Comparative Idioms in Croatian: MWU Approach

Kristina Kocijan, Sara Librenjak
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences,
University of Zagreb

Keywords: phraseology, comparative structures, syntactic grammars, Croatian, NooJ.

Abstract

This article presents the work aiming to describe comparative idioms in Croatian language for computational processing using NooJ linguistic environment. As a part of a larger project concentrated on annotating and extracting different Croatian idioms as multi-word units (MWUs), this work aims to present automated comparative idiom search in any Croatian text. Using NooJ environment, a user can find any comparative structure in a text and use it for translation, language learning or research purposes.

1. INTRODUCTION

Croatian phraseology is a young discipline whose beginning dates in 1970ies and since that time it was analyzed from different perspectives including semantical, syntactic, etymological, sociolinguistic, stylistic and discourse analysis (Matešić, 1978; Menac, 1978; Mence and Mihalić, 2007), and only recently in computational linguistic (Ljubešić et. al., 2014, Kocijan & Librenjak, 2015).

As the idioms are rooted in the tradition of the language and the society from which they hail from, they need a special treatment in computational linguistics. With our tool, Croatian idioms can be successfully detected in order to be properly matched with their translation in corresponding language, which would eliminate awkward and completely wrong automated translations. Thus, this work seeks to aid not only the successful development of additional language resources for Croatian language (Vučković et al., 2010), but a possible assistance in future work relating to machine assisted translation.

In this article, first we introduce the methodology and corpora constructed for the purpose of this work. Then, we explain the technical details of the software which automatically recognizes comparative idioms. Since it was made for the NooJ linguistic environment, we will use the terms pertaining to NooJ software, and thus hope to encourage readers to try and use this environment for their own projects. We will explain the construction of electronic dictionary containing all the necessary comparative idioms in chapter 3, while chapter 4, which explains the types of comparative structures and their treatment in NooJ, takes up the largest portion of this article. For more detailed explanation of NooJ environment see Silberztein (2003). In the last chapter, we will present the evaluation of our tool and comment on its possible uses.

2. METHODOLOGY AND CORPORA

If one wants to make an automated idioms detection tool, one must first collect all the relevant idioms, and construct a digital dictionary. For this purpose, Croatian dictionary of idioms (Menac et al., 2003) was used as a reference, but the list grew in the process of construction, and some additional idioms were added. In our previous work (Kocijan and Librenjak, 2015), we presented an all-around idiom detection tool using the NooJ linguistic environment. This work concentrates on one type of idioms – comparative idioms containing words such as "like" or "as if".

In the second phase, collected idioms were sorted into categories by their syntactic properties. At the same time, we began construction of NooJ files for syntactic processing of texts containing the idioms (called NooJ grammars), which will be discussed in detail in chapter 4.

Simultaneously, three different types of corpora were specifically constructed for the purpose of this work. A smaller corpus of sentences containing only comparisons was made for training the grammars. Then, we constructed two larger corpora for testing. For this research, we specifically made two stylistically different corpora in order to collect statistical data about frequency of comparative idioms in Croatian texts. First test corpus is a general text corpus, collected from the Web, while the second one is a literal works corpus. In the final chapter we will discuss the results in the different corpora and implications about Croatian stylistics which follow from the results.

3. DICTIONARY

There are 533 main entries in the dictionary of comparative idiomatic expressions. Here we define a main entry as the word occurring in the 1st position of an idiom (noun, adjective or a verb) that has a word before *kao* ("like","as"). However, in the null category, i.e. an idiom that starts with *kao*, a main entry is considered the first word that comes after *kao* if it allows the change (*kao heroj – en$^{lit.}$ as a hero*, *kao crvena jabuka – en$^{lit.}$ like a red apple*). The change of the first word is recognized via FLX attribute that is linked to the name of a paradigm, while the change of the second word is recognized via grammar. If no change occurs, entire phrase (without conjunction) is entered as a main entry (*kao ispod čekića – en$^{lit.}$ as if under the hammer*).

