling in the period from 5 February 2016 to 4 May 2016 as the relevant period for concerned case study, defined for the purpose of this work. Samples are chosen in order to get various perspectives of the studied event considering usual practice referring to the level of critical thinking, as well as by basic political orientation. Both criteria, meaning the level of critical thinking and the basic political orientation, are identified qualitatively by observation, together with the assessment of each newspaper/author accordingly. A certain number of articles were included and further exploration stopped when the findings started to repeat themselves unlikely to open a new perspective. All the links are added to the references at the end of the paper.

In addition to the case study, qualitative content analysis and narrative analysis are applied as well as analysis of the appropriate theories and analytical jurisprudence.

According to the claims published on 5 February 2016 on Telegram.hr (Vlašić, 2016), Prime Minister Tihomir Orešković (referred to in the text below, except in quotes, as the Prime Minister) replied to journalists that the President and he would jointly decide on the dismissal of the Director. A communication from the Government later stated that the Prime Minister would decide on the President's initiative the following week.

Going back to the first news in the media regarding the decision of the President (Dnevnik.hr, 2016a), one may notice that at the very beginning the media used different concepts: decision and initiative. The ambiguous and inconsistent use of the notions immediately created the effect of confusion over the situation. As part of the analysis of this moment, it should be stated that the Government confirmed (Vlašić, 2016) that the President had initiated the dismissal procedure according to the Act, which means that she had made her decision on the matter. It was therefore completely wrong to use the word "initiative".

The second element that should be highlighted is that the media failed to report or insufficiently reported about the important fact that there is a legally specified alternative way of initiating the dismissal procedure in terms of the President of the Republic/Prime Minister as it is not explicitly stated that signatures must be simultaneous. The consensus of the two decision-makers is implied in the content of the whole provision. However, the structure of the provision allows them to be successive, which again means that the procedure was actually initiated in accordance with the relevant law.

Further development of the said story encouraged the usual narratives in the Croatian public space, such as the differentiation between the left-wing and the right-wing, defamation and the like. For example, in article entitled "The Life of the Head of the Main Croatian Secret Service, Former Hard-Core Rightist and Supporter of Gojko Šušak has Become Zoran Milanović's Favourite Staff Member" (Korbler, 2016), Director is called a semi-dismissed chief of Croatian intelligence agents. The article gives a rather detailed description of the Director's actions, which seem impeccable, and it subtly introduces doubts into the President's actions and her contacts obviously trying to undermine the credibility of the President's decision as well as her own credibility. It only briefly mentions that ex-president of the Republic of Croatia Josipović did not find her dismissal without a cause disputable and that he stated that during his mandate some secrets had never reached his office. Another article (Toma, 2016) also questioned legality of the President's decision. It claimed that the violation of law would be a basis for a dismissal, but that it should be determined in a disciplinary procedure, criminal proceeding or at least by a conclusion of the Parliamentary Committee, although this is not explicitly set by the Act. It is not clear if that is opinion of the author or introductory paraphrasing of the following statement of Ranko Ostojić, President of the Committee, who insisted on legality and in his opinion lost confidence is not legally justified reason for a dismissal. On 10 February 2016 Kamenjar.hr (2016) also reports professor Jurčević's claims and comments and, among others, lists the reasons for the President's decision confirmed by other media outlets. SOA intercepted the President, and the transcripts were then delivered by the Director to then-Prime Minister Milanović. On 26 January 2016, it is additionally reported (Žabec, 2016) that SOA first informed the State Attorney's Office of the Republic of Croatia (DORH) about all of this, and only then the President. However, the President was mainly dissatisfied with the fact that she had not been warned that the people who, at the time, were surveilled had been trying to approach and contact her.

These few examples show that it is possible to comment the (il)legality of the President's decision in various ways in order to direct contextual perception of that specific information.

On 4 May 2016, Dnevnik.hr (2016b) published that Daniel Markić had been appointed as the new SOA Director, following the opinion given by the Parliamentary Committee and as proposed by the President and the Prime Minister. In the Parliamentary Committee, the ruling party had been against the appointment, whereas the opposition had voted for the appointment.

The gaps (the left vs. the right, the defamation, etc.), materialised as the lack of a meaningful discussion, are the opposite of the essence of the freedom of public communication. The encouragement of a meaningful discussion, on the other hand, would have been a suitable practice of freedom of public communication because, since the story about the indisputably sensitive issue, in regards to the actors and content as well, had already been started in the media. The public may have had a deeper and more coherent knowledge about the matter and this, in turn, would have provided the opportunity to have more public discussions about the relevant legal solution and fewer games with the low, sensationalistic elements of the story.

Thus directed reporting might then have imposed the need to evaluate the provisions regarding the agreement of the responsible people in a socially and politically important matter related to the national security, such as the dismissal of the SOA Director. It also might have started a public discussion regarding the formulation of the alternative procedure initiation and at the end of the day regarding the whole Act. Said issues are stated as examples allowing the possibility of the emergence of new issues regarding this matter, which would steer the discussion towards a constructive, rather than completely destructive information process.

The fact that the participants in the story contributed to the media perception of the event should not be disregarded: (1) The President stated that there was a serious breach of trust; (2) By delaying the signing, the Prime Minister made the citizens distrust the most important state institutions; (3) The President, to whom safeguarding the stability of the state is a basic duty, possibly initiated the procedure without discussing it with the Prime Minister etc. This also grazed the part of freedom which refers to the officials' freedom of expression and the classification of information confidentiality, where media exploitation of (partial) information may or may not be counted on. However, if one only looks at the media perception, it would be irresponsible to claim that. It would also represent the violation of the purpose of the freedom of expression because individual responsibility would be presumed, or even claimed, without a single piece of valid evidence other than "our gut feeling", common partner of the disinformation spiralling which, once it has been started whether intentionally, because of disregard or by coincidence, very often nobody controls.

If at the core of the freedom of the press is the citizen, and at the core of the ethics of journalism is the freedom of expression and the right of the citizen to information and information dissemination (Cayrol, 1997 quoted in Jergović, 2003), texts which report on the said topics are expected to cover more specifically the roots of the created problems, unclear legal provisions and the potential use of the system for the purposes of political clashes. However, one finding – a wide variety of image deformations which cannot be corrected by an exception or two – opposes this idea.

In a parallel simulation, the moment when the story disappeared from the media could be the moment when the story was created. Alternatively, those points in time could be characterised as mitigation of a crisis in which the system was so overexposed to the media that it did not contribute to national security. In the long run, the case – which is by no means the only one and during which a story was being told about the instability of the entire system of government, including its essential part which refers to national security and which especially encouraged the feeling that the relationship between the President and the Prime Minister was not good – creates the image of extremely bad relationships between institutions rather than between people. This is where the truth of the information and the relevance of information, as well as its purpose in the context of public interest in a wider sense, start being questioned. Just how much media representation of bad relationships, verging on an incident, contribute to the security of the country (even when that information is true) should be able to be expressed in certain models of communication, where variables such as the ones stated above, or the variables of time, international relations and internal horizontal and vertical political relationships, would be interconnected.

Freedom and regulation

In his text "Right to the Freedom of Expression of Thoughts", Mato Arlović (2016) elaborates the notion of freedom in terms of the constitution and the law. He perceives it as a higher concept and divides it into several lower concepts; a) freedom of expression of thoughts, b) freedom of speech and speaking in public, c) freedom of the press and other media, and d) freedom of founding any media institutions. Furthermore, he emphasizes that the freedom of thought and expression of thought, guaranteed by the Constitution of the Republic of Croatia, is indirectly supplemented by the rights and the content which are substantially related to it and influence its implementation. They are: a) prohibition of censorship, which should be interpreted as prohibition of official state bodies' oversight of the media, b) constitutional guarantee of the right to the access to information, c) reasons for limiting the right to access to information, d) constitutional guarantee to the right to correction to anyone whose right guaranteed by the Constitution or the law has been infringed. What should be mentioned here is the appearance of journalists' self-censorship, which prof. Smerdel refers to as "danger" (Smerdel, 2013 quouted in Arlović, 2016), an opinion Arlović also agrees with. Contrary to the stated opinion, in this paper self-censorship will be depicted as an institute whose possibility should not be mystified or discarded. A different view mostly derives from a different description of self-censorship. It is understandable that the description – "the danger of self-censorship, whereby the people in the media, aware of the circumstances they are working in and the risks related to free reporting, pay special attention not to hurt the feelings of the government and its officials" – does not seem promising for the journalists' profession or their readers. However, viewed synthetically, this set of words tells us that journalists choose not to write because they do not want to be vulnerable (hurt), or they do not want to hurt somebody important, which does not seem convincing. If the aim of media activity is public interest, it would be useful to perceive self-censorship as an inhibition mechanism for the purposes of ensuring true freedom of reporting, as well as personal filtering and an invitation to deliberation. The issue of the meaning of the notion "public interest" could also be raised, but that goes beyond this paper. However, it is unlikely that subtle differences in the perception of public interest would significantly erase important elements such as truthfulness of information, relevance of information, clarity and coherence of reporting. In addition, it is true that public interest is everything that, as McQuail (2010) states, is widely considered essential for long-term benefits to the society and its members.

Self-censorship is similar to media self-regulation, which will probably not survive on its own, without legal help. However, it may contribute to better legal solutions, better understanding of the problem, and the development of intrinsic motivation for the improvement of quality instead of strict external legal regulation. The aforementioned is closely connected with media accountability defined by McQuail (2003) as the orientation process claiming that "responsible communication exists where authors (gatekeepers) answer for the quality and consequences of publication, are oriented towards the audience and others who are affected by the publication and respond to their expectations and to the expectations of the society as a whole". In a wider context, self-censorship, or more accurately self-regulation, should be connected with media transparency, which can broadly be defined as a transparent relationship between the journalist and the source of information. In that sense, article entitled How Effective Is Media Self-Regulation? Results from a Comparative Survey of European Journalists (Susane Fengler et al., 2015) presents the results of research on media accountability, conducted on journalists from 14 countries. In addition, with its empirical data, the research contributes to the debate on the future of media self-regulation in Europe. For the purposes of this paper, from a series of results the author will highlight the one revealing the attitude towards media transparency. It shows that journalists from Central, Eastern and Southern Europe (the participants were journalists from Romania, Poland, Spain and Italy) are more sceptical about the concept. While the journalists from Northern and Western Europe are convinced that transparency related to journalists' treatment and publications of corrigenda and apologies develop more trust in the media, their counterparts in the parts of Europe listed above believe that this, together with newsroom transparency, damages the trust between the journalist and the audience. Research also shows that, although journalists from all over Europe univocally support the statement "Journalistic responsibility is the precondition for the media freedom", actual support for the concept of media self-regulation is not great as journalists question the efficacy of the existing apparatus (journalists' councils, ombudsmen, etc.) On the other hand, laws (regulation) and company instructions were highly graded.

Normative media theory

McQuail (2010) defines media theory as a complex structure of socio-political and philosophical principles which organises ideas on the relationship between the media and the society. One type of this theory is the normative media theory, which deals with the issue of what the media should do rather than what they are really doing. The premise of the immersion of the media in a concrete society is important in terms of the fact that dominant ideas about media obligations will correspond to other social values and processes, which in liberal societies means freedom, equality before the law, social solidarity and cohesion, cultural diversity, active involvement, and social responsibility. Basic varieties of normative theory are: authoritarian theory, theory of the free press, theory of social responsibility, theory of development, alternative theory. In reality, there are no clean models; rather, what exists in a concrete society is a model combining theoretical elements and media types.

Normative theory is important because it, according to McQuail (2005), plays the role in shaping and legitimizing media institutions. Furthermore, McQuail (2010) claims that there are differences in problem analysis. However, they lie more in the ways of dealing with the problem (regulation/self-regulation, competition). He also believes that basic principles of media activity can be isolated: independence, diversity or pluralism, information quality, preservation of social and cultural order. He warns that some of these principles are conflicting, but that one of the aims of regulation is the very management of tensions and settlement of conflicts. The conflict in focus – the relationship between the concept of secrecy, which the national security system is entitled to, and the request of the public, which is a norm of a democratic society – is only one of the elements appearing in a specific relationship between that system and the media. On the other hand, as stated by authors Peter Gill and Mark Phythian in their book Intelligence in an Insecure World (2018), a part of their relationship can also be marked by a combination of dependence, manipulation, support and praise, which can lead to conflicts of other relevant categories from the group which are related to independence and the quality of information. A similar situation involving the relationship with the source is also conspicuous here – the relationship between the journalist and the source, the system and the source or, also according to the authors mentioned above, the possibility of fraudulent, including illegal, behaviour, in order to obtain the information inaccessible to the public.

Conclusion

Key problems of reporting on national security and variables of that information process have been identified by analysing suitable theoretical premises and one case study. The nature of reporting does not depend solely on the media; it also depends on those involved in the process. However, that does not minimise the responsibility of the media which, thanks to the freedom, are entitled to use when doing their job, to a certain extent control public information space and shape public knowledge. The media are sometimes also considered to conduct oversight sui generis of the intelligence and security system, which should not be a problem in a democratic society. However, by following and analysing media publications, it is safe to conclude that no significant aim beneficial to public interest was actually achieved in most of the analysed units. Moreover, as stated above, what was created was a confusing situation which, at certain points, became more serious information chaos. In his text Regulation by Revelation, Richard Aldrich (2009) offers specific revelation models, induced/allowed by the systems of Western countries, particularly in relation to the end of the Cold War and, in principle, the opening of the system to the public. He also mentions big globalisation trends, internet development, and whistle-blowers. In addition to the models, Aldrich also discusses problems which appeared at the time. He concludes his text with a part entitled Regulation by Revelation? In it, he mentions several similarities between the said factors – the media and the national security system. He offers the functions of informing and enlightening as examples. He also mentions the similarities between work methods and interaction, which is more frequent than it is perceived by the public. The author emphasises that, on the path toward a discretely different legal treatment of reporting about the national security system, future guidelines should provide solutions for cases when journalists believe that the revelation of information is in the public interest, as well as for the opposite cases, when information is not published in order to prevent obvious harm. Considering the said statement, the author of this text believes that the fragile line between constructive public criticism of political institutions and destructive, purposeless incoherent publications with an agenda or negative motivations, and the line between the requirements of the public and the secrecy requirement, should be studied continuously. By compiling examples of the freedom of public communication in the context of satisfying the needs of public interest related to national security and their adequate methodological processing, it is necessary to produce test patterns within the concept of media freedom, or media self-regulation in the best-case scenario, for future research in order to create a model of information analysis in defined context.

