

## Izrada i struktura CROVALLEX-a

Osnovna datoteka u kojoj su pohranjene informacije CROVALLEX-a je XML datoteka, koja je generirana iz Microsoft Access baze u kojoj su te informacije bile pohranjene u međusobno povezanim tablicama<sup>11</sup>.

CROVALLEX v 2.008 je verzija leksikona koja posjeduje čvrstu i eksplicitno iskazanu strukturu podataka kroz definirani odnos među elementima.

Leksikon se sastoji od **natuknica** (*word\_entry*) koje se moraju pojaviti barem jednom, ali mogu i više puta. Element *word\_entry* se sastoji od 3 elementa: *headword\_lemma*, *reflexivity\_example* te *frame\_entry*, od kojih *headword\_lemma* predstavlja glagolsku lemu, nalazi se na prvom mjestu i smije se pojaviti samo jednom unutar elementa *word\_entry*, element *reflexivity\_example* predstavlja primjer moguće refleksivne upotrebe glagola (nije nužan), dok se element *frame\_entry* mora pojaviti barem jednom, ali može i više puta unutar elementa *word\_entry* i predstavlja ono najvažnije – valencijski okvir glagola.

---

<sup>11</sup> XML ne definira kako će podaci izgledati, već strukturu podataka. Izgled podataka na ekranu definira se pomoću dodatne CSS datoteke, gdje je CSS kratica od *Cascading Styles Sheets*, tj. stilskih obrazaca. CROVALLEX ima svoju CROVALLEX .css datoteku stilskih obrazaca. XML koristi Document Type Definition (DTD) za definiranje tipa XML dokumenta. DTD navodi koji atributi i koje oznake se mogu koristiti za opisivanje sadržaja XML dokumenta te gdje se oni mogu pojaviti. Aplikacije koriste DTD da bi mogle pravilno pročitati i prikazati sadržaj dokumenta. Svrha CROVALLEX DTD-a je definirati koje će se oznake i kako koristiti u CROVALLEX XML dokumentu, tj. definira strukturu dokumenata popisom legalnih elemenata i atributa.

Osim podelemenata, *natuknica* (*word\_entry*) sadrži 2 atributa: *aspect* tj. vid glagola (s vrijednostima: *fin*, *inf*, *dual*) i atribut *freq* kojim se bilježi učestalost glagola u Čestotnom rječniku hrvatskoga jezika.

Element *headword\_lemma* također sadrži attribute: indekse za homograf ili leksički homonim (ako se radi o glagolu koji ima svoj homografski/homonimski parnjak) te atribut *reflexive* koji pokazuje može li glagol poprimiti refleksivan oblik.

Element *frame\_entry* se sastoji od 6 podelemenata, od kojih se dva javljaju obvezno: *gloss* – element koji donosi objašnjenje konkretnog značenja leme i *example* – element koji donosi primjer upotrebe leme u tom značenju preuzet iz Hrvatskog nacionalnog korpusa ili Hrvatske jezične riznice. Ostali elementi, kao što su klasa glagola (*class*) ili broj dopuna glagolu valencijskom okviru (*frame\_slots*) mogu se i ne moraju pojaviti.

Element *frame\_entry* također sadrži attribute: indeks okvira i atribut *idiom* – budući da su idiomi glagola smješteni u posebne valencijske okvire za svako značenje glagola. Podelement elementa *frame\_entry* s nazivom *frame\_slots* se može sastojati od nijednog, jednog ili više elemenata naziva *form*, što odgovara činjenice da glagoli mogu biti nevalentni, jednovalentni, dvovalentni, itd.

Element *frame\_slots* sadrži attribute: *type* (koji označava tip glagolske dopune i poprima vrijednosti *obl* ili *typ*, ovisno o tome je li dopuna glagolu obvezna ili izborna tipična) i *functor* (koji poprima jednu od 34 vrijednosti dubinskoga padeža).