- `heroj,NW+Type=50n+FLX=BRATIĆ`
- `crven,NW+Type=50an+FLX=PRESPOR+SUFX=jabuka`
- `ispod čekića,NW+Type=500`

To mark the 2nd part of an idiom, we have used the attributes SUFX (suffix), SUFXX (suffix X), SUFXA (suffix A), SUFXB (suffix B) and SUFXC (suffix C) in the following manner: SUFX holds a single word, SUFXX holds an optional expression (one or more words) that may be omitted (*slobodan kao ptica [na grani] – en$^{lit.}$ free as a bird [on a branch]*), SUFXA and SUFXB split multiple word expression so that we can accommodate more possibilities for the main entry (*šaka u oko – en$^{lit.}$ a punch to the eye* and *šaka u glavu – en$^{lit.}$ a punch in the head*) or to divide the expression into the part that may and may not change (*siromašan kao crkveni miš – en$^{lit.}$ poor as a church mouse*, *mlad kao rosa u podne – en$^{lit.}$ young as a noon dew*) and SUFXC holds the conjunction for coordination (*razlikovati se kao nebo i zemlja – en$^{lit.}$ be different as heaven and earth*).

- `slobodan,NW+Type=5an+SUFX=ptica+SUFX=ptičica+SUFXX=na grani+FLX=DIVAN`
- `šaka,NW+Type=50np+SUFXA=u+SUFXB=oko+SUFXB=glavu`

- `siromašan,NW+Type=5aan+SUFXA=crkveni+SUFXB=miš+FLX=DIVAN`
- `mlad,NW+Type=5anp+SUFXA=rosa+SUFXB=u podne+FLX=MLAD`
- `razlikovati,NW+pov+Type=5vn+FLX=RAZLIKOVATI+SUFXA=nebo+SUFXC=i+SUFXB=zemlja`

It is possible that one main entry has one or more SUFX, SUFXA and SUFXB attributes. So for the dictionary entry:

- `crven,NW+Type=5an+SUFX=paprika+SUFX=rak+SUFX=krv+SUFX=mak+SUFX=paradajz+FLX=PRESPOR`

an adjective 'red' (hr: *crven*) has five SUFX attributes, meaning that it is found in five different idiomatic structures of the same subclass. Thus, all valid expressions *crven kao paprika* (en: red as a pepper), *crven kao rak* (en: red as a lobster), *crven kao krv* (en: red as blood), *crven kao mak* (en: red as a poppy), *crven kao paradajz* (en: red as a tomato) are recognized. This variability in form is usually due to synonymy of possible SUFX parts (*dati znak* | *mig* | *signal*). Although this is not always the case, the meaning still remains the same (*tko* | *vrag bi ga znao*) (Menac, 1978). However, there are occurrences in our dictionary that have SUFX parts which are quite opposites like

- *osjećati se kao riba na suhom* – en[lit] *to feel as a fish on dry land* -> en. feel very bad
- *osjećati se kao riba u vodi* – en[lit] *to feel as a fish in water* -> en. feel very good

Any SUFXA may be matched with any SUFXB if found inside the same main entry description. If SUFXA and SUFXB values must not appear together, they have to be entered as a new main entry (*bježati kao štakori s broda koji tone, bježati kao vrag od tamjana, bježati kao đavo od tamjana*)

- `bježati,NW+Type=5vn+FLX=BOJATI+SUFXA=štakor+SUFXB=s broda koji tone`
- `bježati,NW+Type=5vn+FLX=BOJATI+SUFXA=vrag+SUFXA=đavo+SUFXB=od tamjana`

Thus, although there are 533 main entries in the dictionary, it actually holds 858 different comparative idioms. Considering all the valid possibilities due to the flective property of nouns, verbs, adjectives and pronouns found in CI and the somewhat free order of its constituents, we are able to recognize many more occurrences by building a syntactic grammar in NooJ. The grammar uses the value of an attribute 'Type' from the dictionary entries to define which words can be found in particular expression. We will explain this attribute in more detail in the following section.

The distribution of comparative idiomatic expressions in the dictionary is given in Figure 1 with the following legend:

- the top row holds the names of CI Types;
- the bottom row holds the description of particular subtype;
- the middle row shows how many dictionary occurrences there are, considering their type, subtype but also if the 1st part (+ – null 1st part; ○ – 1st part changes; □ – 1st part doesn't change) or the 2nd part of CI changes (green – 2nd part fixed; purple – 2nd part changeable).