References:

Aldrich, R. (2009). Regulation by revelation? Intelligence, the Media and transparency // Dover, R., Goodman, M. S. // Spinning Intelligence, Why Intelligence Needs Media, Why Media Needs Intelligence. New York: Columbia University Press, 13-37

Arlović, M. Pravo na slobodu izražavanja misli (ustavnopravni okvir i ustavnosudska praksa u Republici Hrvatskoj. // Zbornik radova Pravnog fakulteta u Splitu 53, 2, 377-411

Bajruši, R. (2012). Tko je Dragan Lozančić? Milanović bi na čelu SOA-e bivšeg člana HDZ-a kojem je bio šef prije 12 godina. 11 September. http://www.jutarnji.hr/vijesti/hrvatska/tko-je-dragan-lozancic-milanovic-bi-na-celu-soa-e-bivseg-clana-hdz-a-kojem-je-bio-sef-prije-12-godina/1556970/ (30.4.2019)

Dnevnik.hr. (2016a). Predsjednica potpisala razrješenje ravnatelja SOA-e i čelnog čovjeka Vijeća za nacionalnu sigurnost. 5 February. http://dnevnik.hr/vijesti/hrvatska/predsjednica-kolinda-grabar-kitarovic-smijenila-ravnatelja-soa-e-dragana-lozancica---425582.html, (30.4.2019)

Dnevnik.hr (2016b). Daniel Markić potvrđen za novog ravnatelja SOA-e. 4 May. http://dnevnik.hr/vijesti/hrvatska/sjednica-saborskog-odbora-za-unutarnju-politiku-i-nacionalnu-sigurnost---435844.html (30.4.2019)

Dnevnik.hr (2016c). Tko je Dragan Lozančić, čovjek koji je podijelio vladajuće? 8 February. http://dnevnik.hr/vijesti/hrvatska/biografija-dragan-lozancic-ravnatelj-soa-e---425776.html (30.4.2019)

Fengler, S. et al. (2015). How effective is media self-regulation? Results from a comparative survey of European journalists. // European Journal of Communication 30, 3, 249-266

Gill, P., Phythian, M. (2018). Intelligence in Insecure world. Oxford: Polity Press

Grizold, A. (1994). The concept of national security in the contemporary world. //International Journal on World Peace 11, 3, 37-53

Jelinić, B. (2015). Ekskluzivno: Tajni dosje o kriminalu u Sigurnosno obavještajnoj agenciji. 19 July. http://www.nacional.hr/ekskluzivno-tajni-dosje-o-kriminalu-u-sigurnosno-obavjestajnoj-agenciji/ (30.4.2019)

Jergović, B. (2003). Zakonske promjene i tisak u Hrvatskoj od 1990-2002. // Politička misao: časopis za politologiju 40, 1; 92-108

Kamenjar.hr (2016). Razgovor: Josip Jurčević: Udbaški sustav ne želi se odreći vlasti pa provodi državni udar! 10 February. http://kamenjar.com/josip-jurcevic-udbaski-sustav-ne-zeli-se-odreci-vlasti-pa-provodi-drzavni-udar/ (30.4.2019)

Korbler, J. (2016). Životna priča šefa glavne hrvatske tajne službe- nekad tvrdi desničar i simpatizer Gojka Šuška postao je omiljeni kadar Zorana Milanovića. 7.2.2016. http://www.jutarnji.hr/vijesti/zivotna-prica-sefa-glavne-hrvatske-tajne-sluzbe-nekad-tvrdi-desnicar-i-simpatizer-gojka-suska-postao-je-omiljeni-kadar-zorana-milanovica/96556/ (30.4.2019)

McQuail, D. (2003). Media Accountability and Freedom of Publication. Oxford and New York: Oxford University Press, 2003

McQuail, D. (2005). Mass Communication Theory.Los Angeles, London, New Delhi, Singapore, Washington D.C.: SAGE Publications Ltd., SAGE Publications Inc., SAGE Publications India Pvt Ltd., SAGE Publications Asio-Pacific Pte Ltd.

McQuail, D. (2010). Module 2, Unit 11: Media Regulation. https://www.le.ac.uk/oerresources/media/ms7501/mod2unit11/index.htm. Leicester: University of Leicester, Department of Media and Communication, 2010 (30.4.2019)

Toma, I. (2016). Nema zakonske osnove za smjenu šefa SOA-e? Predsjednica i premijer sve su dogovorili, ali Lozančić mora prvo ponuditi svoju ostavku. 30 March 2016. http://www.jutarnji.hr/vijesti/hrvatska/nema-zakonske-osnove-za-smjenu-sefa-soa-e-predsjednica-i-premijer-sve-su-dogovorili-ali-lozancic-mora-prvo-ponuditi-svoju-ostavku/38143/ (30.4.2019)

Vijesti.hr (2016). Komentar:Cvrtila:'Predsjednica je izvršila veliki pritisak na premijera, osporavanje njezine odluke o smjeni Lozančića bi dovelo do krize'. 6 February. http://www.vijesti.rtl.hr/novosti/hrvatska/1927308/cvrtila-predsjednica-je-izvrsila-veliki-pritisak-na-premijera-osporavanje-njezine-odluke-o-smjeni-lozancica-bi-dovelo-do-krize/ (30.4.2019)

Vlašić, T. (2016). Predsjednica potpisala odluku o smjeni šefa SOA-e, ali premijer za to, čini se, nije znao. 5 February. http://www.telegram.hr/politika-kriminal/predsjednica-razrijesila-ravnatelja-soa-e-dragana-lozancica-osobno-mu-rekla-razlog-smjene/ (30.4.2019)

Žabec, K. (2016). Politički vrh želi promjenu na čelu SOA-e Traži se hitna smjena šefa Dragana Lozančića. 26 January. http://www.jutarnji.hr/vijesti/hrvatska/jutarnji-ekskluzivno-otkriva-politicki-vrh-zeli-promjenu-na-celu-soa-e-trazi-se-hitna-smjena-sefa-dragana-lozancica/91615/ (30.4.2019)

https://doi.org/10.17234/INFUTURE.2019.23

Review paper

Open Source Intelligence (OSINT)
Issues and Trends

Tomislav Dokman
Faculty of Humanities and Social Sciences University of Zagreb, Croatia
dokman.tomislav@gmail.com

Tomislav Ivanjko
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
tivanjko@ffzg.hr

Summary

Open Source Intelligence (OSINT) is an intelligence product which has been processed, analysed and obtained from the publicly available information. It should be actionable and disseminated in a timely manner to the appropriate audience. Open source intelligence transfers specific knowledge to beneficiaries for them to use it in their actions and the decision-making process. Even though intelligence has for years been considered a state activity, a new postmodern intelligence paradigm (resulting from the post-Cold War period), along with the new security threats and technological improvement within the information and communication technology changed the understanding of intelligence and open source intelligence. It is believed that the bulk of intelligence, according to some estimates as much as 80 percent, comes from publicly available sources, which unquestionably indicates that this specific intelligence knowledge can be created outside of the traditional intelligence environment. The positive characteristics of publicly available information is easy availability and velocity, variety of topics, the ethical component and the low cost of its collecting; while on the other hand, in the negative context, it is characterized by manipulative character, disinformation potential, fragmented truth, contradiction and mistrust. Since public domain is a suitable platform for spreading influence, various actors in this domain are trying to have an impact on the information, making it difficult to evaluate, compare and analyse. This paper tries to explore the key advantages and disadvantages of open source intelligence in the context of intelligence activities.

Key words: open source intelligence, information, intelligence, knowledge, disinformation

Introduction

Open source intelligence (OSINT) is a gathering intelligence discipline which consists of collecting raw data accessible to everyone from publicly available sources. By further processing and analysing, this raw data is turned into actionable intelligence knowledge; or in other words, into an intelligence product, which is a crucial segment of intelligence activity. The end intelligence product, most often called intelligence information, is at the same time the crown of complex quality intelligence work since, by timely classification and distribution of intelligence knowledge to its ultimate users, the decision-making process is simplified, made quicker and easier, not only when it comes to various political decisions but also to a variety of business decisions. The traditional intelligence process is finished with this act. What preceded it was getting a request, fact finding, processing and analysis, followed by the beginning of a new cycle. It is also interesting to note here that up to date, there is no concrete record of the first use of open source intelligence. It has been claimed that the term itself, as well as the activity of collecting intelligence from publicly available sources, have by now probably been in use for hundreds of years (Hassan, 2018). Despite the fact that for a long time intelligence activity was considered to be an exclusive function of the state, the new postmodern views that came into being as a consequence of the period after the Cold War, new security challenges and technological progress within the framework of the information and technology channel, changed the understanding of the concept of both intelligence and open source intelligence. The fact that in the course of the World War II, open source intelligence was predominantly associated with the state bears witness to that. Today, the creation of actionable knowledge using open source intelligence tools constitutes an essential quality of various non-state organizations and business sector corporations. Namely, it is against the law and it is a punishable activity for the private sector to collect strictly confidential and/or classified information by means of industrial espionage (Hulnick, 2002: 566). Hence, the creation of new knowledge by using open source intelligence activities has recently been conducted exclusively by non-state participants. It is considered that most intelligence knowledge, according to some estimates even up to 80%, comes from public sources (Wallner, 1993; Hulnick, 2002; Riley et al., 2005), which is undeniable proof that actionable knowledge can be created from publicly available information. In other words, we can discern the level of reach of open source intelligence and its role in the decision-making process. Although it is the non-state actors that have predominantly used the public knowledge, it is wider professional and academic communities that have a controversial view of the part and nature of opensource intelligence in the intelligence work of state agencies while creating an intelligence product by means of information obtained from public sources. A part of the academic and professional community considers the gathering of secret information the quality of intelligence work and intelligence knowledge only (Random, 1958; Warne, 2002; Johnston, 2005; Gil, Phythian, 2012, Warner, 2009). It needs to be pointed out that the authors listed here are primarily experts, who consider the creation of intelligence knowledge from their own specific paradigm, unfoundedly depreciating open source intelligence and giving support to the superiority of classified sources. It is Liaropulus (2006: 8) who points out that in the recent past, information was not sufficiently accessible and it was mainly a product of covert and expensive operations, while in the contemporary world, owing to the Internet, all this information is easily available, cheap and more tangible. Because of this, in this paper I would like to explain what open source intelligence is, what the open source intelligence cycle is, as well as explore key advantages and disadvantages of open source information.

Open source intelligence and the intelligence cycle

When dealing with intelligence, the term is above all determined by a definition controversy, since in its wider sense, it is identified through organization, knowledge, information and process; while in its narrower sense, the idea of intelligence refers to analysed information as the basis for decision-making. On the other hand, Tuđman (2002) states that the term intelligence refers to the knowledge one has about the enemy, the activity of gathering and the organization that deals with intelligence activities, including the procedure of collecting, processing, analysing, connecting and evaluating the information at our disposal. In other words, there is intelligence potential in all collecting intelligence disciplines, no matter what the origin of data collected is. It is crucial that the data collected is structured, processed, evaluated, correlated, analysed and timely disseminated to their end users. The actionable character of an intelligence product or information is among its essential qualities.

The development of the Internet has, in our contemporary world, together with the availability of information in the digital form, had considerable impact on the growth in the quantities of open source information. The way information is gathered has also changed – until recently, it was available mostly in the printed or audio-visual form. On the other hand, with technological progress and the appearance of the Internet, an opportunity was created to gather information desired at a greater speed, no matter whether the content was displayed in one’s mother tongue or a foreign language, from another country or another continent. According to Mercado (2004), it is with one movement and one click of the mouse that “analysts and officials understand the world.” In other words, knowledge has never before been within such an easy reach, it has never been as widespread as today.

As for defining the term, one of the founders and one of the people who have done the most on advancing open source intelligence information, Robert David Steele, emphasises that OSINT is “unclassified information that has been deliberately discovered, discriminated, distilled and disseminated to a select audience in order to address a specific question (Steele, 2006: 129). On the other hand, Williams and Blum (2018: 8) give a definition of OSINT as “publicly available information that has been discovered, determined to be of intelligence value, and disseminated by a member of the IC.” “Open source intelligence is also a specific intelligence collection discipline such as human intelligence (HUMINT), intelligence gathered by interception of signals (SIGINT), imaginary intelligence (IMINT) and the scientific and technical processing of data collected from different moving and immovable sources (MASINT) or measurement and signature intelligence (NATO 2001).” In other words, it is a discipline whose predominant feature is the availability of unclassified information. According to Mercado (2005) “open sources often equal or surpass classified information in monitoring and analysing such pressing problems such as terrorism, proliferation, and counterintelligence.” For Mercado (2005), the space of public and secret information is often interwoven with frequent meandering from the clandestine space to the public area and the other way around. The author states that it is in the public space that we often find information obtained by leaking secret information, while state systems are characterised by obtaining intelligence information from secret sources and are quite often an amalgam of open source information. It is essential to point out here that OSINT “products can reduce the demands on classified intelligence collection resources.” (Steele, 2006: 129) Steele is of the opinion that collecting data in a secret way should be reduced to a minimum, or that they should be gathered in such ways only when we know what it is that is still missing in a certain puzzle. Steele in fact here deals with the problem of optimizing state resources, with quality planning being a precondition for that, which in a concrete case means collecting the publicly available, and only after that the less readily accessible, classified information. Matey (2005: 8) thinks that “OSINT is changing the traditional conception of intelligence”, and as a consequence, there will be more and more situations where the private sector enters the field of intelligence and the process of creating intelligence knowledge. Today, there are numerous organizations which, by using open sources advance intelligence knowledge and create analytical work in the area of national security based on publicly available information. Some of the better-known ones are the American think-tank organization, RAND Corporation, along with Jane’s Information Group, and the British BBC Monitoring. The traditional open source intelligence process (request - collection - processing - analysis - dissemination) shows the cyclically interconnected parts exchanging within a circle. It is a never-ending process; after the dissemination of intelligence knowledge in the form of information is done, it goes on and a new cycle of collecting unclassified data begins. The process is initiated with a previously defined request, or an ad-hoc task, and it is with this aim that raw data is collected again. When consulting open sources, raw data is gathered from the Internet, i. e. various web pages, blogs, social media, social networks, and by accessing the traditional media such as television, radio and newspapers, including grey literature and geographical data (Hassan, 2018). Before open source information is disseminated, it is of key importance to structure, evaluate, explain and correlate the data collected so far, which is done in the data-processing phase, in the phase of analysis, when the end intelligence product is made, prior to its dissemination to the end users. In order to fully comprehend the value of OSINT, it is of crucial importance to point out the American view of it, according to which “information does not have to be secret to be valuable” (CIA, 2010), in other words, an open source piece of information can be equally useful, assuming it is obtained in time and is actionable in its character. In such cases, open source data is collected, mostly raw, the data is then sorted out, structured or processed, thus making them adequate for analytical processing and the creation of an intelligence product for dissemination to ultimate users.

The advantages and disadvantages of OSINT

In order to discuss the positive and negative features of open source information, we need to point out that the value and usefulness of this particular collection discipline outweighs its certain disadvantages and the negative traits of this discipline. These features of OSINT are summarized in Table 1 and Table 2.

Table 1. The advantages of OSINT

Advantage	Characteristics
The simplicity of obtaining information	Technological development and the Internet have resulted in simple ways of obtaining open source information. To get open source information, one needs a PC and Internet access. “This is the reason why OSINT is more accessible, ubiquitous, and valuable.” (Mercado, 2004: 47)
The speed at which a great quantity of information is gathered	Digital data are suitable for generating great quantities of information. They also make it possible to search through a great volume of available

	open source information within a very short time-span, which is not a feature of any other intelligence discipline.
Covering a wide spectrum of content units	The advantage of open source work is the coverage of a variety of national security topics, which is not possible in covert information gathering by using one human source or other intelligence collection disciplines.
Searching for data from the recent past	This discipline is superior when it comes to the investigation of post festum events, when it is necessary to collect a wider information pool about an event from the past, and investigate and/or a reconstruct a certain, especially recent, past event.
Low level of danger (safe model)	Here we are dealing with a risk-free way of information gathering, which is why this discipline can be described as the most acceptable data-collecting technique, as far as matters of security are concerned. In other words, exploiting open source information access can diminish or optimize the necessity to use much more dangerous and much more complex ways to gain factual knowledge with the help of human sources.
Open- source pieces of information do not cost much	Considering the fact that open source data is publicly available, mostly they do not have to be paid for. Still, there are certain magazines, studies and articles in think-tank organizations that can be used on condition they are given an extra fee.
Real time information gathering	This particular data gathering discipline makes it possible for us to keep track of certain events and phenomena in real time. It is possible to follow migrants' routes, investigate the consequences of a terrorist attack, outburst of an epidemic, the organization of violent demonstrations and the like.
Availability of information from various languages	Another advantage of this discipline is also obvious when it comes to looking through content in various foreign languages and having the chosen content translated into our mother tongue by means of machine translation.
Collecting information from a single spot	This secret data collecting discipline from human sources is characterised by the fact that information can be collected without going to some place, environment or society, i. e. the required data is collected directly from our offices.
Simple data dissemination procedures	Considering the fact that here we are dealing with the data readily available to everybody, their dissemination should not be based on spreading them restrictively or in a limited way; such pieces of information are suitable for both horizontal sharing and they can be made available to the broader public in general.
Secret information authentication/ verification	The comparison of classified information with publicly available data is an outstanding method of checking whether there is maybe some "secret" data being circulated in open sources. There is a possibility that some data were previously publicly known, so this method can be used to ascertain the validity and evaluate both the content and the source of covert information, at the same time channelling further collection of classified information.
Business intelligence suitability	Using open source data can definitely be helpful in bringing adequate business decisions, planning future projects, conducting pilot projects, market research, comparing and evaluating competitors, as well as ensuring appropriate publicity for one's own business ideas.