Element *form* koji opisuje površinsku realizaciju pojedine glagolske dopune sadrži niz atributa, ovisno o tome kakvu morfološku realizaciju poprima dubinski padež. Pa tako imamo attribute *direct\_case*, *prepos\_case*, *prepos\_case-infinitive*, *subord\_conj*, *infinitive*, *infinitive-direct\_case*, *adjective*, *adverb* te *nominal\_predicate*. Atribut *direct\_case* se odnosi na sintaktički padež i može poprimiti vrijednosti 0-7 (skriveni subjekt u nominativu, nominativ, genitiv, dativ, akuzativ, vokativ, lokativ, instrumental) i vrijednosti *indeclinabilia\_2* (*indeclinabilia* + imenica u genitivu, npr. *100 godina*) ili *indeclinabilia* (npr. *100*). Atribut *prepos\_case* kombinacija je atributa *prepos\_lemma* i *direct\_case* (npr. *Pao je pri dolasku*). Atribut *prepos\_case-infinitive* kombinacija je atributa *prepos\_lemma* i *infinitive* (npr. *Ne preostaje nego pokušati*). Atribut *subord\_conj* pomoću veznika *subord\_conj\_lemma* uvodi zavisnu rečenicu (npr. *Podgorskoga je pozvao na dvoboj jer je držao da Olgi prijete pogibao*). Atribut *infinitive* uvodi

dopunu u infinitivu (npr. *Učio je čitati*). Atribut *infinitive-direct\_case* kombinacija je atributa *infinitive* i *direct\_case* (npr. *Tvrtka je kanila zakupiti 190 hektara njiva*). Atribut *adjective* uvodi pridjevsku dopunu (npr. *Giga se napravi sitna i uska*). Atribut *adverb* uvodi priložnu dopunu (npr. *Samardžić me je bezočno iskoristio*).

Atribut *nominal\_predicate* uvodi dopunu u obliku imenskog predikata (npr. *Viktor se ćutio krivcem*).

HTML verzija leksikona CROVALLEX v 2.008 napravljena je pomoću Perl programskog koda koji je iz XML-a izvukao elemente i njihove attribute i izgenerirao popis međusobno povezanih html stranica za svaki glagol i njegove valencijske okvire.

U skladu s inicijativom za otvoren pristup primarnim i sekundarnim istraživačkim podacima, koja sve više jača u humanističkim i društvenim znanostima, osigurana je i trajna poveznica na CROVALLEX kao značajan jezični resurs za hrvatski jezik.

CROVALLEX v 2.008 dostupan je na: <http://theta.ffzg.hr/crovallex/>.

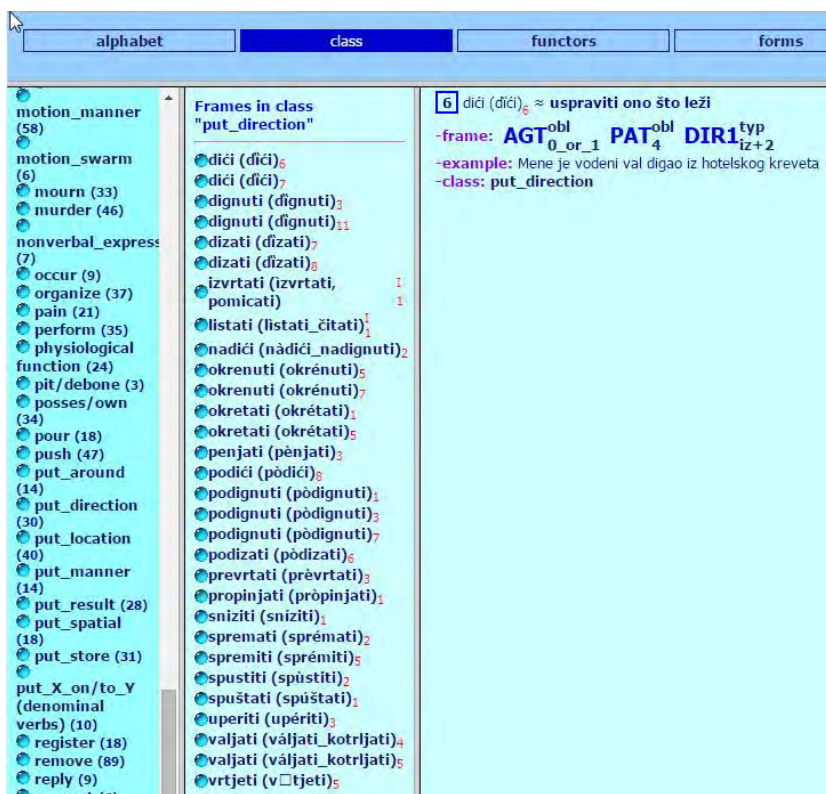
No, leksikon CROVALLEX je također uključen u platformu za pristup i dijeljenje jezičnih resursa META-SHARE ([www.meta-share.eu](http://www.meta-share.eu)), gdje je moguće pristupiti i ostalim jezičnim resursima i alatima za hrvatski jezik.