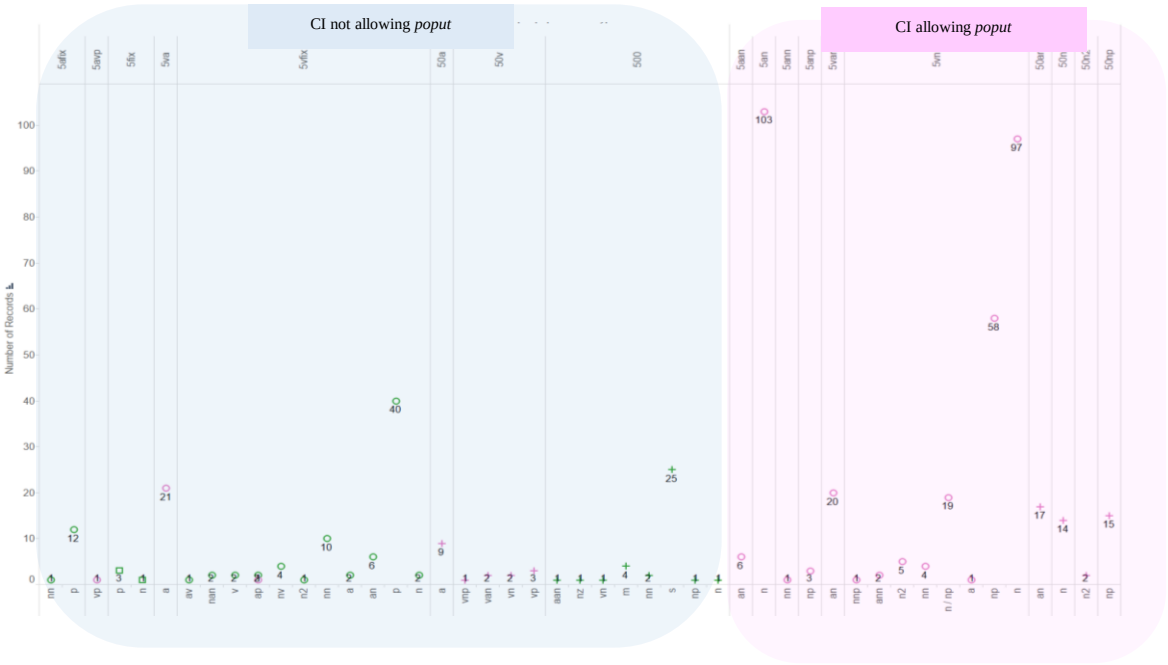

Figure 1: **Distribution of comparative idioms in NooJ dictionary**

4. TYPES OF COMPARATIVE IDIOMATIC EXPRESSIONS

Croatian language is very rich in idiomatic expressions and comparative idioms make one out of five different types of such constructions (Kocijan et al., 2015) that slightly differ from Matešić (1978). We have observed their occurrences in the text and where thus able to differentiate different subtypes inside this category.

Although all of the comparative expressions appear with conjunction *kao* | *ko* (en. as | like), some of them also appear in corpus with preposition *poput* (en. such as). However, this is only true for idioms that in second position (i.e. after conjunction *kao*) have either a single noun in nominative (*bijel kao snijeg* -> en. white as a snow), an adjective and a noun both in nominative (*razbježali se kao rakova djeca* -> en^{lit} to spread around as crab's children, *siromašan kao crkveni miš* -> en^{lit} poor as a church's mouse), a noun in nominative followed by a noun in genitive (*raste kao gruda snijega* -> en^{lit} to grow as a snow ball) or a noun in nominative followed by a prepositional phrase (*drhtati kao list na vjetru* -> en^{lit} to shiver as a lief on the wind; *kao guska u magli* -> en^{lit} as a geese in the fog). This is important since the preposition *poput* does not take nominative construction but genitive one. We have solved this through our grammar since the dictionary allows for only one paradigm to be defined and only for the main entry. Thus, although some expressions were classified at first as fixed or null subtype (Kocijan et al., 2015), we decided to move them into new categories that will allow for the genitive case to be recognized as well.