Other authors have also been studying the good and bad sides of open source intelligence information and their opinions concur with the insights given here. For instance, Annie Ahuja (2018: 470) states the advantages of open source intelligence are: 1) Less Expensive; 2) Accessing Information; 3) Security; 4) Business; 5) Social Media; 6) Updated Data and Metadata; 7) Semantic Understanding; 8) Applications. Simultaneously, Mercado (2005) emphasises OSINT's "value advert, speed, quantity, quality, clarity, ease of use and cost." OSINT is a unique discipline that can share its product with everyone, the profit being greater if we take into consideration the fact that it is also shared with humanitarian aid missions and forces of security, law and order. The fact that the results of OSINT's inquests are used as contributions to early warning systems also needs to be born in mind (Steele, 2006: 133). On the other hand, it is important to point out that OSINT has not been put in place to replace other intelligence disciplines in the production of actionable knowledge, i.e. it is not aimed at one discipline replacing an existing one.

Table 2. The disadvantages of OSINT

Disadvantage	Characteristics
Great quantities of accessible information	Public space is characterised by great quantities of available information. Quite often the data is badly laid out and therefore confusing, and then it is also difficult to find the exact data we are looking for. Hassan (2018) thinks that we live in an information world, with millions of people in constant communication and constantly sharing all kinds of information. Namely, the daily data production, according to some estimates, amounts to about 2.5 quintillion bytes, the assumption being that this number is bound to increase even further with further Internet growth (Marr, 2018)
Information contradiction	The availability of a wide information fund has been caused by the democratization of the net, coupled with the opportunity of an individual to create and spread information without too much difficulty. It is important to objectively view the data collected in order for the product of public intelligence knowledge to be as accurate as possible.
The manipulative character of publicly available information	Public space is obviously an excellent area for spreading disinformation and fake news in order to achieve certain effects. By influencing the corpus of public knowledge, using information and counter information in the public arena, those in charge of various campaigns want to gain information superiority (Akrap, 2009: 80). It is due to this reason that this discipline is characterised by the constant checking and rechecking of all the material at our disposal.

Conclusion

Until the emergence of the Internet and the state-of-the-art communication and information-sharing applications, open source intelligence information was considered to be of low value as an intelligence discipline. Not accepting the real value of open source intelligence mostly corresponded with an a priori attitude that an intelligence product could exclusively come from covert information sources, while work with publicly available data was considered to be less valuable and less interesting for the intelligence activity. However, in the current, above all changing security environment, when traditional threats became more divergent and changed in their configuration, especially when the gathering of secret data became a long-lasting, expensive and complex process, intelligence services began creating open source intelligence knowledge. Besides that, this collecting discipline also proved to be an invaluable source of knowledge apart from state activity and has, therefore, been fully assimilated into other spheres of public life. Moreover, the intelligence activity, including open source intelligence information, is no longer only a feature belonging to the state itself. Gathering, analytical processing and the production of relevant and actionable intelligence knowledge are now typical for both the state and non-state actors. However, it is envisaged that in the future, OSINT will be faced with great challenges caused by immense quantities of unstructured data, because in order to collect them and sort them out appropriately, an ever-increasing development of sophisticated software will be required. Since we are here dealing with the greatest source of information of all intelligence disciplines, great effort will be necessary to extract quality and timely information from a considerable number of sources, in order to attain usable actionable knowledge. At the same time, it needs to be pointed out that public space is an area for spreading influence, propaganda and a variety of interests, which is why the checking of information has to be the main feature of work when dealing with the publicly available content.

References

Akrap, G. (2009). Informacijske strategije i oblikovanje javnoga znanja. // National Security and the Future 10, 2, 77-151

Annie, A. (2018). A Review Paper on Open Source Intelligence: An intelligence sustenance. // International Journal of Recent Trends in Engineering & Research 4, 4, 463-474

Best, C. (2007). Open Source Intelligence. // Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and their Applications to Security / Fogelman-Soulie, F., Perrotta, D., Piskorski, J., Steinberger, R. (eds.). Amsterdam: IOS Press, 331-344

CIA. (2010). INTellingence: Open Source Intelligence. 2010/07/23. https://www.cia.gov/news-information/featured-story-archive/2010-featured-story-archive/open-source-intelligence.html (2019/08/05)

Hassan, N. (2018). An Introduction To Open Source Intelligence (OSINT) Gathering. 2018/08/12. https://www.secjuice.com/introduction-to-open-source-intelligence-osint/ (2019/08/10)

Hulnick, A. S. (2002). The Downside of Open Source Intelligence. // International Journal of Intelligence and CounterIntelligence 15, 4, 565-579

Johnston, R. (2005). Analytic Culture in the US Intelligence Community: An Ethnographic Study. Washington: Central Intelligence Agency

Liaropoulos, A. N. (2006). A (r)evolution in intelligence affairs? In search of a new paradigm. Athens: Research Institute for Europe and American studies (RIEAS)

Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. 2018/05/21. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#12e73d5f60ba (2019/08/11)

Mercado, S. C. (2004). Sailing the Sea of OSINT in the Information Age: A Venerable Source in a New Era. // Studies in Intelligence 48, 3. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol48no3/article05.html (2019/08/15)

Mercado, S. C. (2005). Reexamining the Distinction Between Open Information and Secrets. // Studies in Intelligence. 49, 2. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/Vol49no2/reexamining_the_distinction_3.htm (2019/08/15)

Matey, G. D. (2005). Intelligence Studies at the Dawn of the 21st Century: New Possibilities and Resources for a Recent Topic in International Relations: UNISCI Discussion Papers. May. https://www.ucm.es/data/cont/media/www/pag-72533/Gustavo2.pdf (2019/08/11)

NATO (2001). Open Source Intelligence Handbook. November 2001. https://archive.org/details/NATOOSINTHandbookV1.2 (2019/08/19)

Random, R. A. (1958). Intelligence as a science. 1958, last updated 2011. https://www.cia.gov/library/center-for-the-study-of-intelligence/kent-csi/vol2no2/html/v02i2a09p_0001.htm (2019/08/10)

Riley, J. K.. Treverton, G. F., Wilson, J. M., Davis, L. M. (2005). State and Local Intelligence in the War on Terrorism. Santa Monica: RAND Corporation

Steele, R. D. (2006). Open Source Intelligence. // Handbook of Intelligence Studies / Johnson, Loch K. (ed.). New York : Routledge, 129-147

Tuđman, M. (2002). Informacijska znanost i izvjesnice. // Informatologia 35, 4, 244-251

Wallner, P. F. (1993). Open Sources and the Intelligence Community: Myths and Realities. // American Intelligence Journal 14, Spring/summer, 19-24.

Warner, M. (2002). Wanted: A Definition of Intelligence: Understanding Our Craft. // Studies in Intelligence 46, 3. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol46no3/article02.html (2019/08/11)

Warner, M. (2009). Intelligence as risk shifting. // Intelligence Theory: Key questions and debates / Gill, P., Marrin, S., Phythian, M. (eds.). London: Routledge, 16-32

Williams, H. J., Blum, I. (2018). Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise. Santa Monica: RAND Corporation

https://doi.org/10.17234/INFUTURE.2019.24 Original scientific paper

The Role of New Media in Building Social Skills of Students with and without Disabilities

Eva Brlek
University North, Koprivnica, Croatia
evbrlek@unin.hr

Ljerka Luić
University North, Koprivnica, Croatia
ljluic@unin.hr

Jelena Škoda
University North, Koprivnica, Croatia
jeskoda@unin.hr

Summary

Digital technologies are an integral part of our everyday life, and new media have changed the way we communicate. New forms of communication are most evident among school children, whose upbringing and education is marked by the immersion in the world of social media and internet technologies. The authors of this paper aim to analyze the influence of new media on the development of school children's social skills. It is based on a quantitative study of the impact of the length of daily usage of new media on elementary school students' social skills (with and without disabilities), seen from the perspective of both parents and teachers. The results were gained by the survey conducted on a representative sample of primary schools in the Republic of Croatia in the school year 2018/2019. Comparative analysis of the results obtained with similar research results led to a conclusion of the synergistic impact of new media and new communication channels on the development of social skills of students. The authenticity of this research is established in the approach involving both parents and teachers and the material on which the research was conducted (students with and without disabilities). The scientific contribution was achieved at different levels: at the cognitive level through the development of information concepts of social skills, while at the methodological level through the development of original methodology of qualitative comparison of respondents' attitudes by making a statistical analysis of dynamic ontological parameters. Finally, on the practical level through the application of defined information concepts in the study of the impact of new media on the development of social skills.

Key words: new media, new communication channels, social skills, students with and without disabilities

Introduction

The upbringing and education of children and young people for modern life in the 21st century is directed towards gaining competences that are essentially different from those prevailing in the past. The question that arises in the broader social context, enriched with modern information and communication technologies, is how much and to what extent are social skills of children affected by new media. The term new media most commonly refers to internet content, computer multimedia, video games, and virtual reality. But the definition of the term new media is more complex than it seems. Lievrouw and Livingstone (2006: 2) define new media as 'Information and communication technologies and their social contexts with three key components: devices used to communicate or transmit information; activities and practices in which people engage in communication and social patterns that develop by using these devices and practices." The adjective "new" in the name itself, implies a dichotomy with respect to some "old" media. Analyzing a wide range of changes in media production, distribution, and use, Lister et al. (2009: 13) consider changes in media from a technological, textual, conventional, and cultural point of view, and find that the media we define as new are "digital, interactive, hypertextual, virtual, networked and simulated." Because of these characteristics, the media is an extremely important socializing agent in today's world. Socialization is broadly defined as “the process during which people acquire attitudes and values of a particular culture and learn behaviors that are considered appropriate for individuals as members of a particular society” (Raboteg-Šarić, 1997: 423). The most important immediate agents in the primary socialization process of children and young people are parents, siblings, peer groups, pre-schools and educational institutions.

Table 1. Social skills

Basic social skills	**Secondary social skills**
Listening	Seeking help
Starting a conversation	Asking for information
Holding a conversation	Giving directions
Asking questions	Following directions
Ending a conversation	Providing assurance
Introducing oneself	Apologizing
Introducing others	Self-control

Adapted from Rajić, A., (2011.), http://oc-pgz.hr/clanci_socijalne_vjestine.html

Developed social skills of an individual are the foundation of all relationships and one of the best predictors of healthy functioning in adulthood. In order for a child to maintain good face-to-face relationships with peers and adults, he or she must master basic social skills like listening, holding a conversation, seeking help or playing a game (Miljković, Rijavec, 2002). Children with well-developed social competences, have high self-esteem and self-respect; feel supported and loved by their peers, parents and other adults; they feel included and wanted rather than isolated; they are surrounded by many people who they can spend time with; they are involved in large, organized social groups such as sports clubs, religious groups, classes; they have people in their life who they can share their experiences with, their thoughts and feelings. Generally, they have a positive outlook on life and are open-minded so other people want to be around them (Klarin, 2006). Previous research of the role of new media in the understanding of social learning theory is extremely important, since the likelihood of remembering and assuming a particular rule or model of behavior depends on the strength and complexity of the internalized pattern, which is mainly conditioned by the attractiveness of the observed content (Livazović, 2009: 110). Zgrabljić Rotar (2005: 8) states that children are not only influenced by what is received through the media, but also by the passivity of the environment in which children grow up, that is, the passive family and school, which results in their personal passivity and the absence of emotional and intellectual readiness for life with the media. Ilišin (2001: 176) considers that the media are becoming the main element of socialization, therefore they surpass school, family life, etc. and thus influence the overall formation of values and behaviors of young people. It is possible to distinguish the risk and protective factors, as well as the short- and long-term factors of the influence of new media on the development children and young people’s skills. Alexander and Hanson (2003) find that they threaten to form a lasting system of beliefs, attitudes and habits as a daily pattern of living and behavior in social interactions.

Methodology

Objectives and hypotheses

The aim of the conducted research on the basis of this paper was to explain the impact of new media on social skills of elementary school students, with or without disabilities, aged 7-14 years in the Republic of Croatia. At the same time, the objective was to determine the time and content dimension of media exposure and use, as well as which devices students prefer to use for media content. Also, the objective was to analyze and determine if there was a difference in the perception of social skills development between a teacher and a parent of an individual student. In accordance with the objectives, hypotheses were stated:

H1: Time spent on new media has a significant impact on the development of social skills in children with and without disabilities.

H2: There is a correlation between time spent using new media and parental control.
H3: There is no difference in the perception of social skills development between a teacher and a parent.

Sample

The sample of research participants consists of parents and teachers of 200 students of primary schools and 46 students of education centers from six counties in the Republic of Croatia. Children with disabilities included in the study have been diagnosed with autism spectrum disorder and attend special education programs. For each individual student, the parent and teacher answered the same questions on the survey questionnaire and their answers were processed and compared afterwards.

Parents (N = 266) and teachers (N = 266) answered questions about the habits of using new media of their child/student and assessed the basic social skills of their child/student on variables adapted from Rajić (2011). Only face-to-face interaction was taken into account when assessing the child/student's social skills while family and educational setting were chosen because they are the agents of primary socialization. The parents completed the questionnaire at the parent meeting while the teachers filled it out at school, in the class, and since they could not estimate themselves how much time each student spent using new media, they asked the students themselves and based on that provided an answer. We hypothesized that parents of underage students could fairly accurately state how much time their child spends using new media.

All grades of primary school are included, with age 10, being the average age of students (M = 10.58). The majority of students are 13 years old, 61 (22.9%), one student is 15 years old (0.4%) and one is 18 years old (0.4%) as they are students of Education Centre for children with disabilities who are being educated until they are 21 years old. A total number of teachers (N=266) and parents (N=266) of 220 students (82.7%) without disabilities and 46 (17.3%) students with disabilities participated in the survey where 123 were girls (50.4%), and 132 boys (49.6%). The difference between the sizes of the two groups of participants stems from the lower percentage of students with disabilities in general population.

Results

Media time exposure and the use of devices

Survey results related to the time and content dimension of media exposure and use showed that, on average, respondents spend 2 hours (M = 2.01) daily with new media. The answers show that most children 53.7% spend 1 to 2 hours a day with new media, 25.2% spend less than 1 hour, while 19.9% spend 3 hours or more using media. To determine whether there was a statistically significant difference in media use between children with and without developmental disabilities, a nonparametric Mann-Whitney test was selected because previous Q-Q plots analysis in the SPSS Statistical Data Processing Program determined that the data were not normally distributed. It was found that there was no statistically significant difference in time spent on new media between students with or without disabilities (Table 2). Also, both groups of respondents spend, on average, between one and two hours of watching television and playing video games. As can be seen from Table 1, The Mann-Whitney shows no statistically significant difference between the groups was found (p> 0.05).

Table 2. Total time spent with media daily

Variable	Mean rank		Summ of ranks		Mann Whitney	p
	Regular development	Disabilities	Regular development	Disabilities		
Total time spent with media daily	133,42	133,87	29353,00	6158,00	5043,00	.969
Total time spent watching TV daily	132,57	137,96	29165,00	6346,00	4855,00	.632
Total time spent playing videogames	136,53	118,99	30037,50	5473,50	4395,50	.110

Note: *p < .05, **p < .01, ***p < .001; N=266 (Regular development=220, Disabilities=46)

The number of hours spent playing video games increases with the students' age, and 13-year-olds, are the age group included in the research that spends the most time playing videogames, then using the highest number of technological devices, they also watch television the most and spend the highest total number of hours with new media. The largest number of students, 103 (38.72%), prefer using cell phones and laptops, the number of devices used by students is increasing according to the age of the students. Students with disabilities prefer to use a cellphone, while students of a regular development a cellphone and a laptop. In addition to the devices offered in the survey: a cellphone, a computer, Wii and Xbox, 4 parents indicated that children were using Play Station 4 while 1 parent indicated that the child was also using Play Station 3.

Media content control

It can be seen that there is no difference between the parents of a child with disabilities (M = 1.154, SD = 0.51792) and regular development (M = 1.1739, SD = 0.38322) in controlling media content. As many as 88.35% of the parents of the respondents control the content that the children use. Although a smaller percentage of parents do not control media content, students without control spend more time with new media, television and video games (Table 3).

Table 3. Media content control

Variable	Mean rank		Summ of ranks		Mann Whitney	p
	Control	No control	Control	No control		
Total time	126,09	174,97	29379,00	5074,00	2118,00	.000***

Note: *p < .05, **p < .01, ***p < .001; N=266 (Media content control=235, The lack of media content control=31)

The assessment of social skills by parents and teachers

A Mann-Whitney nonparametric test was conducted to compare the assessment of the social skills of students with regular development and disabilities by parents, because the analysis found that the data in the variables did not correspond to the normal distribution. The differences were not determined on the following variables: the child waves and greets other people, the child accepts that he did not win the game. Significant differences were found in the variables: the child politely says no, makes eye contact when talking to other people, and has a difficult time making new friends.