Leksikon pri otvaranju u pregledivom obliku izgleda kao na Slici 1.

**Slika 1.** Prikaz abecednog popisa glagola (prvi stupac), glagola sa slovom A (drugi stupac) i valencijskih okvira prvog glagola na popisu slova A (treći stupac)

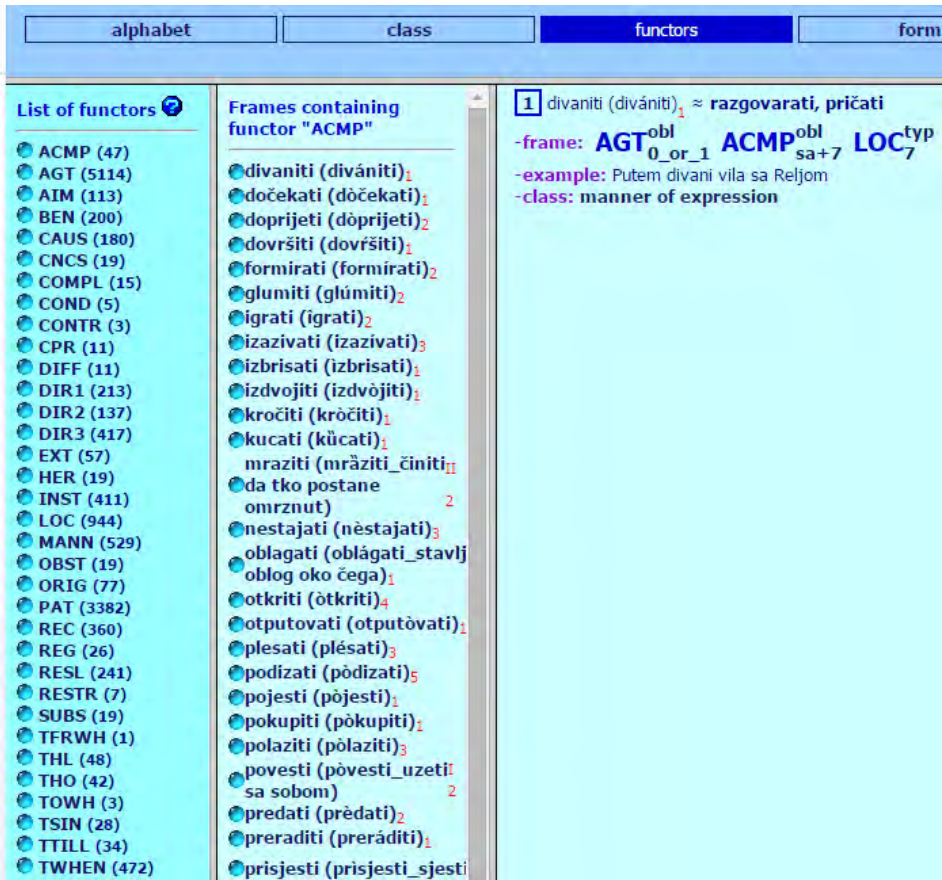
Početno sučelje leksikona u krajnje lijevom stupcu donosi abecedni popis s brojem glagola koji počinju pojedinim slovom u zagradi uz slovo abecede. Po