The second characteristic taken into account was to divide comparative idioms regarding the possibility to change case, number, gender or tense of either only first part of the expression or both first and second part. This required a detailed investigation of all the dictionary entries in order to build the grammars that will recognize all and only valid constructions. Matešić (1978) states that only the verb, adjective and noun positioned before *kao* may change in tense (verbs), gender (verbs and adjectives), number (all three) and case (adjectives and nouns). In the case that a change occurs on any of the words that come after *kao*, the structure ceases to be an idiom. So while *živjet ćemo kao bubreg u loju* (en: we will live as kidney in suet) is an idiom, the composition *živjet ćemo kao bubrezi u loju* (en: we will live as kidneys in suet) is not yet a fully

recognized idiom. Still, it can be considered an idiomatic neologism, and is treated as such in our work. This type of language creativity is quite rare according to the corpus, but we predicted it and they are recognized by grammars in case they appear.

There are 18 possible values (50a, 50v, 500, 5fix, 5afix, 5vfix, 5avp, 5va, 5an, 5aan, 5ann, 5anp, 5vn, 5van, 50an, 50n, 50n2, 50np) for the attribute 'Type' coded in the following manner:

- 1st position – **5** : denotes the comparative type of idioms;
- 2nd position – **0** : denotes the null subclass | **v** : denotes the verbal subclass | **a** : denotes the adjectival subclass | **fix** : denotes no change in any part of CI;
- 3rd position – **fix** | **0**[1] : denotes no change in 2nd part of CI | **a | an | ann | n | nn | np | nnp | n2 | vp** : first letter of word category[2] found in 2nd part of CI in the order of the appearance.

First we will look deeper into the comparative idioms (CI) that do not allow *poput* and second into those that allow it.

4.1. CI not allowing poput

There are 165 dictionary entries that do not allow poput. Of that number, 53 have null 1st position (36 fixed 2nd position and 17 change 2nd position occupied by an adjective or a verb), 4 have a fixed 1st position (and fixed 2nd position) and remaining 108 have changeable 1st position (85 fixed 2nd position and 23 change 2nd position occupied by an adjective) with 92 verbs and 14 adjectives in the 1st position.

The fixed 2nd position is occupied either with a prepositional phrase (*bježati kao od kuge – en*$^{lit.}$ *run like from a plague*), or an adjective (*doći kao naručen – en*$^{lit.}$ *come as ordered*), or Dative noun and Nominative noun (*pristajati kao kravi sedlo – en*$^{lit.}$ *fit as a saddle fits a cow*), or Nominative noun and Genitive noun (*pun kao šipak koštica – en*$^{lit.}$ *full like a grenadine is full of seed*).

We recognize eight subtypes in this group of CI coded as 500, 50a, 50v, 5fix, 5afix, 5vfix, 5va and 5avp. Inside the grammar, we have grouped them together if they show similar patterns in their usage.

4.1.1. Type 500

Type 500 CI are all **null CI**, with no 1st position and no changes in the 2nd position. This category has mainly prepositional phrases after conjunction *kao* | *ko* and since there are no changes, entire phrase is entered in the dictionary as a main entry (*kao na iglama, kao od šale, kao u raju*):

- `na iglama,NW+Type=500`
- `od šale,NW+Type=500`
- `u raju,NW+Type=500`

[1] Where there is no change of 2nd part of CI, there was no need to mark the word category so we used word 'fix' or '0' instead to mark this section. However, the distribution of word categories found can be seen in Figure 1.

[2] Word categories found are: a – adjective, n – noun, p – prepositional phrase, n2 – coordination of 2 nouns, v – verb.

4.1.2. Types 50a and 50v

Types 50a and 50v are similar in a way that they are both **null CI**s but, contrary to the type 500, they have a first word in the 2nd part (adjective or verb) that changes. We were able to accommodate for this change via dictionary and the attribute FLX.