Children of regular development stand up for themselves more commonly as they say no to others more often or ask someone to stop doing something that bothers them than the children with disabilities. Also, children of regular development make better eye contact while talking to others, and make new friends more easily.

Table 4. The assessment of social skills by parents

Variable	Mean rank		Summ of ranks		Mann Whitney	p
	Regular development	Disabilities	Regular development	Disabilities		
Student greets	135,47	124,07	29804,00	5707,00	4626,00	,333
Student says no	143,15	87,37	31492,00	4019,00	2938,00	,000***
Student makes eye contact with people	144,48	80,97	31786,50	3724,50	2643,50	,000***
Student makes friends	123,99	187,97	27278,50	8232,50	2968,50	,000***
Students accepts not winning in a game	133,73	132,42	29419,50	6091,50	5010,50	,913

Note: *p < .05, **p < .01, ***p < .001; N=266 (Regular development=220, Disabilities=46)

A nonparametric Mann-Whitney test was conducted to compare the assessment of students' social skills of regular development and with disabilities by teachers. No differences were found on the following variable: the child accepts that he did not win the game. Significant differences were found in the variables: the student politely says no, makes eye contact when talking, and has difficulty making new friends. A slightly smaller but statistically significant difference is also evident in the variable child waving to others and greeting them.

Children of regular development stand up for themselves more commonly as they say no to others more often or ask someone to stop doing something that bothers them than the children with disabilities. Also, children with regular development make better eye contact while talking to others, and make new friends more easily as it was shown by the parents' assessment as well.

Table 5. The assessment of social skills by teachers

Variable	Mean rank		Summ of ranks		Mann Whitney	p
	Regular development	Disabilities	Regular development	Disabilities		
Student greets other people	137,63	113,73	30279,50	5231,50	4150,00	,048*
Student says no	145,55	75,85	32022,00	3489,00	2408,00	,000***
Student makes eye contact with people	142,32	91,34	31309,50	4201,50	3120,50	,000***
Student makes friends	128,24	158,64	28213,50	7297,50	3903,50	,000***
Students accepts not winning in a game	130,79	146,47	28773,50	6737,50	4463,50	,190

Note: *p < .05, **p < .01, ***p < .001; N=266 (Regular development=220, Disabilities=46)

There are almost no differences in assessments between parents and teachers in the field of social skills. The only difference is seen in the waving, greeting, and healing variables, where parents rated their children more positively than did the teachers.

The correlation between hours spent with new media and social skills

To determine whether there is a correlation between total time spent with new media and parents' assessment of social skills, we used Spearman's correlation coefficient. The results showed that, according to parents, there is a significant correlation between hours spent on new media and social skills (Table 6). There is a statistically significant association with hours spent on new media and students making eye contact with other people. Students who spend more than 1 hour with new media rarely make eye contact when talking, and find it difficult to make new friends.

Table 6. The correlation between total time spent with new media and parents' assessment of variables

Variable	Student greets other people	Student says no	Student makes eye contact with people	Student makes friends	Students accepts not winning in a game
Student greets other people	-	,003**	.001***	,389	,747
Student says no	,003**	-	,000***	,013*	.306
Student makes eye contact with people	,001***	,000***	-	,000***	,084
Student makes friends	,393	,014*	,000***	-	,301
Students accepts not winning in a game	,725	,338	,106	,284	

Note: *p < .05, **p < .01, ***p < .001

Table 7. The correlation between total time spent with new media and teachers' assessment on variables

Variable	Student greets other people	Student says no	Student makes eye contact with people	Student makes friends	Students accepts not winning in a game
Student greets other people	-	,980	.190	,061	,613
Student says no	,980	-	,000***	,115	.000***
Student makes eye	,190	,000***	-	,000***	,001***

contact with people					
Student makes friends	,061	,115	,000***	-	,002**
Students accepts not winning in a game	,613	,000***	,001***	,002**	

Note: *p < .05, **p < .01, ***p < .001

Spearman's correlation coefficient was used to determine whether there is a correlation between the variables: the total time spent with new media and the teachers' assessment of students' social skills (Table 7). The results show that, according to the teachers' assessment, there is a significant correlation between the hours spent on new media and the adoption of social skills. There is a statistically significant correlation between hours spent with new media as it making eye contact with other people is more difficult for students, they also tend not to stick up for themselves and do not say no to others when they disagree. Teachers also believe that students who spend more time using new media find it more difficult to make friends and become more involved in playing with other students.

Discussion

Changes in the media landscape are caused not only by the development of new media, but also by the transformation of traditional media (Valkenburg, Taylor Piotrowski, 2017: 2). Our research shows that, on average, students watch television for 1 to 2 hours a day, as much time as they spend on average playing video games daily mostly on smartphones or tablets because of touch sensitive technology that affects the way children are engaged in a play. Playing games is a very important source of learning social patterns and skills in elementary school, such as: acceptance in the environment, waiting for a turn, making friendships, conflict resolution skills (Katz, McClellan, 1999: 19), it is a very important factor in building social skills in children. The research conducted confirmed the hypothesis (H1) that time spent on new media significantly influences the adoption of social skills in children of regular development and children with disabilities. The results of the study show that students who spend more time using new media are less likely to express social patterns such as making eye contact with others, indicating that video games cannot replace the game through personal contact that encourages communication. Considering the role of the media, Ilishin (2003: 15) points out that children who are less satisfied with life and less socially adjusted are more likely to use the media, therefore the media can be understood as a compensation for real life deficiencies. The media do their real task only when they encourage learning, developing talent, good behavior, and emphasizing positive values in children (Rožić, 2015: 153).

Media education is crucial in preventing negative media influences. The research preceding this paper shows that students whose parents do not control media content spend more time with the media, thus confirming the hypothesis (H2) that there is a correlation between time spent using new media and parental control. In cases where parents do not understand the use of their child's media content, it is possible for children to adopt undesirable behaviors. Today, the media is an extremely important socializing agent, and a particular danger can be caused by the inability to identify inadequate and manipulative content in the media. According to the study by Bickham and Rich (2006: 388) on the impact of exposure and television viewing on social isolation, children who watch television with peers, regardless of time spent, are found to be engaged in more social interactions with peers in the social environment. The American Academy of Pediatrics (2009) also proposes parents' involvement to spend time with their children while using media, and comment on the content of media with children to avoid negative media influences.

Learning non-verbal communication is only possible through face-to-face interaction, through which social skills that are crucial in adulthood (Bosacki, Astington, 1999: 242). The results of the study show that the assessments between parents and teachers in the field of students' social skills are almost identical, thus confirming the hypothesis (H3) that there is no difference in the perception of the adoption of social skills between teachers and parents.

Conclusion

The impact of new media on the development of social skills in children is inevitable. The organization of children's free time spent with their peers is increasingly being replaced by the use of media and technology. Consequently, for the quality media use, it is extremely important parents and teachers to assist and control the content that children use and to be actively involved in spending time with children and new media. It can be stated that the media is a potentially important socializing agent, but that their influence depends on a number of factors, such as: media selection, time of use, content selection, conditions of use and subgroup characteristics. The students with autism spectrum disorders, included in this study, dominantly exhibit discrepancies in the field of communication and social skills, so this research shows that special attention should be paid precisely to the use of media and to involve children with different types of disabilities in the future research.

Since parents and teachers are equally well acquainted with students' social skills, fostering their development of is one of the basic tasks of parents as well as contemporary educational institutions. In developing future curricula, the media aspect, especially the development of media literacy, needs to be closely addressed. Observing today's children, it is indisputable that they spend more and more time using new media, and it would be relevant, given the impact of these media on their social skills, to consider whether new media also affect the dimension of their emotional development.

References

Alexander, A., Hanson, J. (2003). Taking sides-mass media and society, McGrawHill/Dushkin, Connecticut

American Academy of Pediatrics (2009). Council on Communications and Media. Policy Statement-Media Violence. Pediatrics, 124, 1495-1503

Bickham D. S., Rich M. (2006). Is television viewing associated with social isolation? Roles of exposure time, viewing context, and violent content. // Arch Pediatr Adolesc Med 160, Apr, 4, 387-392

Bosacki, S., Astington, J. W. (1999). Theory of mind in preadolescence: Relations between social understanding and social competence. // Social Development 8, 237-255

Hoza, B. (2007). Peer functioning in children with ADHD. // Journal of Pediatric Psychology 32, 6, 655-663

Ilišin, V. (2003). Media for Leisure Time of Children and Youth, Media Research 2, 9-34

Ilišin, V., Bobinac Marinović, A., Radin, F.(2001). Children and the Media: The Role of The Media in Everyday Life of Children. DZOMM/IDIZ

Katz, L. G., McClellan, D. E. (1999). Encouraging the Development of Children's Social Competence, Zagreb, Educa

Klarin, M. (2006). The Development of children in a Social Context. Jastrebarsko: Naklada Slap

Lievrouw, L. A. and Livingstone, S. (2006). (eds.), Handbook of new media: social shaping and social consequences-fully revised student edition. London, UK : SAGE Publications, 1-14

Lister, M., Dovey, J., Giddings, S., Grant, I. and Kelly, K. (2009). New Media: a critical introduction. Routledge 270 Madison Ave, New York, NY 10016

Livazović, G. (2009). Theoretical and Methodological Features of Media Influence on Adolescents. // Life and School 21, 108-115

Manovich, L. (2002). New Media from Borges to HTML. The New Media Reader, edited by Noah Wardrip-Fruin and Nick Montforl, The MIT Press

Miljković, D., Rijavec, M. (2002). Better be the Wind than a Leaf. Zagreb: IEP.

Raboteg-Šarić, Z. (1997). Socialization od Childern and Youth. // Društvena istraživanja: časopis za opća društvena pitanja 6, 4-5, 30-31

Rajić, A. (2011.). Social skills. Obiteljski centar Primorsko-goranske županije, Rijeka. http://oc-pgz.hr/clanci_socijalne_vjestine.html (Accessed 21.10.2019.)

Rožić, I. The Influence of the Media on the Value System of Young People in Split and Padua. http://e-lib.efst.hr/2012/2102773.pdf (4.7.2019)

Valkenburg, P. M., Taylor Piotrowski, J. (2017) Plugged // How Media Attract and Affect Youth. New Haven: Yale University Press

Zgrabljić Rotar, N. (2005). The Media-Media Literacy, Media Content and Media influence; Media Literacy and Civil Society, Sarajevo, Media Centar

https://doi.org/10.17234/INFUTURE.2019.25

Original scientific paper

The Usage of Social Media for Higher Education Purposes

Tihana Babić
Algebra University College, Zagreb, Croatia
tihana.babic@algebra.hr

Gordana Vilović
Faculty of Political Science, University of Zagreb, Croatia
gordana.vilovic@fpzg.hr

Ljubica Bakić Tomić
University of Applied Sciences Baltazar Zaprešić, Croatia
ljbakictomic@bak.hr

Summary

Social media today are an inevitable part of everyday life; networking of individuals and new ways of communication and interaction have emerged in the academic world. They are available 24 hours a day, 7 days a week across a range of devices and from different locations, thus depending only on the availability of the internet and the will of their users. By enabling the creation and sharing of information, ideas, and interests through virtual communities and networks, whose purpose is to turn communication into interactive dialogue, they have influenced society, economics, politics, science, and education. Students of today are generations who have started virtual life on social media since their early age and do not know the world where there are no computers, cell phones, and social media. Social media is an example of technology students have widely adopted, but the digital gap between students and their educational institutions is evident. Numerous researches show that there is a need for integration of social media into the teaching process because students experience them as self-explanatory. There are many advantages and positive effects on the teaching process, but teachers are less likely to use them because they are more focused on their shortcomings and dangers. The use of technology can be influenced by numerous factors, and the current study is conducted to identify the current status of teachers' usage of social media in higher education and detection of factors which influence the usage of social media in higher education. The questionnaire is based on the Unified Theory of Acceptance and Use of Technology named UTAUT model as a theoretical framework, and data is collected on a sample of teachers at the Algebra University College in Zagreb and the University of Applied Sciences Baltazar Zaprešić. This research is the first research conducted among the higher education teachers in the Republic of Croatia, concerned not only with the social media usage for the higher education purposes but also with factors that influence the willingness of higher education teachers to use social media for the higher education purposes.

Key words: social media, higher education, teachers, communication, UTAUT Model

Introduction

Social media sites (SMS) are based on Web 2.0 technology, enabling the creation and sharing of information, ideas, and interests across virtual communities and networks that aim to turn communication into interactive dialogue. There are 13 subtypes: blogs, microblogs, business, and social networking tools, collaborative projects, forums, photo sharing tools, business collaboration tools, product and service reviews, research networks, social games, photo/video sharing tools, and virtual worlds (Aichner, Jacob, 2015). They are available 24 hours a day, 7 days a week, across a range of devices and from different locations. Therefore, they depend only on the availability of the internet and the will of their users. They have influenced society, economy, politics, science, and education.

Students of today are members of a generation that, inherently, have grown up on social media. Moreover, they do not know a world where computers, cell phones, and social media do not exist. Social media is an example of technology that has been widely adopted by students and, consequently, has the potential to become a valuable resource for supporting educational communication and student collaboration with the faculty (Arsović, 2012). Nevertheless, numerous studies point to the digital gap between students and their educational institutions, as well as the trend towards non-adaptation of new technologies in higher education institutions; students are willing to use them, and faculty employees are not (Roblyer et al., 2010).

There is no universal point of view and "the jury hasn't reached a verdict yet" about the use of social media for higher education purposes. Although their influence can be seen as something that enriches and makes our lives easier, we may have a more pessimistic view, but what we certainly cannot do is ignore the changes that have taken place.

Previous researches

Social media sites (SNS) and their users are a frequent topic of research, both in Croatia and worldwide. Statistics show that in January 2018, there were 48% of active social media users signing up for social media services (Statista, 2018a). The power of social networking is so great that the number of active social media users worldwide will grow from the current 2.62 billion to 3.02 billion by 2021, which is about a third of the total Earth population (Statista, 2018b).

Research on the impact of social media is being conducted in all areas of human activity, and the academic environment is no exception. Research into computer-mediated communication and the possibilities of social media application in higher education are most often concerned with students' activities on social media (Vivian et al., 2014; Echeng et al., 2016), confirming that they represent a new generation of students who are gaining knowledge in new ways (Selwyn, 2011). Moreover, research goes to prove that students ‘values often conflict with the traditional values of higher education institutions (Ulbrich et al., 2011).

Social media tools and applications are challenging the concept of formal education as we know it today (Selwyn, 2011), and the role of teachers is changing. Research conducted on the attitudes and experiences of social media integration into higher education institutions (Echeng et al. 2016; Josefsson, 2017) shows that there is a need to integrate social media into the teaching process (Okello-Obura, 2015)- In addition, research is showing the numerous benefits and positive effects on the learning and teaching process(Alsolamy et al., 2017; Coleman et al., 2018) such as teaching quality (Silvestru et al., 2016) and students ‘academic achievement (Tamayo et al., 2014). While students perceive social media as something self-explanatory, teachers are more inclined to reflect on the shortcomings and dangers (Raut et al., 2016; Willems et al., 2018).

Unified theory of acceptance and use of technology (UTAUT model)

The of usage of technology, in general, can be influenced by a number of factors, so in 2003 Venkatesh et al. developed and empirically validated a Unified Technology Acceptance and Use Theory (UTAUT) based on a review of 8 models of earlier theories and the consolidation of constructs. According to the UTAUT model (Venkatesh et al., 2003), behavioral intention (intention to use in the next 12-24 months) and (actual) behavior i.e. use of technology are distinguished. It is assumed that behavioral intention is significantly influenced by expected work performance, expected effort, and social impact, while behavioral intention and facilitating conditions significantly contribute to the actual use of the system. In addition to the predictors of use, four moderator variables of key relationships were included in the model: age, gender, the experience of use, and voluntary use.