zadanom se odmah u srednjem stupcu učitavaju glagoli s naglašenim oblikom koji počinju prvim slovom abecede, a u krajnje desnom stupcu se učitavaju svi valencijski okviri prve natuknice koja počinje prvim slovom abecede (u slučaju glagola *analizirati* radi se o samo jednom valencijskom okviru). Uz samu natuknicu u leksikonu se nalazi vid (za glagol *analizirati* je *dual* jer se radi o dvovidnom glagolu) te broj koji pokazuje učestalost glagola u Hrvatskome čestotnom rječniku. Kratice AGT, PAT, LOC i TWHEN u valencijskom okviru glagola *analizirati* označuju dubinske padeže, *obly/typ* uz te kratice znače tip dopune (obvezna ili izborna tipična), brojevi znače morfološki padež, prijedlog + broj označava prijedložni izraz (prijedlog + morfološki padež imenice).



**Slika 2.** Prikaz popisa klasa glagola (prvi stupac), glagole iz klase *put\_direction* (drugi stupac) i valencijskih okvira prvog glagola iz klase *put\_direction* (treći stupac)

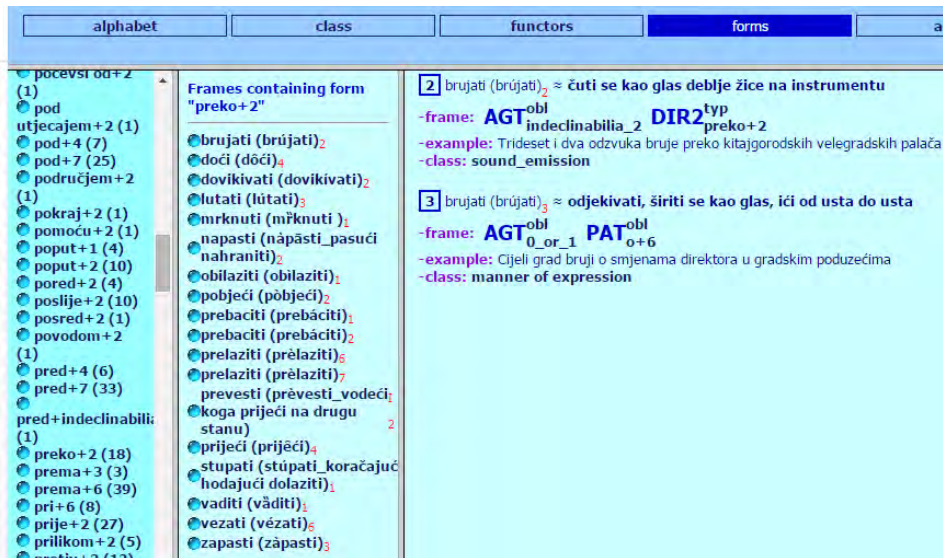
Kada korisnik klikne na gumb *class* (Slika 2), opet mu se otvaraju 3 stupca: s tim da se sada u prvom nalazi popis svih klasa glagola, u drugom popis svih glagola koji pripadaju klasi koju je korisnik izabrao klikom na nju, a u trećem se otvara valencijski okvir glagola na kojega je korisnik kliknuo u drugom stupcu.



**Slika 3.** Prikaz popisa dubinskih padeža (prvi stupac), glagole koji sadrže dubinski padež ACMP, koji je prvi na popisu (drugi stupac) i valencijskih okvira prvog glagola koji sadrži dubinski padež ACMP (treći stupac)

Kada korisnik klikne na gumb *functors* (Slika 3), opet mu se otvaraju 3 stupca: u prvom je abecedni popis kratice svih obveznih i izbornih tipičnih dubinskih padeža, u drugom popis svih glagola koji sadrže dubinski padež koji je korisnik izabrao, a u trećem se otvaraju samo okviri glagola koji sadrže dubinski padež

na kojega je korisnik kliknuo u drugom stupcu (za konkretan primjer otvara se okvir s padežom ACMP).