- `lud,NW+Type=50a+FLX=MLAD`
- `pasti,NW+Type=50v+FLX=SJESTI+SUFXA=s+SUFXB=Marsa+SUFXB=neba`

The remaining parts are recognized via grammar (see Figure 1.) so that entire expression is marked as a phraseme, i.e. as <PHR+FRAZEM=*kao da je pao s Marsa*+Type=50v> where **PHR** (code for phraseme) denotes that the string is a phraseme, **+FRAZEM** holds the recognized string and **+Type** holds the type of the recognized string which is inherited from the type of the main CI word placed inside the variable **F**.

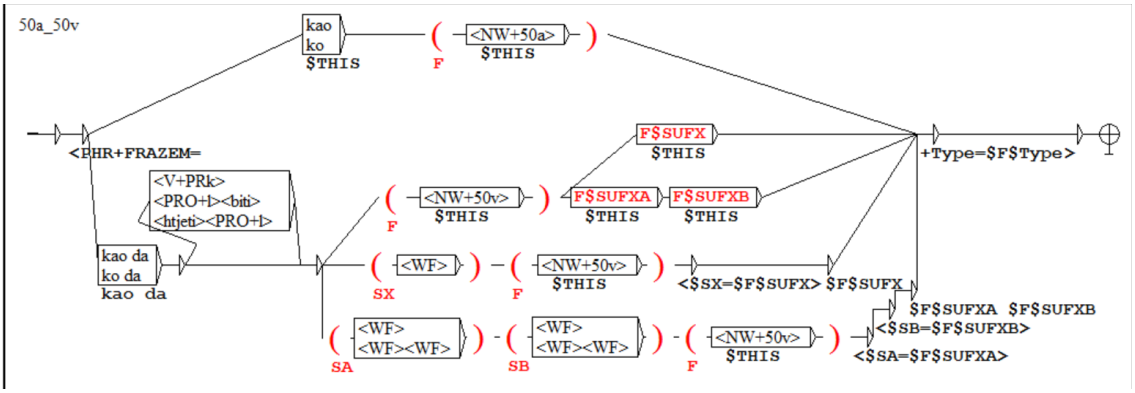

Figure 2 - Grammar recognizing null CI with changeable 2nd part (Types 50a and 50v)

4.1.3. Type 5fix

Although there are not many CI that start with a noun (**nominal CI**), we have decided to put them in a separate category. We marked it as '5fix' since neither section, i.e. one prior to and one following conjunctions '*kao*' or '*ko*', changes its form or position inside the expression. Thus the dictionary entries for this category do not require any paradigm to be connected to them (no FLX attribute).

Dictionary entries for subtype 3 (*mrak kao u rogu – en*$^{lit.}$ *dark as if you're inside a horn, tišina kao u crkvi – en*$^{lit.}$ *silence like in the church, tišina kao u grobu – en*$^{lit.}$ *silence as if you were in a grave*):

- `mrak,NW+Type=5fix+SUFXA=u+SUFXB=rogu`
- `tišina,NW+Type=5fix+SUFXA=u+SUFXB=crkvi+SUFXB=grobu`

4.1.4. Types 5va and 5avp

The category 5va belongs to **verbal CI** while 4avp belongs to **adjectival CI**. They both have changeable parts in the 1st and 2nd part of an expression and do not allow *poput*. The change in the 1st part is recognized via dictionary as in the previous two categories:

- `osjećati,NW+Type=5va+pov+FLX=SJATI+SUFX=preporođen`
- `gol,NW+Type=5avp+FLX=CRN+SUFXA=od majke+SUFXB=rođen`

The change in 2nd part is recognized via grammar (see Figure 3.) either by using the variable $S or variables $SA and $SB. We can check that what is inside these variables agrees with the SUFX or SUFXA and SUFXB of the main entry placed inside the variable $F (for example: $SA=$F$SUFXA checks that whatever is in the variable $SA is equal to the SUFXA of the word found in variable $F).