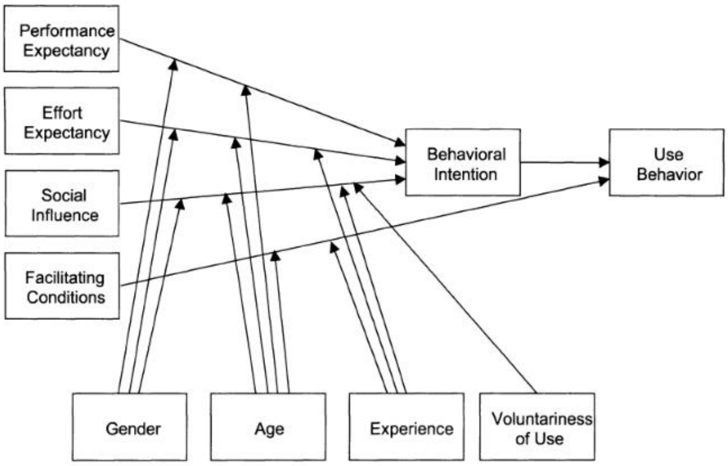


Figure 1. UTAUT Research Model (Venkatesh et al. 2003; 447)

The UTAUT model has different fields of application and has thus served as a theoretical basis for scientists in some countries to implement social computing in higher education institutions (Hussein, 2013) and to identify factors that influence students' use of social media platforms for academic purposes (Ali et al., 2017).

The research goals

The aim of the research was to determine and address whether teachers use social media for higher education purposes, and whether there is an influence of age, gender, scientific-teaching area and/or associate profession, teacher attitudes toward usage, social influence and anxiety when using social media to the actual teaching of usage of social media for higher education purposes.

The research questions are:

1. Do teachers use social media for higher education purposes?
2. Do demographic characteristics such as age, gender, and field of science of teaching and/or associate profession of teachers influence their use of social media for higher education purposes?
3. Do teachers' attitudes, social impact, and anxiety when using social media influence their usage of social media for higher education purposes?

The research sample

A sample of this research was pertinent. There were 73 participants which have scientific-teaching and associate titles and are employed as teachers or external teaching assistants at the Algebra University College Zagreb (52 % of respondents) and the Baltazar Zaprešić University of Applied Sciences (48 % of respondents).

Research methodology

To examine teachers' perceptions and usage of social media in higher education, a structured questionnaire was designed according to a tailored UTAUT model (Venkatesh et al., 2003). The model is adjusted in such a way that, with the following independent variables: age, gender, anxiety, and social impact on the social media usage for higher education, the conceptual model is expanded by 2 more variables: attitude towards the social media usage for higher education purposes and the scientific field of scientific-teaching or associate titles.

The survey was conducted during the summer semester of the 2018/2019 academic year and was conducted using a specially designed questionnaire in the Google Forms tool, which was distributed to teachers via e-mail. Respondents participation was voluntary and anonymous. The questionnaire consisted of two wholes of grouped closed-ended questions about general demographics and social media usage for higher education. To ensure a clear understanding of the term social media before the question, a descriptive definition of the term social media is specified in the questionnaire.

Results of an empirical study with discussion

1. Do teachers use social media for higher education purposes?

The results of this study on teachers' social media usage for higher education purposes are presented concerning the frequency of use and subtypes of social media used by teachers.

1.1. Frequency of teachers' social media usage for higher education purposes

Table 1 shows that almost half of the respondents, 47.9% of them, use social media for higher education purposes daily, but less than an hour a day, while the smallest percentage (9.6%) do so for several hours a day. 32.8% of respondents do this in the range of several hours per week to several hours per month, and 13.7% of respondents do not use social media at all.

Table 1. Distribution of the answer to the question (N = 73): How often do you use social media for higher education purposes? Source: questionnaire and author's analysis.

Using social media for higher education purposes	**Percentage of respondents in the sample**
A few hours a day	9.6 %
Less than an hour a day	47.9 %
A few hours a week	16.4 %
A few hours a month	16.4 %
I don't use social media	13.7 %

1.2. Social media subtypes used by teachers for higher education purposes

Respondents indicated that they mostly use social networks such as Facebook, Google+ (49.3%), social media for video sharing such as YouTube, Vimeo (46.6%) and business networks such as LinkedIn (43.8%) for higher education purposes, while least of them use social games such as World of Warcraft or Mafia Wars (2.7%) or social media for bookmarking such as Pinterest and Reddit (12.3%). 13.7% of survey participants do not use any of the social media subtypes for higher education purposes.

2. Impact of age, gender and scientific field of scientific-teaching and/or associate profession of teachers on their use of social media for higher education

In the second part of the analysis, demographic differences in the total sample (N = 73) were examined. The demographic statistics of the sample are shown in Table 2.

Table 2. Distribution of respondents by age, gender and scientific field of scientific-teaching and/or associate profession (N = 73). Source: questionnaire and authors' analysis.

Variable	**Category**	**Percentage**
Gender	Men	56.2 %
	Women	43.8 %
Age	Less than 20 years	0 %
	21 - 30 years	8.2 %
	31 – 40 years	39.7 %
	41 – 50 years	30.1 %
	51 – 60 years	13.7 %
	61 years and over	8.2 %
Scientific area	Natural Sciences	13.7 %
	Technical Sciences	24.7%
	Biomedicine and Healthcare	0 %
	Biotehncical Sciences	0 %
	Social Sciences	54.8 %
	Human Sciences	11 %
	Art field	4.1 %
	Interdisciplinary fields	15.1 %

To answer the second research question, the following hypothesis is defined:
H1: There are statistically significant differences in the acceptance of the social media usage for the higher education purposes by students and teachers concerning the following indicators: age, gender and scientific area of scientific-teaching and/or associate profession.
The data were analyzed by chi-square test (Table 3) and showed that the variables age (p = 0.341), gender (p = 0.405) and scientific area (p = 0.117) were not statistically significant for teachers' social media usage for higher education purposes and therefore hypothesis H1 is rejected.

Table 3. Indicators of age, gender and scientific area of scientific-teaching and/or associate profession; Pearson Chi-Square Test, Source: Questionnaire, SPSS Author Processing

Indicator	**Value**	**df**	**Asymp. Sig. (2-sided)**
Teacher gender	4,008a	4	,405
Teacher age	17,717 a	16	,341
Scientific area of scientific-teaching and/or associate profession of teacher	27,705a	20	,117

3. The impact of teachers 'attitudes, social influence, and anxiety when using social media for higher education purposes

3.1. Teachers' attitudes and social media usage for higher education purposes
To determine whether teachers' attitudes influence their actual social media usage for higher education purposes, the main hypothesis is defined:
H2: The social media usage for higher education purposes is significantly influenced by the teacher's attitude towards social media usage for higher education.
The grouping variable “Attitude of teachers towards the social media usage” had 4 sub-variables, auxiliary hypotheses were defined, and the data were analyzed by chi-square test (Table 4):
H2.1. The social media usage for higher education purposes is significantly influenced by the ATTITUDES of teachers: "Teachers should use social media more actively to teach students."
The stated attitude is not statistically significant (p = 0,120), auxiliary hypothesis H2.1. is discarded.
H2.2. The social media usage for higher education purposes is significantly influenced by the ATTITUDE OF THE TEACHER: Institutions of higher education should adopt a policy of using social media for study purposes.
The stated attitude was not statistically significant (p = 0.143), auxiliary hypothesis H2.2. is discarded.
H2.3. The social media usage higher education purposes are significantly influenced by the ATTITUDES OF THE TEACHER: Institutions of higher education should implement teacher education on social media usage.
The stated attitude was statistically significant (p = 0.003), auxiliary hypothesis H2.3. is accepted.
H2.4. The social media usage for higher education purposes is significantly influenced by the ATTITUDE OF THE TEACHERS: Higher education institutions should educate students on social media usage.
The stated attitude is statistically significant (p = 0.002), thus supporting hypothesis H2.4. is accepted.

Table 4. Attitude toward social media usage for higher education purposes; Pearson Chi-Square Test, Source: Questionnaire, SPSS Author analysis

Attitude toward social media usage for higher education purposes	**Value**	**Df**	**Asymp. Sig. (2-sided)**
Teachers should use social media more actively to teach students.	22,782^{a}	16	,120
Higher education institutions should adopt a policy of using social media for study purposes.	22,004^{a}	16	,143
Colleges should spend on educating teachers on social media usage.	35,728^{a}	16	,003

Colleges should spend on educating students on social media usage.	36,903[a]	16	,002

It can be concluded that there is a statistically significant connection between the teachers' attitude that higher education institutions should educate both teachers and students on the social media usage and the actual teaching of social media usage for the higher education purposes.

3.2. Anxiety when using social media and teaching of social media usage for higher education purposes

To determine whether anxiety when using social media influences teachers' social media usage for higher education purposes, the following main hypothesis is defined:

H3: Teachers' social media usage for higher education needs is significantly influenced by the construct from the adapted UTAUT model: ANXIETY when using social media.

The grouping variable "Anxiety when using social media" had 4 sub-variables, auxiliary hypotheses defined, and a chi-squared test used for testing (Table 5):

H3.1. Teachers' social media usage for higher education is significantly influenced by: "Feeling concerned about the social media usage for higher education."

The sub variable is not statistically significant (p = 0.424), auxiliary hypothesis H3.1. is discarded.

H3.2. Teachers' social media usage for higher education is significantly influenced by: "Feeling scared that he/she might lose a lot of information when using social media for higher education if he/she does something wrong."

The sub variable is not statistically significant (p = 0.400), auxiliary hypothesis H3.2. is discarded.

H3.3. Teachers' social media usage for higher education is significantly influenced by: "The hesitation in using social media for higher education from fear of errors that cannot be corrected."

The sub-variable is not statistically significant (p = 0.634), auxiliary hypothesis H3.3. rejects.

H3.4. Teachers' social media usage for higher education is significantly influenced by: "The feeling that using social media for higher education is a little scary."

The sub-variable is statistically significant (p = 0.033), auxiliary hypothesis H3.4. accept.

Table 5. Anxiety while using social media for higher education purposes; results Pearson Chi-Square Test, Source: questionnaire, SPSS author analysis

Anxiety while using social media for higher education purposes	**Value**	**Df**	**Asymp. Sig. (2-sided)**
Feeling concerned about using social media for higher education purposes.	16,427[a]	16	,424
Feeling scared that she could / could lose a lot of information when using social media for higher education purposes if she did something wrong.	16,775[a]	16	,400
The hesitation of using social media for higher education purposes for fear of errors that cannot be corrected.	13,532[a]	16	,634
The feeling that using social media for higher education needs is a little daunting.	27,849a	16	,033

Based on the results, it can be concluded that there is a statistically significant connection between the teachers' sense that the social media usage for higher education is a bit scary and the actual teacher's social media usage for higher education purposes.

3.3. Social influence and teachers' social media usage for higher education purposes

To determine connection between social influence and teachers' social media usage for higher education purposes, the following main hypothesis is defined:

H4: The teachers' social media usage for higher education purposes is significantly influenced by the construct from the adapted UTAUT model: SOCIAL IMPACT.

The grouping variable “SOCIAL IMPACT” had 4 sub-variables, auxiliary hypotheses were defined, and a chi-squared test was used to test the auxiliary hypotheses (Table 6):

H4.1. Teachers' social media usage for higher education is significantly influenced by: " Colleague teachers using social media for higher education."
The stated social influence was not statistically significant (p = 0.127), auxiliary hypothesis H4.1. is discarded.
H4.2. Teachers' social media usage for higher education is significantly influenced by "Students who feel that teachers should use social media for higher education."
The stated social influence was statistically significant (p = 0.035), auxiliary hypothesis H4.2. is accepted.
H4.3. Teachers' social media usage for higher education is significantly influenced by: "Friends and close acquaintances of teachers who use social media for higher education."
The indicated social influence was not statistically significant (p = 0.348), ancillary hypothesis H4.3. is discarded.
H4.4. Teachers' social media usage for higher education is significantly influenced by "Faculty management that supports the use of social media for higher education purposes."
The indicated social influence is statistically significant (p = 0.022), auxiliary hypothesis H4.4. is accepted.

Table 6. Social influence on using social media for higher education purposes; Pearson Chi-Square Test, Source: questionnaire, SPSS author analysis

Social influence on using social media for higher education purposes	**Value**	**Df**	**Asymp. Sig. (2-sided)**
College teachers using social media for higher education purposes.	22,539[a]	16	,127
Students who feel that teachers should use social media for higher education purposes.	27,584[a]	16	,035
Friends and close acquaintances of teachers using social media for higher education purposes.	17,599[a]	16	,348
Faculty management that supports the use of social media for higher education purposes.	29,300[a]	16	,022

It can be concluded that there is a statistically significant connection between the influence of students who believe that teachers should use social media for higher education and the influence of the faculty administration that supports the social media usage for higher education on the actual teaching social media usage for higher education.
Finally, it can be pointed out that this research has shown that gender, age and the scientific field of teaching are not significantly affected by the actual social media usage of teachers, but it is certainly not neglected that there are statistically significant relationships between particular attitudes of teachers towards the social media usage. Social influences and variations in anxiety when using social media on actual teacher social media usage, as summarized in Table 7.

Table 7. Summary of Findings; Source: questionnaire, SPSS author analysis

Dependent Variable	**Independent Variables**	**Explanation**
SMS usage for higher education purposes	Gender	There is no statistically significant connection
SMS usage for higher education purposes	Age	There is no statistically significant connection
SMS usage for higher education purposes	Academic Field	There is no statistically significant connection
SMS usage for higher education purposes	Attitude toward using SNS for higher education purposes	There is no statistically significant connection between: Teachers should use social media more actively to teach students. Higher education institutions should adopt a policy of using social media for study purposes. There is a statistically significant connection between: Colleges should spend on educating teachers on social media usage. Colleges should spend on educating students on social media

		usage.
SMS usage for higher education purposes	Social Influence	There is no statistically significant connection between: College teachers using social media for higher education purposes. Friends and close acquaintances of teachers using social media for higher education purposes. There is a statistically significant connection between: Students who feel that teachers should use social media for higher education purposes. Faculty management that supports the use of social media for higher education purposes.
SMS usage for higher education purposes	Computer Anxiety when using SNS for higher education purposes	There is no statistically significant connection between: Feeling concerned about using social media for higher education purposes. Feeling scared that she could / could lose a lot of information when using social media for higher education purposes if she did something wrong. The hesitation of using social media for higher education purposes for fear of errors that cannot be corrected. There is a statistically significant connection between: The feeling that using social media for higher education purposes is a little daunting.

Conclusion

This paper examines whether teachers' demographic characteristics, attitudes toward usage, social influences, and anxiety when using social media influence their actual social media usage for higher education purposes.

The results show that the age, gender and scientific field of the teaching profession do not affect their actual social media usage for higher education purposes, as is the case with particular attitudes towards use, social influences, and anxiety when using. However, it is not negligible that at the same time higher education institutions would have a significant and positive impact on their actual social media usage through the implementation of education for teachers and students on the social media usage, which could also reduce their sense that the social media usage for the needs of high education is little scary. For teachers' usage of social media for higher education purposes, significantly, students believe teachers should use social media for higher education purposes. Moreover, for them, support from faculty administration regarding social media usage for higher education is also crucial.

The research conducted has certain shortcomings as well that could be eliminated in future research. This research was conducted at two higher education institutions and thus the results of the research do not necessarily apply to other higher education institutions. Also, the sample was relatively small. The study included a limited number of relevant variables, and more accurate results could be obtained by examining the extent to which a single variable is statistically significant and to what extent it affects the behavior of teachers regarding the actual social media usage for higher education needs. Future research could extend the analysis to other variables such as the extent to which perceived strengths and perceived disadvantages affect the actual social media usage. The results can also be compared between private and public higher education institutions.

