# Sketch Grammar: RegEx-over-POS or Dependency Parser?

## A Comparison Of Two MWE Extraction Methods

Simon Krek*, Kaja Dobrovoljc**

*Jožef Stefan Institute, Artificial Intelligence Laboratory; **Trojina, Institute for Applied Slovene Studies, Ljubljana, Slovenia

PARSEME 2nd general meeting, 10–11 March 2014, Athens, Greece; WG 1 and WG 3

## Introduction

Word sketch is a corpus-based summary of a word's grammatical and collocational behavior that enables the extraction of collocations (and corpus examples) using the Sketch Engine tool. A detailed sketch grammar for Slovene, based on regular expressions over POS tags, was developed for the extraction of lexical data from the Gigafida corpus for the purposes of compiling Slovene Lexical Database. Since the adaptation of the MSTParser for Slovene, lexical data based on the same or similar grammatical patterns can also be extracted from parsed corpus data. We compare the difference between the two „sketch grammars“ both in terms of general syntactic analysis (1) and MWE extraction and evaluation (2).

## RegEx-over-POS based word sketches

### RegEx-WS

- a series of grammatical relations (gramrels) using regular expressions over POS-tags in a tagged corpus
- developed for the extraction of lexical data from the 1 billion Gigafida corpus for the purpose of compiling **Slovene Lexical Database (SLD)**
- number of gramrels: **105** (v.16)
- e. g. gramrel describing adjectival premodification of nouns:

```
=modifier/head
2: [tag="A.*"][tag!="[VNCS].*" & word!="[,:;()-]"]{0,5} 1: [tag!="N.*"]
```

## Dependency parser based word sketches

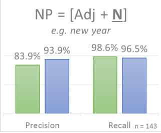
### DepPars-WS

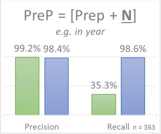
- Minimum-Spanning Tree Parser (**MSTParser**)
- trained on the ssj500k corpus (235.864 tokens, **~11.400 sentences**)
- **10 labels** (5 for phrases, 4 for sentence elements, 1 for root)
- overall accuracy **90.43%** (unlabelled), **87,52%** (labelled)
- ssj500k: http://eng.slovenscina.eu/tehnologije/ucni-korpus (CC BY-NC-SA 2.5 SI)
- Dependency Parser: http://eng.slovenscina.eu/tehnologije/razclenjevalnik (Apache License v2.0)

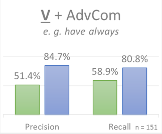
## 1. Comparison of Parsing Precision/Recall

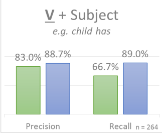
### 1.1 Method

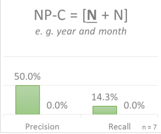
For the most frequent lemmas in the gold ssj500k corpus, we compared the recall and precision of both sketch grammars for extraction of collocates (i. e. dependents) within the given set of grammatical relations (4 for noun, 4 for verb and 2 for adjectives), regardless of their MWE status.

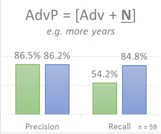


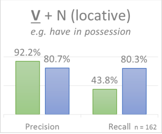


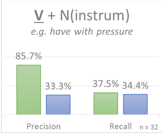


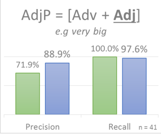


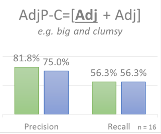








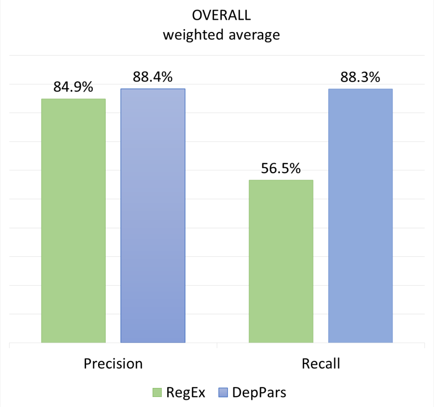




N = leto (*year*); V = imeti (*have*), Adj = velik (*big*)

### 1.2 Results

Overall (see →), for currently comparable set of grammatical relations (10) between selected heads (lemmas) and their dependents (collocates), dependency parser gives slightly higher precision and significantly higher recall. However, the results for both methods vary considerably depending on the type of grammatical relations (see ←). The most significant differences can be observed in the recall for discontinues syntactic relations (e. g. prepositional phrases), where dependent is often farther away from the head.



## 2. Comparison of MWE Extraction

### 2.1 What MWEs were we interested in?

**Phraseological units** in SLD are defined as word combinations whose meaning or communication function is **not deducible from its parts** and have **metaphoric meaning**, as opposed to multi-word units, whose meaning remains non-metaphorical.