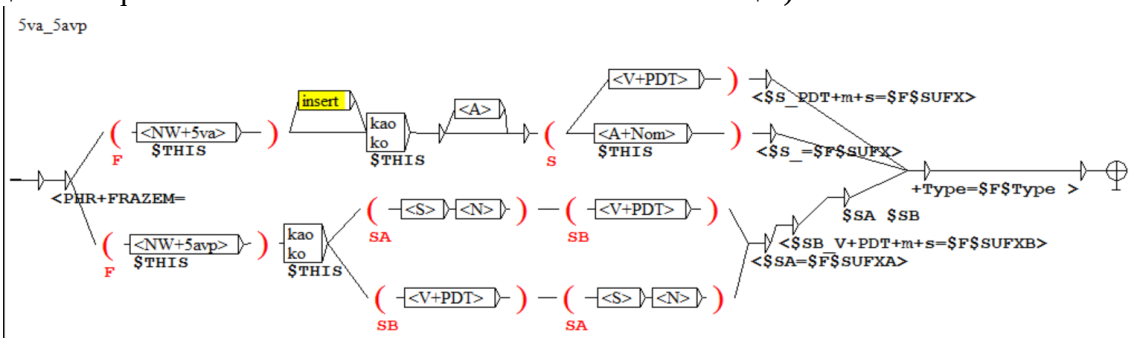

Figure 3: Grammar recognizing subtypes 5va and 5avp

Now, for the two last examples, we can recognize the following forms as well:

- *osjećao se kao preporođen* – en. he felt as if reborn
- *osjećala se kao preporođena* – en. she felt as if reborn
- *gol je kao od majke rođen* – en[lit]. he is naked as born by mother
- *gola je kao od majke rođena* – en[lit]. she is naked as born by mother

4.2. CI allowing poput

There are 368 dictionary entries that allow *poput*. Of that number, 48 have null 1st position (all change 2nd position occupied by a single noun or an adjective+noun or noun+prepositional phrase or coordination of two nouns) and remaining 320 have changeable 1st position (all change 2nd position) with 207 verbs and 113 adjectives in the 1st position.

In this category, regardless of the 1st position (null, adjective or verb), the 2nd position may be a single noun, a noun and a prepositional phrase, an adjective and a noun or coordination of two nouns both of which change in number and case (nominative or genitive).

There are no fixed 1st or fixed 2nd positions in this category with 10 subtypes coded 50n, 50an, 50np, 50n2, 5an, 5ann, 5ann, 5anp, 5vn and 5van. They are also grouped together regarding their similarities.

4.2.1. Types 50n, 50np and 50an

These are all subtypes of **null CI**s with changeable 2nd section that allow *poput.* Type 50n has only one word (noun), while the types 50np and 50an have more than one word segments. Their dictionary entries look like the following examples (*kao zmaj – en[lit.] like a dragon, kao grom iz vedra neba – en[lit.] like a thunder from a clear sky, kao otvorena knjiga – like an open book*):

- `zmaj,NW+Type=50n+FLX=KRALJ`
- `grom,NW+Type=50np+FLX=BAT+SUFXA=iz+SUFXB=vedra neba`
- `otvoren,NW+Type=50an+SUFX=knjiga+FLX=PRESPOR`

Here, as well, the change of the first word is recognized via the attribute FLX which was enough for the types 50n and 50np. The type 50an however, has a noun that also may change in case and number. In addition, in this category, the main word that usually comes before *kao* | *ko* | *poput* may also appear after it. We solved both these

problems via grammar (see Figure 4) where we allow for any noun (in nominative or genitive case) inside the variable $S, and then check if it matches the value of the SUFX attribute found in the dictionary.