References

Ali, M., Yaacob, R.; Ednut, B., Makki, B. (2017). Determining the academic Use of Social Media with technology Acceptance Models. Pakistan: NFC Institute of Engineering and Fertilizer Research

Alsolamy, F. (2017). Social networking in higher education: academics' attitudes, uses, motivations, and concerns. Doctoral, Sheffield Hallam University

Arsović, B. (2012). Društvene mreže – izazovi mogućnosti za obrazovanje. // Tehnika i informatika u obrazovanju: 4. Internacionalna konferencija, Čačak: Tehnički fakultet Čačak

Coleman, B.; Petitt, S.; Buning, M. (2018). Social Media Use in Higher Education: Do Members of the Academy Recognize Any Advantages? // The Journal of Social Media in Society 7, 1, 420-442

Echeng, R., Usoro, A., Ewuzie, I. (2016). Factors to Consider when Enhancing the Use of Web 2.0 Technologies in Higher Education: Students' and Lectures' Views for Quality Use. // International Journal of Digital Society 7, 1

Josefsson, P. (2017). Higher education meets private use of social media technologies. Doctoral, KTH Royal Institute of Technology

Mabić, M. (2014). Društvene mreže u obrazovanju: Što misle studenti Sveučilišta u Mostaru, Opatija: XXI. Međunarodni znanstveni skup Društvo i tehnologija – Dr. Juraj Plenković

Nikolić, G. (2013). Cjeloživotno učenje, potrebne promjene u obrazovanju odraslih. Opatija: predavanje na IV. Susretu ustanova za obrazovanje odraslih, Arhiva Zajednica ustanova za obrazovanje odraslih

Nikolić, G. (2014). Nove tehnologije donose promjene. // Andragoški glasnik 18, 2, 25-43

Okello-Obura, C. (2015). Web 2.0 technologies application in teaching and learning. // Library Philosophy and Practice 1248

Raut, V., Patil, P. (2016). Use of Social Media in Education: Positive and Negative impact on the students. // International Journal on Recent and Innovation Trends in Computing and Communication 4, 1, 281-285

Roblyer, M. D., McDaniel, M., Webb, M., Herman, J., Witty, J. V. (2010). Findings on Facebook in higher education: A comparison of college faculty and student uses and perceptions of social network sites. // Internet and Higher Education 13, 134-140

Selwyn, N. (2011). Social media in higher education. Education and Technology Continuum, London - New York.

Silvestru, C. I., Lupescu, M. E., Draistaru, A. S. (2016). The Impact of Using Social Media in Adult Education, Studies from Education and Society

Statista (2018a). https://www.statista.com/statistics/295660/active-social-media-penetration-in-european-countries/. (21.10.2018)

Statista (2018b). https://www.statista.com/topics/1164/social-networks/. (21.10.2018)

Sugimoto, C. R., Work, S., Lariviere, V., Haustein, S. (2016). Scholarly use of social media and altmetrics: a review of the literature. // Journal of the association for Information Science

Tamayo, J. D., Cruz, G. S. G. (2014). The Relationship of Social Media with the Academic Performance of Bachelor of Science in Information Technology Students of Centro Escolar University – Malolos. // International Journal of Scientific and Research Publications 4, 5, 1-9

Ulbrich, F., Jahnke, I. Martensson, P. (2011). Special Issue on knowledge development and the net generation. // International Journal of Sociotechnology and Knowledge Development

Venkatesh, V., Morris, M. G., Davis, G. B., Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly 27, 3, 425-478

Vivian, R., Barnes, A., Geer, R., Wood, D. (2014). The academic journey of university students on Facebook: an analysis of informal academic-related activity over a semester // Association for learning technology: Open Access Journal

Willems, J., Adachi, C., Bussey, F., Doherty, I., Hujiser, H. (2018). Debating the use of social media in higher education in Australasia: Where are we now? // Australasian Journal of Educational Technology 34, 5

https://doi.org/10.17234/INFUTURE.2019.26

Review paper

Software Visualization in Education

Vedran Juričić
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
vedran.juricic@ffzg.hr

Summary

Software visualization technology provides better understanding and more efficient creation and use of computer software by graphically representing its components, functionality or algorithms. There are various uses and benefits of software visualization, from detecting logical bugs, errors and performance bottlenecks to their use in simulations and e-learning. This paper focuses on algorithm visualization tools, approaches and languages, which show various states, transitions and data structures in more abstract and clearer way than algorithms presented with traditional programming code. It also shows the importance of visualization in learning algorithms, those that are taught at the very beginning of programming courses, as well as advanced cryptographic algorithms. Visualization tools differ in the level of engagement they provide to their users: some of them provide only simplest interactivity, while other provide questions and quizzes, changing input data or constructing custom visualizations. The paper analyses modern and most popular tools and approaches in algorithm visualization and compares their characteristics, advantages and disadvantages in order to show the most important characteristics of today's visualization tools and approaches, and their suitability for different problem areas and scenarios. The paper also analyzes the impact of software visualization on students' knowledge, motivation and efficiency.

Key words: software visualization, algorithm visualization, tools, education, analysis

Introduction

Software visualization encompasses multiple terms like program visualization, algorithm visualization and visual programming. They are commonly used in the literature and although they are similar or related, should not be used as synonyms (Price, Baecker, Small, 1993). A formal definition of program visualization states that it is a mapping from programs to graphical representations (Gruia-Catalin, Cox, 1993). Its goal is to simplify an explanation of a certain program segment, program structure or flow control and requires three participants: the programmer who develops a program, the animator who constructs the mapping and the viewer who observes the representation. On the other hand, visual programming is programming that uses multiple dimensions for transferring and presenting semantics (Burnet, 2001). Typical examples include a time dimension that defines before and after relationships or multidimensional objects like diagrams, icons or sketches that provide additional meaning. A programming language whose syntax consists of visual expressions is called a visual programming language and an environment where a code is written is called a visual programming environment (Ingaiis et al., 2008). Therefore, program visualization and visual programming are two entirely different concepts. In visual programming, the program is written by using graphics, while in program visualization is specified in a regular or textual form and the graphics is only used for clarification and description (Myers, 1986).

Algorithm visualization is a process of visualizing a high-level description of a piece of software (Price, Baecker, Small, 1993). It simplifies the understanding and analysis of programs and its elements by showing various states, transitions and data structures in more abstract and cleaner way than algorithms presented with traditional programming code. This paper analyses modern and most popular or influential tools and approaches in algorithm visualization and compares their characteristics, advantages and disadvantages. Also, it analyses their potential usage and impact in educating students with and without prior programming knowledge.

Visualization in education

The information and communication technologies bring new methods and opportunities for more attractive and productive teaching and learning. They are used to create and study learning materials, track students' progress, write online quizzes; they increase accessibility of learning materials, provide virtual learning environments and enable a development of student's individual work and qualities by individualization of tasks, where each student can be assigned different problem difficulty or different amount of time needed for its solution (Majhertová, Palásthy, Gunčaga, 2014; Noor, 2013). ICTs have potential to accelerate and deepen students' skills, to motivate and stimulate students, and to make their learning more interesting and engaging.

An important benefit of using ICT in teaching mathematics and informatics is a possibility of visualization, animation and simulation. From student's perspective, visualization provides a new approach for discovering the structure and component dependencies, to solve problems and to discover and analyse results and data, enabling acquisition of new information and knowledge. From teacher's perspective, it provides a new teaching style that complements standard verbal presentation, enabling more dynamics and reducing a time required for explaining a complex problem (Majhertová, Palásthy, Gunčaga, 2014).

A core challenge in teaching programming is helping a student to reason or infer about execution of program code and not to think about the syntax itself. Students should learn how a static textual representation or programming code maps to a dynamic process or program execution (Guo, 2013; Sorva, 2012). Learning programming language is the minor problem in computer courses, where beginners learn about variables, branches and loops. The main problem is how to develop student's thinking and appropriate approaches, and how to use his knowledge of programming language for solving higher level problems. A language, i.e. language syntax and rules, can be learnt in a couple of days, but it can take years to become a good programmer.

Robins et al. (2003) show that the average student does not make much progress in an introductory programming course. Most teachers confess that large number of graduates are beyond standard of programming competence and are not able to program. Guzdial (2011) discovered that only 14% of Yale's students that completed introductory programming course were able to solve the Rainfall problem, a simple problem with reading input numbers and calculating their average. Every study that has used this problem has found similar low performance (Guzdial, 2011). Another study shows that most students of introductory programming courses either fail miserably or pass with extremely high results, with only few that pass with average results (Dehnadi, Bornat, 2006).

The purpose of program visualization is to enhance students' understanding of program behaviour on level higher than understanding programming code. The analysis of three surveys (Naps et al., 2002) based on more than 140000 respondents who were involved in visualization supported programming course showed that 90% find teaching experience more enjoyable and more than 76% of the students showed improved level of participation and motivation. More importantly, for 72% of the students was shown that their learning was improved. Another study was conducted on a control group of 94 students and a test group of 78 student that were using visualization tool. It showed that students achieved better results with visualization, that is they were able to predict program behaviours more accurately.

Although the results of this studies show positive impact of visualization on teaching and learning process, some authors argue that a visualization does not have a significant educational value if it does not engage its users, i.e. students, in an active learning activity (Naps et al., 2002). It can be achieved by writing their own input data sets, predicting future visualization states, programming an algorithm, answering questions or constructing their own visualizations (Hundhausen, Douglas, 2000). Byrne, Catrambone and Stasko (1999) observed that pure visualization and animation does not improve students' knowledge and that there exists no statistically significant difference between animation and PowerPoint presentations, but they confirmed the improvement when student were able to enter their own data into the experiment or program. In other words, in order for a visualization technology to be successfully applied in education, it should not refer to passive images or observations, but must be interactive and enable student to experiment with states and input data.

Visualization tools

Program visualization is studied for many years now and there have been developed numerous tools and platforms that support the learning of programming languages and programming in general. This chapter describes the most popular tools according to the size of their community but also tools that are unique or have special impact on this technology.

UUhistle (Uuhistle.org, 2019) is a program visualization system made for visualizing the execution of small, single-threaded programs in introductory programming courses. It has been designed to support active learning and encourages interaction between a student and visualization. UUhistle can be used in numerous use cases (Sorva and Sirkiä, 2011): animated debugging of programs that students wrote themselves, making presentations as a help when analysing program execution, visual program simulation, interactive mode and program animation, quizzes that can be embedded in example programs etc. System is a subject of many scientific papers and even doctoral thesis because of its amazing interaction and positive effect on students. It is free, written in Java for Python programming language and available for offline usage and as an applet. Unfortunately, it is no longer developed and its latest version was launched in 2013. An example of tool's appearance when visualizing a method invocation is shown in Figure 1.

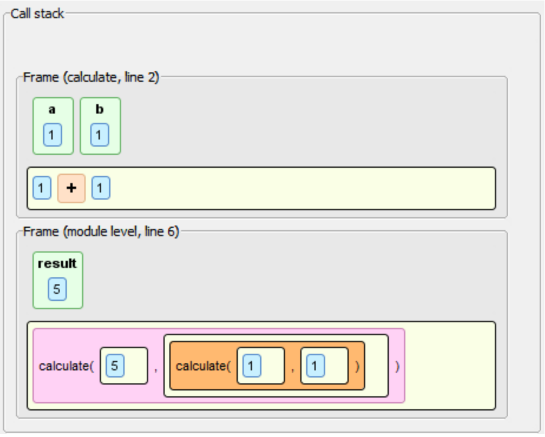
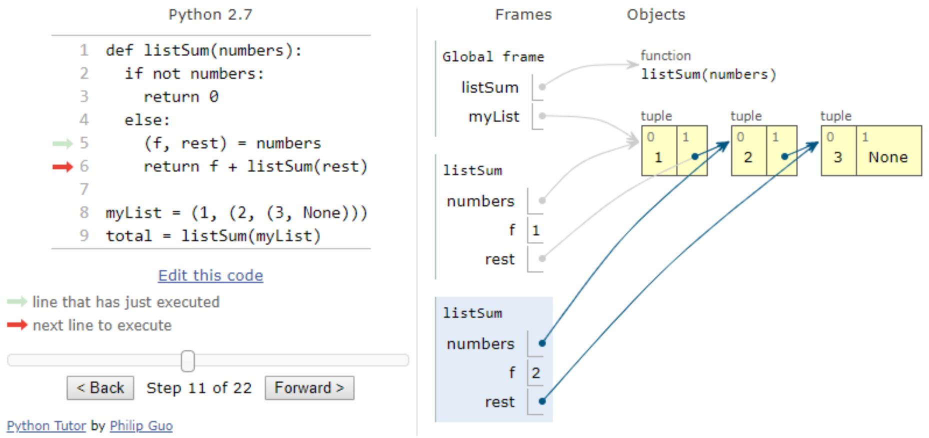

Call stack
Frame (calculate, line 2)
a
1
b
1
1 + 1
Frame (module level, line 6)
result
5
calculate(5 , calculate(1 , 1))

Figure 1. UUhistle showing simple method invocation, source: Uuhistle.org, 2019

Jsvee & Kelmu (Sirkiä, 2018) are two successors of UUhisle system. Jsvee is a JavaScript library that helps teachers to visualize notional machines and create expression-level animations, while Kelmu is a toolkit that enables writing more advanced explanations. They are not developed for a certain language and can be customized and extended. There are numerous online examples for Java and Python programming languages. They are being actively developed and their latest version was launched in August 2019.

Ville (Cs.utu.fi, 2009) is also a program visualization tool that supports multiple programming languages. It comes with syntax and example editor, but teachers can also write, define and add their own languages, which enables its maximum extensibility. Students can edit code, view call stack and variables, set breakpoints and have complete control over visualization and animation. It also integrates quizzes so that students can test their knowledge of observed problem. It is free, written in Java and available as desktop version and as an applet.

Python tutor (Pythontutor.com, 2013) is an advanced program visualization tool that supports Python, Java, C, C++, JavaScript and Ruby programming languages. It has a large user community, including students and instructors, with over five million people in 180 countries. It provides interactive environment where students can experiment with program execution, including live programming mode, where students can analyse recursions and complex programming problems. It stands out with live help module, supported by many volunteers around the world, that allows anyone to join user's session and help him debug program by writing code or explaining a problem using integrated chat. Python tutor is free and accessible through a web application. Figure 2 shows an example of recursion visualization.



Python 2.7

```
def listSum(numbers):
  if not numbers:
    return 0
  else:
    (f, rest) = numbers
    return f + listSum(rest)

myList = (1, (2, (3, None)))
total = listSum(myList)
```

Figure 2. Python tutor visualizing a recursion, source: Pythontutor.com, 2013

BlueJ (Kouznetsova, 2007) is a tool that differs from above described ones because it is not designed as an aid for learning basics, but for object-oriented programming, which is more difficult for students than procedural programming. It is integrated development environment for Java programming language, that comes with textbook, teacher support and detailed documentation. The tool is open-source, under GNU General Public licence and available on Windows, Linux and MacOS operating systems. It offers an overview of object activities and their dependencies and provides an interaction with objects and methods, including their invocation with custom parameters. An example of object and its methods is shown in Figure 3.

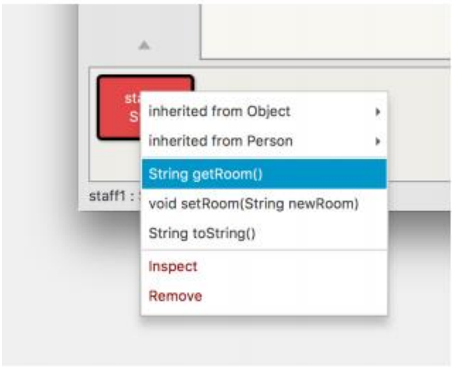


Figure 3. BlueJ displaying possible method invocations

Tools that were described in this paper are only a small subset of available ones. There exists over 50 program visualization tools that naturally differ in their features and possible usages. Most of them share on important characteristic, interactivity, which is in accordance with results of previously conducted studies that showed its positive effect on students' satisfaction and knowledge. They have the option to perform step-by-step analysis of target program execution and experiment with input data and current memory state, like variable values on stack and heap. Most of them are free, written by developer community and with partially available source code. They commonly support two programming languages, Python and Java. Currently they are the most represented languages in education, Python in introductory programming courses and Java in object-oriented programming.

Tools differ in their access and run methods. Some of them are available only as a standalone application for Windows, MacOS or Linux, and some of them through web browser, which gives them features like possibility to save examples in the cloud, to share them with community and to request help and tutoring. The most obvious difference is a result of their purpose; some of them are created as an aid with basic programming, some of them for algorithm visualization and some of them show dependencies on a higher level, like objects and their interconnections.

Conclusion

This paper describes software visualization taxonomy and shows advantages and suggestions for using program visualization tools in education. It is shown that using visualization technology is not sufficient in order to improve students' knowledge; it only improves their satisfaction and interest for programming. The real progress occurs when visualization is broadened with interactivity, including experimenting with input data, variables and program flow. Paper describes some of the available visualization tools that are often used in teaching programming, in order to show their common features and characteristics.