**Slika 4.** Prikaz popisa površinskih oblika (prvi stupac), valencijskih okvira koji sadrže prvi površinski oblik iz prvog stupca: prijedložna sintagma *preko* + 2 (drugi stupac) i valencijskih okvira prvog glagola (*brujati*) koji sadrži površinski oblik *preko* + 2 (treći stupac)

Kada korisnik klikne na gumb *forms* (Slika 4), otvaraju se 3 stupca: u prvom popis kategorija mogućih površinskih kategorizacija (9 kategorija), u drugom popis svih glagola koji sadrže površinsku strukturu koju korisnik izabere u stupcu jedan, a u trećem stupcu okviri glagola koji sadrže odabranu površinsku strukturu. Budući da za navedeni primjer, glagol *brujati*, od 3 okvira tek drugi okvir sadrži prijedložni izraz *preko* + 2, za ovaj glagol se prikazuju tek značenja od toga okvira nadalje.

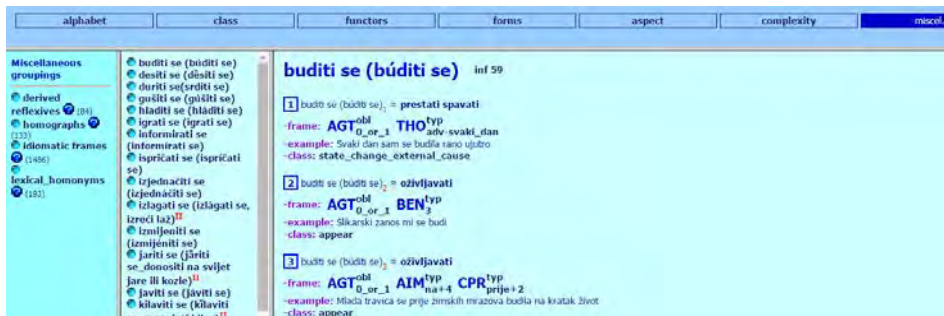
Kada korisnik klikne na gumb *aspect* (Slika 5), otvaraju se 3 stupca: u prvom se nalazi tip glagolskog vida: dual, fin ili inf (dvovidni, svršeni i nesvršeni glagoli), u drugom popis svih glagola koji sadrže vid koji korisnik izabere u stupcu jedan, a u trećem okviru glagola na koji je korisnik kliknuo u drugom stupcu.

Slika 5. Prikaz popisa glagolskih vidova (prvi stupac), valencijskih okvira koji sadrže dvovidni glagolski vid iz prvog stupca (drugi stupac) i valencijskih okvira prvog glagola koji ima dvovidni glagolski vid (treći stupac)

Slika 6. Opadajući popis broja valencijskih okvira koje pojedini glagol sadržava (prvi stupac), glagol koji sadrži najveći broj valencijskih okvira (drugi stupac) i popis valencijskih okvira glagola s najvećim brojem okvira (treći stupac)

Kada korisnik klikne na gumb *complexity* (Slika 6), otvaraju se 3 stupca: u prvom se načazi opadajući broj valencijskih okvira pojedinih glagola, počevši od glagola s najviše okvira (*ići*), u drugom je prikazan popis svih koji imaju određen broj okvira, ovisno o tome na što je korisnik kliknuo u prvom stupcu, a u trećem okviri glagola na koji je korisnik kliknuo u drugom stupcu.

Kada korisnik klikne na gumb *miscel.* (Slika 7), otvaraju se 3 stupca. U prvom stupcu nalaze se 4 elementa: *derived\_reflexives*, *homographs*, *idiomatic frames* i *lexical homonyms*. Kad korisnik klikne na *derived reflexives*, u srednjem stupcu se otvara popis svih glagola ovog tipa, a u trećem okviri onog glagola na koji je korisnik kliknuo u srednjem stupcu. Isto je i za ostala 3 elementa: *homographs*, *lexical homonyms* i *idiomatic frames*.



**Slika 7.** Popis dodatnih sintaktičkih informacija u leksikonu: izvedeni povratni glagoli, homografi, idiomi i leksički homonimi (prvi stupac), glagoli koji su homografi u leksikonu (drugi stupac) i popis valencijskih okvira prvog glagola koji je homograf u leksikonu (treći stupac)