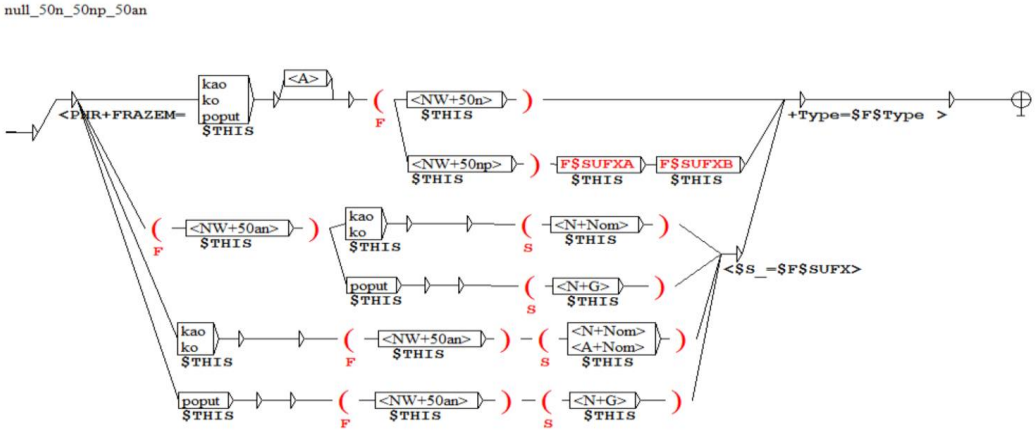

Figure 4: Grammar recognizing subypes 50n, 50np and 50an

4.2.2. Types 5an, 5aan, 5ann and 5anp

Types 5an, 5aan, 5ann and 5anp are all subtypes of and adjectival CI which means that they all have an adjective in the first position. In all of these cases, both the adjective in the first position and the word in the second position (adjective or a noun) change.

- `dug,NW+Type=5an+SUFX=vječnost+FLX=DUG`
- `ubog,NW+Type=5aan+SUFXA=crkveni+SUFXB=miš+FLX=PRESPOR`
- `jak,NW+Type=5ann+FLX=JAK+SUFXA=kraljević+SUFXB=Marko`
- `slobodan,NW+Type=5anp+SUFXA=ptica+SUFXB=na grani+FLX=DIVAN`

All the valid possibilities are defined with the grammar that recognizes:

- *dug kao vječnost, duga kao vječnost, dug poput vječnosti*
- *ubog kao crkveni miš, uboge kao crkveni miš, ubogi poput crkvenog miša*
- *jak kao kraljević Marko, jaki poput kraljevića Marka*
- *slobodna kao ptica nag rani, slobodni kao ptice na grani*

4.2.3. Type 5vn

Idioms with Type 5vn are **verbal CI**s that have a verb in 1st position and one noun in the 2nd position that may change. It is also valid that the verb moves to the last position or that the idiom is split between the verb and *kao | ko | poput.*

Thus, for the dictionary entry (*kretati se kao kornjača*):

- `kretati,NW+pov+Type=5vn+FLX=KRETATI+SUFX=kornjača`

the grammar recognizes the following examples:

- *kretao se kao kornjača, kretala se poput kornjače, kao kornjača se kretao…*

4.2.4. Type 5van

Comparative idioms of Type 5van are also all **verbal CI**s that have a verb in the 1st position and an adjective + noun in the 2nd position and all three words may change. To check if the right adjective and noun are used (even when they change case, number

an/or gender), both an adjective and a noun had to be nominalized and as such placed as SUFXA and SUFXB values.

- `planuti,NW+Type=5van+FLX=BLJESNUTI+SUFXA=živi+SUFXB=vatra`
- `razići,NW+Type=5van+FLX=DOĆI+pov+SUFXA=rakov+SUFXB=djeca`

It is also possible that the verb moves from the first to the last position of the expression or that, while in the first position, is interrupted with a noun phrase, prepositional phrase, reflexive pronoun or an auxiliary verb (defined by the node 'insert' in Figure 5).

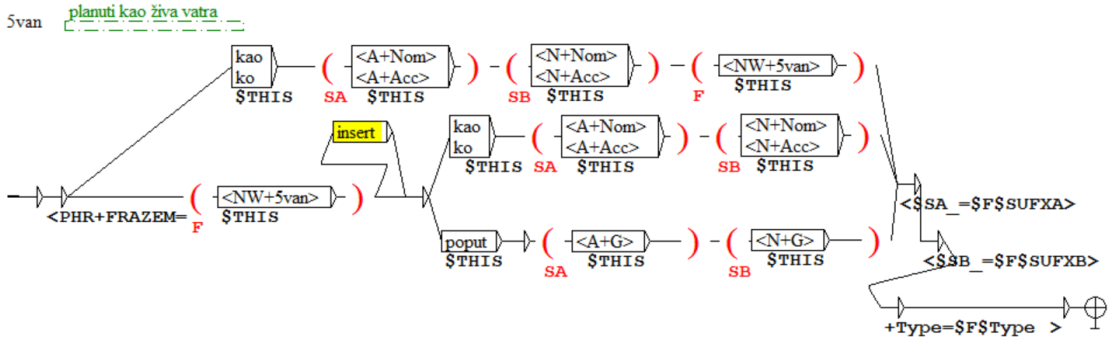

Figure 5: Grammar recognizing subtype 5van CIs

The grammar recognizes also:

- *planuti kao živa vatra, planula je poput žive vatre, kao živa vatra planuše*
- *razišli su se kao rakova djeca, razišle su se poput rakove djece*

5. RESULTS AND CONCLUSIONS

After training phase on a controlled corpus containing only the sentences with comparative idioms, we found both precision and recall to be 100%. Testing corpora (general-style corpus and corpus of literal works) gave us results of respectively 100% and 100% for precision, while we had 98% and 100% for recall. Table 1 shows all the results, as well as the f-measure.

	Kilo-words (Kw)	Number of structures found	Precision	Recall	F-measure
Training corpus	58 223 w	312	100%	100%	100%
Corpus 1 (web)	2247 Kw	103	100%	98%	98,9899%
Corpus 2 (books)	774 Kw	208	100%	100%	100%

Table 1. Results from the corpora

As it can be seen from the table above, these comparisons are not frequent in the corpora at all. The fact that these comparisons are generally less frequent in oral communication then in texts is supported by linguists (Fink Arnovski et al., 2006), but we were also wondering about the differences by style and purpose of the text. Menac (1978) lists several styles of phrasemes from stylistically neutral to vulgar, among which is a literary style characteristic for written forms of expression with 4 subtypes (literary and artistic, journalistic, scientific, business and administrative). Our examples of corpora belong to the first two subtypes and we have observed differences both in number and type of recognized phrasemes in these two corpora. Namely, in the web

generated general texts corpus, there was a frequency of 0.000045. On the contrary, in specialized literally texts corpus their frequency was 0.00026 or approximately six times more frequent.

We can conclude that this work covers almost all comparative idioms in Croatian language and successfully recognizes them, and can be applied for purposes of computer assisted translation, language learning, computational understanding of Croatian language and many other purposes as a language resource.

References

FINK ARNOVSKI, Ž., 2006. Višejezični rječnik poredbenih frazema, *Hrvatsko-slavenski rječnik poredbenih frazema*, Knjigra, Zagreb, p.439.

KOCIJAN, K. AND LIBRENJAK, S., 2015. The Quest for Croatian Idioms as Multi Word Units. To appear in J. Monti, R. Mitkov, G. Corpas Pastor and V. Seretan (eds.) In *Multiword Units in Machine TRanslation and Translation Technology*, John Benjamins Publishing. [in print]

LJUBEŠIĆ, N., DOBROVOLJC, K., KREK, S., PERŠURIĆ ANTONIĆ, M.& FIŠER, D., 2014. hrMWELex – A MWE lexicon of Croatian extracted from a parsed gigacorpus. In *Language technologies: Proceedings of the 17th International Multiconference Information Society IS2014*. Ljubljana, Slovenia. pp. 25-31.

MATEŠIĆ, J., 1978. O poredbenom frazemu u hrvatskom jeziku. In *Filologija 8.* Zagreb, pp. 211-217.

MATEŠIĆ, J., 1982. *Frazeološki rječnik hrvatskoga ili srpskog jezika*. Zagreb: Školska knjiga.

MENAC, A., 1978. Neka pitanja u vezi s klasifikacijom frazelogije. In *Filologija 8*, Zagreb, pp. 219–226.

MENAC, A., FINK-ARSOVSKI, Ž. AND VENTURIN, R., 2003. *Hrvatski frazeološki rječnik.* Zagreb: Naklada Ljevak.

MENAC-MIHALIĆ, M., 2007. Hrvatski Dijalektni Frazemi S Antroponimom Kao Sastavnicom. In *Folia Onomastica Croatica*, no. 12/13, pp. 361–85.

SILBERZTEIN, M. 2003. *NooJ Manual*. Available at < www.nooj4nlp.net> [Accessed July 2015]

VUČKOVIĆ, K., TADIĆ, M., BEKAVAC, B. 2012. Croatian Language Resources for NooJ. In: *CIT. Journal of computing and information technology 18*. pp. 295-301.