References

Burnett, M. (2001). Visual Programming. Wiley Encyclopedia of Electrical and Electronics Engineering

Byrne, M. D., Catrambone, R. C., Stasko, J. T. (1999). Evaluating animations as student aids in learning computer algorithms. // Computers & Education 33, 4, 253-278

Cs.utu.fi. (2009). VILLE - the visual learning tool. Available at: https://ville.cs.utu.fi/old/?p=2 (10.9.2019)

Dehnadi, S. and Bornat, R. (2006). The camel has two humps (working title). Middlesex University, UK, 1-21

Gruia-Catalin, Cox, K. (1993). A Taxonomy of Program Visualization Systems. // Computer 26, 12, 11-24

Guo, P. J. (2013). Online python tutor: embeddable web-based program visualization for cs education. // Proceeding of the 44th ACM technical symposium on Computer science education, 579-584

Guzdial, M. (2011). From Science to Engineering – Exploring the Dual Nature of Computing Education Research. Communications of the ACM 54, 2, 37-39

Hundhausen, C. D., Douglas, S. (2000). Using Visualizations to Learn Algorithms: Should Students Construct Their Own, or View an Expert's? IEEE Symposium on Visual Languages, Los Alamitos, California, 21-28

Ingaiis, D., Wallace, S., Chow, Y.-Y., Ludolph, F., Doyle, K. (2008). Fabrik A Visual Programming Environment. // ACM SIGPLAN Notices 23, 11

Kouznetsova, S. (2007). Using BlueJ and Blackjack to teach object-oriented design concepts in CS1. // Journal of Computing Sciences in Colleges 22, 4, 49-55

Majhertová, J., Palásthy, H., Gunčaga, J. (2014). Educational Software and Visualization in Teaching. Future Learning

Myers, B. A. (1986). Visual programming, programming by example, and program visualization: a taxonomy. // ACM SIGCHI Bulletin 17, 4, 59-66

Naps, T. L., Rößling, G., Almstrum, V., Dann, W., Fleischer, R., Hundhausen, C. and Velázquez-Iturbide, J. Á. (2002). Exploring the role of visualization and engagement in computer science education. // ACM Sigcse Bulletin 35, 2, 131-152

Noor, S. U. (2013). An Effective use of ICT for Education and Learning by Drawing on Worldwide Knowledge, Research, and Experience: ICT as a Change Agent for Education. // Scholarly Journal of Education 2, 4, 38-45

Online Python Tutor (2013). Embeddable Web-Based Program Visualization for CS Education. Philip J. Guo. ACM Technical Symposium on Computer Science Education (SIGCSE)

Price, B. A., Baecker, R. M., Small, I. S. (1993). A Principled Taxonomy of Software Visualization. // Journal of Visual Languages & Computing 4, 3, 211-266

Pythontutor.com. (2013). Python Tutor - Visualize Python, Java, C, C++, JavaScript, TypeScript, and Ruby code execution. Available at: http://pythontutor.com/ (11.9.2019)

Robins, A., Rountree, J. and Rountree, N. (2003). Learning and Teaching Programming: A Review and Discussion. // Computer Science Education 13, 2, 137-172

Sirkiä, T. (2018). Jsvee & Kelmu: Creating and tailoring program animations for computing education. // Journal of Software: Evolution and Process, 30, 2, e1924

Sorva, J. (2012). Visual program simulation in introductory programming education. Aalto University

Sorva, J. and Sirkiä, T. (2011). Context-sensitive guidance in the UUhistle program visualization system. // Proceedings of the 6th Program Visualization Workshop (PVW'11), 77-85

Uuhistle.org. (2019). UUhistle.org. Available at: http://www.uuhistle.org (10.9.2019)

https://doi.org/10.17234/INFUTURE.2019.27

Original scientific paper

How to Measure Digital Literacy?
A Case of Croatian Adult Learners

Dario Pavić
Department for Croatian Studies, University of Zagreb, Croatia
dpavic@hrstud.hr

Iva Černja
Department for Croatian Studies, University of Zagreb, Croatia
icernja@hrstud.hr

Summary

This study aims to present the rationale, the instrument and the main findings of the assessment of digital literacy of adult learners in Croatia, as a part of the project Implementation of EU Agenda for Adult Learning 2017-2019, by Ministry of Science and Education. The concept of digital literacy is at the heart of EU development policies and encompasses skills, competences, and dispositions for using a wide range of technologies, from personal computers to smartphones, vending machines, and ATMs. The digital technologies are also being used for different purposes (work, leisure, communication, health, news, etc.) and differently by social groups (by age, gender socioeconomic status, etc.). Thus, testing digital literacy in the population is of great importance given the pervasiveness of digital technologies and the apparent inadequacy of self-reported measures of digital literacy. An original test was developed, testing the usage of Windows, Word, Excel, PowerPoint, internet browser and Google Maps. The test was implemented through the LimeSurvey software with the additional files available on the participant's computer desktop. The questions were devised in a way that a participant had to solve a specific problem using the specific software, thus including the cognitive component of digital literacy. The difficulty level of the questions is approximate of those of the levels "Below Level 1" and "Level 1" of PSTRE PIAAC research. Additionally, the participants answered the question about their socio-economic status, the frequency of technology usage, and their self-perception of digital skills. The test was administered to 92 attendees of the specialization and qualification programs in seven adult learning facilities throughout Croatia (Knin, Karlovac, Zagreb, Split, Čakovec, Koprivnica, and Virovitica). The results indicate relatively low levels of digital literacy of the tested sample and the different patterns of technology usage by different socio-economic groups. Also, the questionnaire's metric characteristics show that it can be successfully used for testing digital literacy outside the laboratory setting. Conclusion: The low levels of digital literacy among the adult attendees of the specialization and qualification programs implicate the need to include the digital literacy curriculum in their learning programs, according to individual needs and experiences. Also, this research is one of the very few that did not assess only the self-perceptions of the skills or test the skills in the research facility setting, but rather in the learning facilities of adult education. Since the questionnaire performed well under these circumstances, it can be modified to be implemented in multiple environments.

Key words: digital competence test, digital skills, adult education, Croatia

Introduction

Digital skills and digital literacy, in general, have been in the center of both academic interest and national and international policies, yet the reliable measures of these skills are scarce having in mind how digital technology permeates the everyday life of modern people. Part of the reason why this is so is the multitude of definitions and conceptualizations of what digital skills, digital competence, and digital literacy are. Presenting a thorough review of the development of these concepts lies outside of the scope of this paper, yet some historical and conceptual points are in place to show why measuring digital skills has been a difficult endeavor.

Historically, Martin and Grudziecki (2006) point that computer, IT or ICT literacy has been identified as a need as early as the late 1960s, and divide the evolution of these concepts in three phases; from the emphasis on technical knowledge on how computers operate in the earliest phase, to the critical and reflective evaluation of information technologies in the post-1990s period (Martin, Grudziecki, 2006). Also, several related concepts used during the formative years of digital technologies (e.g. media literacy, ICT literacy, computer literacy, visual literacy) coalesced in one nowadays ubiquitous concept of "digital literacy", especially from the 1990s (Chinien, Boutin, 2011), popularized by Paul Glister's seminal book "Digital literacy" (Gilster, 1997). In his definition, Gilster emphasizes the cognitive and life-related aspects of digital literacy as "[the] ability to understand and use information in multiple formats from a wide range of sources […]" (Gilster, 1997: 1-2).

Over time, the other definitions of digital literacy have also embraced the cognitive aspect of literacy as a leading one and have conceived digital literacy as a multidimensional concept. Regardless of whether the definition is of digital literacy, digital competence or digital skill, most of them involve technical operation, but also the information management, collaboration, communication and sharing, creation of content and knowledge, ethics and responsibility, and evaluation and problem solving (Ferrari, 2012). These definitions also include the use of different digital tools and technologies, different domains of learning, and different modalities and goals of technology use (Ferrari, 2012). It must be pointed out that the terms "literacy", "competence" and "skills" are not synonyms. "Skills" are a specific and measurable application of knowledge to attain a goal, so in a digital area, it can be e.g. to open an attachment to the e-mail. "Competence" is a wider concept than "skill" (Rychen, Tiana, 2004, in Halász, Michel, 2011), and it combines knowledge (European Parliament and the Council of the European Union, 2006), skills and attitudes. "Digital literacy" is an over-reaching concept different from the "competence" by being situationally embedded (Martin, 2008). For an extensive review of the definitions of these concepts see (Chinien, Boutin, 2011).

Most of the research on the assessment of digital literacy has been focused on people's self-perception of the skills (Hargittai, 2005), especially until the mid-2000s. Although Hargittai (2005) posits that self-reported measures under certain circumstances can be used as indicators of people's real digital literacy, other studies present a series of problems regarding self-reported measures. There is a concern that people with poor ICT skills overestimate their actual skills (Danish Technology institute, n.d., in Chinien, Boutin, 2011). This overestimation of actual skills has also been shown on a small sample of US university students (Merritt, Smith, Renzo, 2005), and university students in Denmark, Finland, Germany, India, and Singapore (ECDL Foundation, 2018), dispelling the myth of young "digital natives". The same patterns of overestimated self-reported skills were found in Austria and Switzerland on a sample of participants aged 15 to 65 (ECDL Foundation, 2018).

More direct measures of digital literacy have been conducted by testing the skills of participants in a real-life or simulated computer environment. Hargittai (2002) tested the information retrieval from the Internet to discover the inequalities in digital literacy between social groups, a so-called "second digital divide" (Hargittai, 2002). In a series of articles van Deursen and associates (van Deursen, van Dijk, Peters, 2011; van Deursen, van Dijk, 2014; van Deursen, van Dijk, 2008) presented the results of testing the internet skills of Dutch population, performed in university's computer laboratory. The assignments thematic issues covered governmental information, leisure-related information, and health-related ones. The most significant finding from these studies is that the construct of "Internet skills" is composed of five parts: operational, navigation information, social, creative, and mobile (A. van Deursen, Helsper, Eynon, 2016). Jara et al. (2015) used a simulated computer environment to test the students' digital skills on a theme of ecology. The students were able to use a word processor, e-mail, spreadsheet program, and internet browser to complete the tasks. Gui and Argentin (2011) combined theoretical multiple-choice questions with operational and evaluation skill tasks carried out on a computer to test the digital skills of Italian high school students. Apart from these studies, there are numerous digital literacy assessment frameworks carried out by national or international organizations and private corporations. The most notable of these frameworks are PIAAC PS-TR, iSkills, ECDL, SAILS (review of these and other frameworks in Ferrari, 2012 and (Sparks, Katz, Beile, 2016) and ICILS (Fraillon, et al., 2014) The emerging feature of digital literacy from these studies is that it comprises of dimensions of information definition, accessing, evaluating, managing, integrating and creating, as well as from communication, problem-solving, ethical issues and technology use (Sparks et al., 2016).

This article aims to show the usefulness of the test of digital skills made for adult learners and to present its main findings. Testing the digital skills of adult learners in Croatia is an interesting and relevant topic. Most of these learners are of lower socio-economic and educational status, females, unemployed for a long time, but willing to get a degree in vocational education. They represent a particularly vulnerable group since less educated people have lower levels of digital skills (Gui, 2007; Hargittai, 2002; van Deursen, van Dijk, 2011; van Deursen, van Dijk, 2008) and tend to use digital technology less for human capital enhancement, and more for leisure and entertainment (Hargittai, Hinnant, 2008). Furthermore, neither primary adult education in Croatia nor the adult vocational education does not include compulsory courses on digital literacy. This is expected to be changed by the newly proposed program of adult education which is still in the preparatory phase. The information gathered by this research would be invaluable to the policymakers working on the new adult education curriculum.

Data and methods

Our test of digital skills was devised specifically for the population of adult learners in Croatia. Although the contemporary notion of digital skills includes a wide variety of platforms and software, we opted for the most ubiquitous office suite of Microsoft (Word, Outlook, Excel and PowerPoint), Windows OS, Google Chrome or Mozilla Firefox web browsers and Google Maps service. The main reasons for the inclusion of these programs were their wide-spread use and their usefulness in both private and professional areas. The test itself consisted of eighteen questions, each testing a problem in one of the aforementioned programs. The difficulty of these problems was equivalent to levels “Below level 1” and “Level 1” of the PSTRE competencies of the PIAAC research (Kirsch, Yamamoto, Garber, 2013). Solving the problems on these levels involves only a small number of steps, minimal navigation between the pages and simple inference about the goal of the problem. A test folder with test files was prepared for each participant and available on each test computer. The questionnaire and the answer forms were implemented in the online survey software LimeSurvey on the server of the University of Zagreb Computer Centre (SRCE).

The problems were divided into four difficulty groups. The easier questions required little or no navigation by the participant. The participants were usually faced with a screenshot of a program file and asked to identify a key information on the screenshot (e.g. identify the number of words, font size or font type in Word), or they were asked to open a test folder and identify an information about the files in the folder (e.g. the size of the file, the number of .pdf files in the folder, etc.). The more difficult questions asked the participant to perform one or more tasks (e.g. to copy and paste some text from the Word to the answer form, to calculate auto sum in Excel, to find the address of a government body on the internet or to find the distance between two places in Google Maps). The easiest questions were marked with one point, all the way to the most difficult ones (4 points), reflecting the number of distinct steps needed for the solution to the problem. The maximum score on the test was forty. Also, the participants were surveyed on their socio-demographic characteristics, their experience in using digital technologies and self-perception of their digital skills.

The participants were included in the sample by convenience. The MZO provided the list of twenty-five active people’s universities in Croatia that perform the adult education programs. The institutions were contacted by e-mail and phone to arrange the testing date. The testing was performed in seven institutions throughout Croatia (cities Knin, Karlovac, Zagreb, Split, Čakovec, Koprivnica, and Virovitica). The reasons for the exclusion of other institutions were our inability to get their response, the lack of computer classrooms on the premises and the lack of active adult education programs. The number of participants per institution ranged from five to twenty, totaling 92 participants who started and finished the test. The different adult education programs that participants were enrolled in were caretaker, CNC machine operator, shoemaker, intermediate English and German, EU funds specialist, bookkeeper, and ECDL operator. The standard ethics protocol was followed. Each participant was informed about the goal of the survey, the anonymity measures and the right to terminate the participation in the survey at any time.

All the institutions had computer classrooms furnished with desktop computers with Windows 10 OS, Microsoft Office, and internet access. the person who conducted the testing responded only to technical questions and adjusted the accessibility features of the user’s interface. In the cases where the participant didn't possess even the basic digital skills, the survey assistant recorded the participant's answers on all the questions except those from the test itself, so the socio-demographic and other characteristics of this kind of participant remained recorded, but the test score was assigned to zero. The maximum amount of time for the whole questionnaire was 60 minutes which was not exceeded by any of the participants.

Results

Reviewing the results includes evaluating the structure and reliability of the digital skills test. To determine the level of digital skills, the results were analyzed for the overall test and individual questions. Some basic sociodemographic determinants of test performance were also examined.

The correlation matrix of 18 items of the test indicates the suitability of the matrix for the factor analysis. The value of the Kaiser Meyer Olkin (KMO) coefficient is 0.895, which is large enough to conclude the adequacy of the matrix for the analysis (Kaiser, 1970). Bartlett's sphericity test indicates that the extra diagonal elements of the correlation matrix are statistically significantly different from zero ($\chi2$ =96.502, df = 153, p <.001). Out of 153 correlations among the items, all were statistically significant (p <0.05). All item-total correlations were statistically significant (p <.05) and ranged from .55 to .77.

The method of dimensionality reduction was the analysis of principal components. The initial analysis extracted 4 components with characteristic root greater than 1 (characteristic roots were 8.54; 1.36; 1.15 and 1.05). The first principal component has much greater characteristic root than others. The results indicate a clear single-factor structure supported by the results of the Scree test (Figure 1.). We performed principal component analysis with an only one-factor solution. The extracted component explains 47.43 % of the total variance. All items on the corresponding components have saturations greater than .50. The component saturations on the first component are shown in Table 1.