### 2.2 Where did we extract them from?

The **100 million word Kres corpus** is an extensive collection of Slovene texts with a balanced genre structure. It was sampled from the 1BW Gigafida corpus, with random paragraphs as basic sampling units to ensure better representation of the original Gigafida material.

### 2.3 What was our gold?

Slovene Lexical Database (SLD) consists of lexical data of various degrees of compositionality: 44.626 collocations, 7.151 grammatical patterns, 8.298 syntactic combinations (compositional), as well as 2.053 multi-word units and **1.446 phraseological units** (non-compositional).

### 2.4 Method

For each of the 10 comparable grammatical relations (see 1), we chose 3* random MWEs from SLD, whose phraseological core (bigram) can be described by such relation. For the head node of every bigram, we then compared the word sketch for the given gramrel, in particular: a) the **position** of the MWE collocate within the gramrel sketch (rank)**; b) the attributed **collocational strength** (logDice score); and c) the number of **matched corpus concordances**. Note that the latter does not imply recall, as the retrieved examples may or may not be relevant.

| Gramrel | MWE | head-dependant bigram | rank in sketch | | | salience (logDice) | | | examples (potential) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | RE | DP | diff | RE | DP | diff | RE | DP | diff |
| NP = [Adj + N] | ***besedna driska*** **(lexical diarrhea)**<br>to talk a lot without saying anything important | driska - beseden | 13 | 14 | -1 | 5.4 | 5.2 | 0.3 | 4 | 3 | 1 |
| | ***oslova senca*** **(donkey's shadow)**<br>something insignificant | senca - oslov | 3 | 4 | -1 | 8.1 | 8.2 | -0.1 | 15 | 15 | 0 |
| | ***dvorezen meč*** **(double-edged sword)**<br>a benefit that is also a liability | meč - dvorezen | 1 | 1 | 0 | 11.9 | 12.0 | -0.1 | 82 | 81 | 1 |
| PreP = [Prep + N] | ***do absurdnosti*** **(to absurdity)**<br>to the utmost | absurdnost - do | 1 | 1 | 0 | 0.9 | -0.2 | 1.1 | 8 | 8 | 0 |
| | **od časa do časa (from time to time)**<br>occasionally | čas - od | 5 | 11 | -6 | 7.8 | 7.3 | 0.5 | 1154 | 1563 | -409 |
| | ***(prebrati kaj) na dah*** **(in one breath)**<br>to read something in very short time | dah - na | 2 | 2 | 0 | 0.2 | -0.6 | 0.8 | 35 | 30 | 5 |
| AdvP = [Adv + N] | ***zastava na pol droga*** **(flag at half-mast)**<br>sign of one's death and their mourning | drog - pol | 1 | 1 | 0 | 4.8 | 4.7 | 0.0 | 18 | 18 | 0 |
| | ***malo denarja, malo muzike*** **(little money, little music)**<br>cliché similar to 'you get what you pay for' | muzika - malo | 1 | 1 | 0 | 3.5 | 3.8 | -0.3 | 20 | 21 | -1 |
| NP-Coord = [N + N] | ***bog in batina*** **(God and bludgeon)**<br>absolute authority | bog - batina | 1 | 1 | 0 | 9.0 | 9.3 | -0.4 | 19 | 23 | -4 |
| | ***kruha in iger*** **(bread and games)**<br>superficial means of appeasement | kruh - igra | 41 | 39 | 2 | 5.1 | 6.1 | -1.0 | 50 | 56 | -6 |
| | ***sama kost in koža*** **(nothing but skin and bone)**<br>extremely thin | kost - koža | 6 | 8 | -2 | 7.2 | 7.0 | 0.1 | 116 | 77 | 39 |
| Verb + AdvCom | ***spretno sukati pero*** **(to spin pen in a skilfull manner)**<br>be good at writing | sukati - spretno | 2 | 2 | 0 | 7.4 | 7.8 | -0.4 | 10 | 11 | -1 |
| | ***športno prenesti poraz*** **(bear defeat in a sportsmanlike manner)**<br>accept defeat gracefully | prenesti - športno | 3 | 7 | -4 | 7.0 | 6.8 | 0.1 | 5 | 6 | -1 |
| | ***ustreliti mimo*** **(shoot past something)**<br>make a wrong choice or decision | ustreliti - mimo | 5 | 7 | -2 | 6.5 | 6.6 | -0.2 | 11 | 13 | -2 |
| Verb + Subject | ***naj me koklja brcne*** **(let the hen kick me)**<br>intensifier similar to 'I'll be damned (if)' | brcniti - koklja | 1 | 1 | 0 | 8.6 | 9.7 | -1.1 | 3 | 8 | -5 |
| | ***kri ledeni (v žilah) komu*** **(blood freezes up in one's veins)**<br>sentiment of fear or horror | ledeneti - kri | 1 | 1 | 0 | 4.6 | 5.1 | -0.5 | 13 | 13 | 0 |
| | ***kri vre (komu)*** **(blood boils)**<br>excitement and impatience, either positive or negative | vreti - kri | 4 | 3 | 1 | 4.6 | 5.4 | -0.8 | 13 | 16 | -3 |
| Verb + Noun (locative) | ***utapljati kaj v alkoholu*** **(drown someting in alcohol)**<br>to drink alcohol in order to forget your problems | utapljati - alkohol | 2 | 2 | 0 | 5.5 | 5.6 | -0.1 | 12 | 10 | 2 |
| | ***pustiti (koga) na cedilu*** **(leave someone on a strainer)**<br>to leave someone at a time when they need you to stay | pustiti - cedilo | 1 | 1 | 0 | 12.6 | 11.2 | 1.4 | 466 | 505 | -39 |
| | ***(ne) priplavati po juhi*** **(not  swim the soup)**<br>not stupid and cannot be easily deceived | priplavati - juha | 2 | 2 | 0 | 3.9 | 4.8 | -1.0 | 4 | 4 | 0 |
| Verb + Noun (instrumental) | ***zamahniti s čarobno paličico*** **(to wave a magic wand)**<br>to solve a difficult problem with no effort | zamahniti - paličica | 4 | 3 | 1 | 6.9 | 7.3 | -0.4 | 8 | 7 | 1 |
| | ***obrisati se pod nosom za denar*** **(wipe under one's nose for the money)**<br>to not get the awaited money | obrisati - nos | 7 | 7 | 0 | 8.5 | 8.3 | 0.1 | 98 | 71 | 27 |
| | ***opletati z jezikom*** **(swing one's tongue)**<br>to gossip or to chatter | opletati-jezik | 2 | 3 | -1 | 3.4 | 3.8 | -0.4 | 16 | 16 | 0 |
| AdjP = [Adv + Adj] | ***povedati komu nekaj krepkih*** **(to tell someone some strong ones)**<br>to speak angrily to someone because they have done | krepek - nekaj | 3 | 3 | 0 | 3.4 | 2.9 | 0.5 | 53 | 30 | 23 |
| AdjP-Coord = [Adj + Adj] | ***jasen in glasen*** **(clear and loud)**<br>expressed in a determined and straightforward manner | jasen - glasen | 7 | 7 | 0 | 7.3 | 7.4 | -0.2 | 20 | 19 | 1 |
| | ***slep in gluh za kaj*** **(blind and deaf for something)**<br>desribes a person unwilling to consider or do something they find unpleasant | slep - gluh | 2 | 2 | 0 | 10.5 | 10.3 | 0.2 | 81 | 66 | 15 |

*For gramrels with less than 3 MWEs overall, all instances were chosen. ** Default settings for word sketches: minimum_frequency = 3, minimum_score = -9999

### 2.5 Results

See Table (←)

🟨 Identical rank/score/no. of concordances for both
🟩 Higher rank/score/no. of concordances by RegEx-WS
🟦 Higher rank/score/no. of concordances by DepPars-WS

The two sketches give **very similar** results, i.e. **high precision** for extracting MWEs. The dependency-parser based word sketches attribute the MWEs a slightly greater collocational strength (logDice score), although this does not usually change the rank position of the collocate in question, as both sketches usually display the MWE collocates in the same (top-level) positions.

## Future work

- determine syntactic patterns for all MWE types in the SLD database
- for these patterns, define comparable grammatical relations in both sketches
- develop procedures for automated comparison of the two methods in terms of MWE parsing, extraction and evaluation (beyond core bigrams)
- on the basis of results, build a hybrid model that combines the best features of both methods
- further explore the SLD gold standard of more than 60.000 MWEs for machine learning (extraction and MWE type classification)

## References

- Dobrovoljc, K., Krek,S., Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino = Slovene Dependency Parser. Proc. Language Technologies Conference. Ljubljana, Slovenia.
- Kilgarriff, A., Rychly, P., Smrž, P., Tugwell, D. (2004). The Sketch Engine. Proc. Euralex. Lorient, France. 105-116.
- Gantar, P., Krek, S. (2011): Slovene Lexical Database. Proc. Natural language processing, multilinguality : sixth international conference. Modra, Slovakia. 72-80.
- Kosem, I., Krek, S., Gantar, P. (2013). Automatic extraction of data: Slovenian case revisited. SKEW-4: 4th International Sketch Engine Workshop. Talinn, Estonia.