

Glagolske večbesedne enote v učnem korpusu ssj500k 2.1

Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman,
Teja Kavčič

Uvod

- projekt označevanja kategorij glagolskih večbesednih enot (GVE)
- mednarodna COST akcija PARSEME Shared Task
- kvantitativna in kvalitativna analiza GVE

Večbesedne enote

- več definicij:
 - besedne zveze z idiomatičnim pomenom (Atkins in Rundell, 2008: 166)
 - celostni pomen ni vsota pomenov sestavin besedne zveze
- kategorizacija in označevanje glagolskih večbesednih enot (COST akcija PARSEME Shared Task 1.1)

Kategorizacija glagolskih večbesednih enot (GVE)

- Smernice projekta PARSEME Shared Task 1.1 (Bathia et al., 2017)
- jezikovni testi na podlagi pomenskih in skladenjskih lastnosti (drevesnice)

Generic decision tree

If you are annotating **Italian** or **Hindi**, go to the Italian-specific decision tree ([?page=it-dectree](#)) or Hindi-specific decision tree ([?page=hi-dectree](#)). For all other languages follow the tree below.

- ↳ Apply test S.1 ([?page=structural#test-s1](#)) (prev. 6) - [**1HEAD**: Unique verb as functional syntactic head of the whole?]
 - ↳ **NO** ⇒ Apply the VID-specific tests ([?page=vid#decision-tree-vid](#)) ⇒ *VID tests positive?*
 - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
 - ↳ **NO** ⇒ It is not a VMWE, **exit**
 - ↳ **YES** ⇒ Apply test S.2 ([?page=structural#test-s2](#)) (prev. 7) - [**1DEP**: Verb *v* has exactly one lexicalized dependent *d*?]
 - ↳ **NO** ⇒ Apply the VID-specific tests ([?page=vid#decision-tree-vid](#)) ⇒ *VID tests positive?*
 - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
 - ↳ **NO** ⇒ It is not a VMWE, **exit**
 - ↳ **YES** ⇒ Apply test S.3 ([?page=structural#test-s3](#)) - [**LEX-SUBJ**: Lexicalized subject?]
 - ↳ **YES** ⇒ Apply the VID-specific tests ([?page=vid#decision-tree-vid](#)) ⇒ *VID tests positive?*
 - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
 - ↳ **NO** ⇒ It is not a VMWE, **exit**

Kategorije GVE

- Univerzalne kategorije:
 - zveze z glagoli v pomensko oslavljeni rabi (Light Verb Constructions)
 - celostne (LVC.full)
 - vzročnostne (LVC.cause)
 - glagolski idiomi (Verbal Idioms)
- Jezikovno specifične kategorije:
 - inherentno povratni glagoli (Inherently Reflexive Verbs)
 - zveze glagola z izpredložnim morfemom (Verb-Particle Constructions)
 - predložnomorfemski glagoli (Inherently Adpositional Verbs)

Označevanje

- Učni korpus ssj500k 2.0 (Krek et al., 2017): 28.000 povedi iz FidaPLUS (Arhar Holdt in Gorjanc, 2007)
- Prva faza: 11.411 povedi, dve označevalki
- Druga faza: 2.100 povedi
- Kategorizacija 3.364 glagolskih večbesednih enot

Razpored označenih enot glede na kategorije

Kategorija	Vse GVE	Delež	Različne GVE
Inherentno povratni glagoli	1.627	48 %	345
Predložno morfemski glagoli	710	21 %	154
Glagolski idiomi	724	22 %	457
Vzročnostni glagoli v pomensko oslabljeni rabi	64	2 %	27
Celostni glagoli v pomensko oslabljeni rabi	239	7 %	103
Skupaj	3.364	100 %	1.086

Inherentno povratni glagoli

- Glagoli s prostim morfemom *se/si*:
 - a) brez morfema glagol ne obstaja (*zdeti se*) – pogostejši
 - b) morfem glagolu spreminja funkcijo in/ali pomen (*dati se* = ‚moči‘):
- Zveze, kjer morfem *se/si* izraža pasiv in povratni glagoli (*umiti se* = ‚umiti sebe‘), niso označene.

Zveze z glagoli v pomensko oslabljeni rabi (LVC)

- Glagol + samostalnik oz. samostalniška besedna zveza (tudi v obliki predložne zveze)
- Samostalnik povezan s stanjem ali dogajanjem
- Podkategoriji:
 - Celostne zveze: glagol prispeva k pomenu le na kategorialni ravni (*biti v pomoč*)
 - Vzročnostne zveze: osebek je vzrok ali vir stanja/dejanja (*imeti učinek*)
- Delitev na podkategorije glavna razlika z dotedanjimi slovenskimi opredelitvami

Poskus delitve samostalnikov v pomenske skupine (LVC)

- Vzročnostne zveze:
 - opredelitev vrste rezultata (*imeti učinek*)
 - pozitivna (*biti v užitek*) ali negativna (*imeti preglavice*) posledica
- Celostne zveze – načrtovanje in ocenjevanje uspeha:
 - komunikacija (*imeti mnenje*)
 - potencial za (ne)uspeh (*imeti prednost, imeti napake*)
 - začetni koraki (*dajati napoved*)
 - negativno stanje (*biti strah*)
 - pozitivne lastnosti (*imeti moč*)
 - doseženi rezultati (*imeti status*)
 - odnos do še nerealiziranih ciljev (*imeti ambicije*)

Predložno morfemski glagoli

- Glagoli s predložnim morfemom:
 - a) brez katerega ne obstajajo (*simpatizirati z*)
 - b) ki jim predložni morfem spremeni pomen (*biti za – ,strinjati se`, priti do – ,zgoditi se`*)
- Leksikalizirani predložni morfem običajno sledi glagolu: neposredno ali v neposredni bližini.
- Kadar je morfem pred glagolom, je razdalja med njima občutno večja → upoštevanje široke okolice glagola pri strojnem prepoznavanju

Predložno morfemski glagoli (nadaljevanje)

- Razdelitev glede na pomen:
 - izražanje funkcije/lastnosti (*veljati za (favorita)*)
 - izražanje (ne)strinjanja (*biti za (globalizacijo)*)
 - izražanje izhajanja/upoštevanja (*temeljiti na (dejstvu)*)
 - začetek ali spremembo stanja/dejanja, spremembo lastnosti (*pasti v (komo)*)
 - preživljanje, prestajanje (*iti skozi (proces)*)
 - aktivno delovanje (*ukvarjati se z*)

Glagolski idiomi

- Zveza vsaj dveh leksikaliziranih sestavin z glagolom kot skladijskim jedrom
- Razlika s slovensko tradicionalno opredelitvijo: možnost opravljanja povedkove funkcije zveze kot celote (problem: zveza v vlogi predmeta: *(ne spodobi se) voditi za nos*).

Glagolski idiomi (nadaljevanje)

- Najpogostejše zveze glagola in predložne zveze (*biti na dosegu roke*).
- Pogoste kombinacije z glagoloma *biti/imeti*, ki jih lahko zaradi strukturne ustaljenosti ali pomenske izpraznjenosti obravnavamo kot ustaljene glagolske zveze (*biti všeč, imeti pojma*)
- Poleg tega še kombinacije glagola in
 - samostalniške besedne zveze
 - prislova (*ustreliti mimo*)
 - zaimenskega morfema (*zagosti jo (komu)*)
- Stavčne strukture (*samo osel gre dvakrat na led*)

Zaključek

- Prve kvantitativne in kvalitativne analize kategorij GVE
- Seznam GVE predstavlja nabor enot in izhodišče pravil za avtomatsko prepoznavanje in strojno učenje

Načrti za prihodnje delo

- Izdelava tipologije neglagolskih večbesednih enot (*kaplja v morje, v skladu z*)
- Rezultati uporabni pri izdelavi elektronskega leksikona VE in drugih jezikovnih priročnikov (Slovar sodobnega slovenskega jezika (Gorjanc et al., 2015) in na korpusu temelječa znanstvena slovnica)

Zahvala

Raziskava je potekala v okviru projekta ARRS J6-8256 Nova slovnica sodobne standardne slovenščine: viri in metode ter sodelovanja v okviru IC1207 PARSEME COST Action in IS1305 ENeL COST Action.

Hvala