Table 1. Results of analysis of principal components

Question	Component Saturation
q1	.677
q2	.547
q3	.741
q4	.768
q5	.777
q6	.679
q7	.649
q8	.542
q9	.762
q10	.732
q11	.578
q12	.776
q13	.687
q14	.668
q15	.631
q16	.702
q17	.694
q18	.719
Characteristic root	8.538
% of Variance	47.431

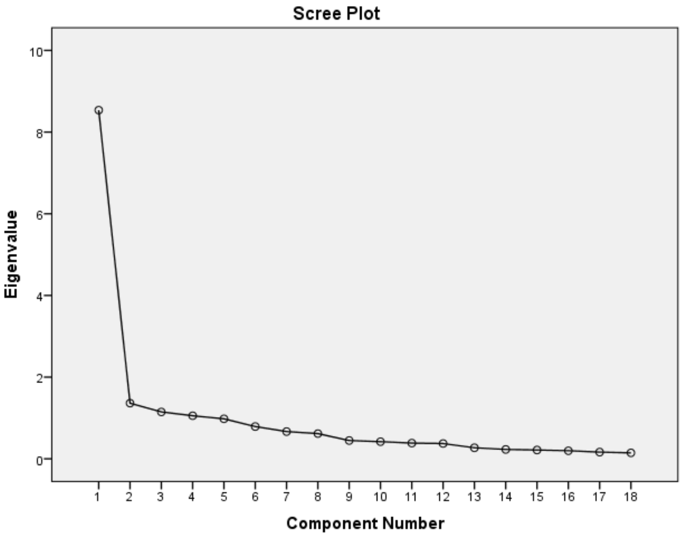


Figure 1. Scree plot

To determine the adequacy of the obtained structure, confirmatory factor analysis was performed on one factor. The results indicate that the one-factor model is well-suited to the data ($\chi 2$/df <3, RMSEA = .103, CFI = .877, SRMR = .080). Since the values fit indices are within the range of the acceptable values (Hu, Bentler, 1999; Tabachnick, Fidell, 2007) we conclude that confirmatory factor analysis confirmed the obtained structure.

Internal reliability of the test was very high (α = .93) which indicates good internal homogeneity of the test. We also performed split-half analysis of reliability with Spearman-Brown prediction formula, and all indicators suggest good reliability of the test (Table 2).

Table 2. Results of split-half analyses of reliability

Cronbach's Alpha	Part 1	Value	.882
		N of Items	9[a]
	Part 2	Value	.889
		N of Items	9[b]
	Total N of Items		18
Correlation Between Forms			.785
Spearman-Brown Coefficient			.879

a. The items are: q1, q2, q3, q4, q5, q6, q7, q8, q9.

b. The items are: q10, q11, q12, q13, q14, q15, q16, q17, q18.

The Digital Competency Test consisted of 18 questions, with the score being weighted depending on their difficulty and the number of actions that participants had to perform to find the correct answer. The point test range was from 0 to 40 (Table 3). A total of 11 respondents did not score a single point in the Digital Competence Test. These respondents are mostly female elementary school graduates, currently unemployed. They generally do not possess a personal computer in their household and state that they rarely use one.

The average score on the Digital Competence Test is in the middle of the theoretical range of scores on the test (M = 20.55), and we can conclude that on average, the participants have correctly solved about half of the test. The highest percentage of correct answers was on questions related to the use of e-mail technology (Table 4). The most incorrectly answered questions were those related to the usage of Excel. Also, the task of requiring the attendees to find the official web site of the Government of the Republic of Croatia was rarely answered correctly.

The results of the t test show that men and women do not differ significantly in their success on the Digital Competence Test (t=-1.39, df=88, p>.05). The result of the digital competence test is also not related to the age of the participants (p>0.05). There was a statistically significant moderate correlation between the level of education and the year of completion of the last degree of education with the result on the Test. The result in the digital competence test was higher for people with higher the level of education (r=.341, p<.01) and those who completed the education more recently (r=.296, p<.01)

Table 3. Descriptive statistics of results in The Digital Competence Test

Variable	N	Min	Max	M	SD	SE	C	Q1	Q3
Results in The Digital Competence Test	92	0	40	20.55	13.629	1.421	7	7	33

Table 4. Results on questions in The Digital Competence Test

Question in test	Score in test	incorrect	correct
		%	%
Win - document size	2	45,70 %	54,30 %
Win – type of file	2	63,00 %	37,00 %
Win – e-mail entry	2	27,20 %	72,80 %
Win – Outlook (responding to e-mail)	1	22,80 %	77,20 %
Word – font size	1	37,00 %	63,00 %
Word – word count	1	42,40 %	57,60 %
PowerP – number of slides	1	50,00 %	50,00 %
Internet – Govt. e-mail address	2	72,80 %	27,20 %
Internet – web page recognition	1	25,00 %	75,00 %
Win – File’s modification date	2	43,50 %	56,50 %
Win - calculator	3	46,70 %	53,30 %
Word – copy-paste	3	34,80 %	65,20 %
Excel – Locating Sheet1	3	48,90 %	51,10 %
Internet – mail address University of Zagreb	3	51,10 %	48,90 %
Excel – AutoSum	4	64,10 %	35,90 %
Internet – locating e-mail address	4	52,20 %	47,80 %
Internet – Google Maps - distance	3	48,90 %	51,10 %
Win/Int – Uploading the file	2	46,90 %	53,10 %

Table 5. shows the correlations of the digital competence test and the digital skills self-assessment. Correlations are moderate, the highest one being with self-reported Word skills. The overall self-assessment of digital skills is moderately correlated with the score on the digital competence test (.599). The two measures share 35.8% of the total variance, which proves that the digital skills measure yields a unique variance through the knowledge test.

Table 5. Correlation between Results in The Digital Competency Test and self-assessment of digital skills

	Results in The Digital Competency Test
Windows	.554**
Word	.630**
Excel	.511**
PowerPoint	.433**
Internet browser	.544**
Internet banking	.498**
Buying on the internet	.344**
General grade of computer knowledge	.599**

Discussion and conclusion

The one component structure of the Digital Competence Test suggests that the test measures general digital skill and can be used for its initial detection of digital competences level. The purpose of the test was not to verify or establish the structure of the construct of "digital literacy" or "digital skill", but rather to be a simple and reliable measure of general digital skill, measured through several most widely used computer technologies. In this respect, the test was successful given the special target population it was performed on. Apart from adult learners with low digital skills, the test could be used to detect the level of digital skills in other vulnerable groups (unemployed, minority groups, etc.)

The significant result of this research is that almost 10% of the test takers didn't possess even the most basic digital skills and could not use the personal computer at all. As stated before, this group consists mainly of low educated women. This is hardly surprising since in Croatia there are 48% of males with basic or above basic digital skills, but only 34% of corresponding females (Eurostat, 2018). The situation is similar with respect to education. There are less than 20% of regular computer users with basic education, compared to around 80% of those with higher education (Državni zavod za statistiku Republike Hrvatske, 2018). None of the programs the respondents involve any type of digital skills instruction, which leaves these individuals without even basic digital skills even though it will certify them with the vocational education diploma. This situation is expected to be amended with the new curriculum for adult education in Croatia where digital skills will be treated as transversal skills.

The problems that have been unanswered the most (type of files, Excel, government web page) reflect the interests of the participants and their most frequent use of digital technologies. The participants use the technologies mostly for job hunting, information and leisure (data not shown). These activities rarely include technologies such as Excel, the technicalities like file size, type and location, and internet pages outside of their interest. The patterns of how respondents use digital technologies and consequently display digital skills can be explained by the differences between "lifestyle" and "workplace" skills. The skills adopted through the use in the lifestyle domain do not automatically transfer to the productivity or workplace domain (ECDL Foundation, 2018).

The observed difference in digital skills between the individuals of different levels of education is expected, given the experience in technology use the higher educated individuals gain through education and more complex jobs. It also might reflect the phenomenon of the "second digital divide" (Hargittai, 2002). While the original notion of "digital divide" emphasized the difference in access to digital technology between the social groups ("haves" and "have-nots"), the "second digital divide" posits that the difference is no more in the access to technology, but in the skills and motivation to use it (Hargittai, 2002). The technology like computers, smartphones, tablets, etc. have become accessible even to the poor and low educated individuals, but not the skills needed to embrace the full potential of participating in a digital society. On the methodological side, this result points to the convergent validity of the test since the difference in digital skills between different education groups are expected both from the theory and the past research.

The significant correlations between the self-assessment of the skills and their measured levels support Hargittai's assertion that self-assessment can be used as an indicator of real digital skills (Hargittai, 2005), yet the Digital Competence Test brings the unique variance that cannot be explained by the self-assessment alone. The answer to the question of why the self-assessment is positively correlated with the measured skills is outside the scope of this article and should be investigated separately. However, the uniqueness of the studied sample points to the possible answer. The subjects tested in this study are a part of a selected group of those who decided to educate themselves further or to change their profession altogether, to have better prospects of finding a job both in Croatia or abroad. Personal characteristics of the subjects may account for the correlation between self-assessment and real score. Those who have low digital skills are aware of it and do not exaggerate their self-assessment, as much as are aware of those with higher digital skills. It is possible that both groups high motivation and the dedication to learning prevent them from making exaggerated or false self-assessment claims.

As noted above, this research is relevant for testing the digital skills of adult learners and other subpopulations. Given the problems were conceived as real-life tasks of problem-solving through digital technology, the test can be expanded with other programs and more complicated problems tailored for the specific groups and their life experiences. It should be noted, however, that the results of this research have their shortcomings. The sample is not a probabilistic one and the number of participants is rather limited. The studied group is a specific one, unrepresentative of the general population, and some aspects of digital literacy (e.g. ethical aspect) have been omitted on purpose. The test should be applied to more groups to ensure its reliability and validity. The last point is also the direction for further research. If proven reliable and valid on other social groups, the test could be widely used as a simple and easy-to-implement initial screening tool for digital skills.

References

Chinien, C., Boutin, F. (2011). Defining Essential Digital Skills in the Canadian Workplace: Final Report. Montreal

Državni zavod za statistiku Republike Hrvatske. (2018). Statistički ljetopis Republike Hrvatske 2018 [Statistical Yearbook of the Republic of Croatia 2018]. Zagreb

ECDL Foundation. (2018). Perception & Reality—Measuring Digital Skills Gaps in Europe, India and Singapore. http://ecdl.org/media/perception__reality_report_-_ecdl_foundation_-_2018_1.pdf (12.8.2019)

European Parliament and the Council of the European Union. (2006). Recommendation of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning. Official Journal of the European Union, L 494, 10-12

Eurostat. (2018). Eurostat—Tables, Graphs and Maps Interface (TGM) table. https://ec.europa.eu/eurostat/tgm/refreshTableAction.do?tab=table&plugin=1&pcode=tepsr_sp410&language=en (12.8.2019)

Ferrari, A. (2012). Digital Competence in practice: An analysis of frameworks. https://ictlogy.net/bibliography/reports/projects.php?idp=2263 (12.8.2019)

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., Gebhardt, E. (2014). Preparing for Life in a Digital Age: The IEA International Computer and Information Literacy Study International Report

Gilster, P. (1997). Digital Literacy. New York: John Wiley and sons

Gui, M. (2007). Formal and substantial Internet information skills: The role of socio-demographic differences on the possession of different components of digital literacy. // First Monday 12, 9

Gui, M., Argentin, G. (2011). Digital skills of internet natives: Different forms of digital literacy in a random sample of northern Italian high school students. // New Media & Society 13, 6, 963-980.

Halász, G., Michel, A. (2011). Key Competences in Europe: Interpretation, policy. // European Journal of Education 46, 3, 289-306

Hargittai, E. (2002). Second-Level Digital Divide: Differences in People's Online Skills. // First Monday 7, 4

Hargittai, E. (2005). Survey Measures of Web-Oriented Digital Literacy. // Social Science Computer Review, 23, 3, 371-379

Hargittai, E., Hinnant, A. (2008). Digital Inequality: Differences in Young Adults' Use of the Internet. // Communication Research 35, 5, 602-621

Hu, L., Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. // Structural Equation Modeling: A Multidisciplinary Journal 6, 1, 1-55

Jara, I., Claro, M., Hinostroza, J. E., San Martín, E., Rodríguez, P., Cabello, T., Ibieta, A., Labbé, C. (2015). Understanding factors related to Chilean students' digital skills: A mixed methods analysis. // Computers & Education 88, 387-398

Kaiser, H. F. (1970). A second-generation little jiffy. // Psychometrika, 35, 4, 401-415

Kirsch, I., Yamamoto, K., Garber, D. (2013). Chapter 1: PIAAC Assessment Design. // Technical Report of the Survey of Adult Skills (PIAAC). Paris: OECD, 1-42

Martin, A. (2008). Digital Literacy and the "Digital Society." // Digital Literacies: Concepts, Policies and Practices / Lankshear C; Knobel, M; (eds.) New York: Peter Lang Publishing, 151-176

Martin, A., Grudziecki, J. (2006). DigEuLit: Concepts and Tools for Digital Literacy Development. // Innovation in Teaching and Learning in Information and Computer Sciences 5, 4, 249–267

Merritt, K., Smith, K. D., Renzo, J. C. D. (2005). An Investigation of Self-Reported Computer Literacy: Is it Reliable? // Issues in Information Systems 7, 289-295

Sparks, J. R., Katz, I. R., Beile, P. M. (2016). Assessing Digital Information Literacy in Higher Education: A Review of Existing Frameworks and Assessments with Recommendations for Next-Generation Assessment: Assessing Digital Information Literacy in Higher Education. ETS Research Report Series 2, 1-33

Tabachnick, B. G., Fidell, L. S. (2005). Using multivariate statistics. 5th ed. Boston: Pearson/Allyn & Bacon

van Deursen, A. J. A. M., Helsper, E. J., Eynon, R. (2016). Development and validation of the Internet Skills Scale (ISS). // Information, Communication & Society, 19, 6, 804-823

van Deursen, A. J. A. M., van Dijk, J. A. G. M.; Peters, O. (2011). Rethinking Internet skills: The contribution of gender, age, education, Internet experience, and hours online to medium- and content-related Internet skills. // Poetics 39, 2, 125-144

van Deursen, A. J. A. M., van Dijk, J. A. G. M. (2014). The digital divide shifts to differences in usage. // New Media & Society, 16, 3, 507-526

van Deursen, A. J. A. M., van Dijk, J. A. G. M. (2008). Using Online Public Services: A Measurement of Citizens' Operational, Formal, Information and Strategic Skills. // Electronic Government / Wimmer, M. A., Scholl, H. J., Ferro, E. (eds.) Berlin: Springer, 195-206

Reviewers

All papers were reviewed by at least two reviewers. INFuture relies on the double-blind peer review process in which the identity of both reviewers and authors as well as their institutions are respectfully concealed from both parties.

INFuture2019 gratefully acknowledges its reviewers:

Petra Bago, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Silvie Cinková, Faculty of Mathematics and Physics, Charles University, Czech Republic

Ivan Dunđer, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Sanjica Faletar Tanacković, Faculty of Humanities and Social Sciences, Josip Juraj Strossmayer University of Osijek

Goran Hajdin, Faculty of Organization and Informatics, University of Zagreb, Croatia

Ivana Hebrang Grgić, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Gordana Hržica, Faculty of Education and Rehabilitation Sciences, University of Zagreb, Croatia

Tomislav Ivanjko, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Vedran Juričić, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Shadrack Katuu, Information Management Officer, United Nations Mission in South Sudan

Sanja Kišiček, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Kristina Kocijan, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Shana Kopaczewski, Indiana State University, USA

Tomislava Lauc, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Vlatka Lemić, University Archives, University of Zagreb, Croatia

Nikola Ljubešić, Jožef Stefan Institute, Slovenia

Basma Makhlouf-Shabou, Haute école de gestion de Genève (HEG-Genève), University of Applied Sciences Western Switzerland (HES-SO), Switzerland

Dario Malnar, Croatian Military Academy *Dr. Franjo Tuđman*, University of Zagreb

Nives Mikelić Preradović, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Željka Miklošević, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Christine Möhrs, Leibniz-Institute for the German Language, Germany

Marko Odak, University of Mostar, Bosnia and Herzegovina

Ivana Ogrizek Biškupić, University of Applied Sciences Baltazar, Croatia

Dario Ogrizović, Faculty of Maritime Studies, University of Rijeka, Croatia

Krešimir Pavlina, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Ana Pongrac Pavlina, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Arian Rajh, Agency for Medicinal Products and Medical Devices, Croatia

Angela Repanovici, Transilvania University of Brașov, Romania

Alexis Sancho-Reinoso, Faculty of Social Sciences, University of Vienna, Austria

Sanja Seljan, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Dragan Soleša, Faculty of Economics and Engineering Management, University Business Academy, Serbia

Helena Stublić, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Mikko Tolonen, Department of Digital Humanities, University of Helsinki, Finland

Radovan Vladisavljević, Faculty of Economics and Engineering Management, University Business Academy, Serbia

Radovan Vrana, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Goran Zlodi, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia