Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 60–67
Marseille, 20 June 2022

Resolving Inflectional Ambiguity of Macedonian Adjectives

Katerina Zdravkova
University Ss, Cyril and Methodius, Faculty of Computer Science and Engineering
Rudjer Boshkovikj, 16, 1000 Skopje, N. Macedonia
katerina,zdravkova@finki.ukim.mk

Abstract

Macedonian adjectives are inflected for gender, number, definiteness and degree, with in average 47.98 inflections per headword. The inflection paradigm of qualificative adjectives is even richer, embracing 56.27 morphophonemic alterations. Depending on the word they were derived from, more than 600 Macedonian adjectives have an identical headword and two different word forms for each grammatical category. While non-verbal adjectives alter the root before adding the inflectional suffixes, suffixes of verbal adjectives are added directly to the root. In parallel with the morphological differences, both types of adjectives have a different translation, depending on the category of the words they have been derived from. Nouns that collocate with these adjectives are mutually disjunctive, enabling the resolution of inflectional ambiguity. They are organised as a lexical taxonomy, created using hierarchical divisive clustering. If embedded in the future spell-checking applications, this taxonomy will significantly reduce the risk of forming incorrect inflections, which frequently occur in the daily news and more often in the advertisements and social media.

Keywords: inflectional ambiguity, lexical taxonomy, linguistic linked open data (LLOD), non-verbal and verbal adjectives

1. Introduction

Macedonian language as a South Slavic language is highly inflective (Bonchanoski and Zdravkova, 2018). Verbs have the richest inflectional paradigm that embraces seven tenses: present, past or aorist (depending on the verb aspect), past undetermined, pluperfect, future, past future, and future told); a conditional form; positive and negative imperative; and a construction with the particle neka (Cyrillic: нека / English: let it), each producing different forms for the three persons and the two numbers (http://vigna.mk/). Verbs have three aspects: progressive, perfective and bi-aspectual (Ljubešić et al, 2021a).

Verbal adjectives can be derived from all the verbs, independently of their aspect (Zdravkova and Petrovski, 2007). Whenever their root is identical with the root of a non-verbal adjective, they trigger the inflectional ambiguity, which is the main subject of this paper.

Inflectional paradigm of Macedonian adjectives is also rich, although unlike most Slavic languages, it does not use cases. Their function in the sentence is determined by the prepositions (Körtvélyessy, 2016). Adjectives are inflected for gender, number, definiteness and degree. In the lexicon MKLex that was annotated according to MULTEXT-East version 4 (Erjavec, 2017), Macedonian adjectives have in average 47.98 inflections per adjectival stem (Table 1.). The lexicon is available from CLASSLA CLARIN knowledge centre for South Slavic languages (https://www.clarin.si/info/k-centre/faq4macedonian/).

The inflection paradigm of qualificative adjectives is even richer, with 56.27 morphophonemic alterations. MKLex is not extended with the verbal adjectives, which introduce more than 30000 headwords (Zdravkova and Petrovski, 2007). Although frequently used, these adjectives are not entered in the official Macedonian dictionaries, which are the core sources of the Digital dictionary of Macedonian language (throughout the paper: DRMJ, http://drmj.eu).

MULTEXT-East version 6 introduced two categories: verbal adjectives, which are participles; and the category general that unites the adjectives such as: takov (таков / such), and gotov (готов / ready), which cannot be classified into any of the previously four mentioned groups. In the project described in this paper, all the adjectives are divided into two threads: verbal, i.e., participle adjectives and non-verbal, i.e., the adjectives belonging to remaining types.

Type	Headwords	All infections
Qualificative	7048	396591
Possessive	2172	65953
Ordinal	307	5200
Total	9749	467744

Table 1: Macedonian adjectives in MKLex

The inflectional base (Laudanna et al., 1992) of non-verbal adjectives is created by dropping the most right vowel before it gains the inflection suffixes (Table 2). When the headword ends with the consonant ~n (~н), which is preceded by two vowels, they are altered to ~jn (~јн). Dropping of the rightmost vowel of the adjectives ending with: ~dok (~док), ~zok (~зок) and ~zhok (~жок) causes morphonemic alterations: ~tk (~тк), ~sk (~ск) and shk (~шк). Exclusion are the endings: ~sten (~стен), which are transformed into ~sn (~сн), and ~on (~он), while the suffix remains unchanged, equally to all verbal adjectives.

Headword endings	Base endings	Headwords - base
~aen	~jn	traen – trajn
~ar	~r	dobar – dobr
~dok	~tk	redok – retk
~een	~jn	ideen - idejn
~ien	~jn	stihien – stihijn
~en	~n	temen - temn
~oen	~jn	bezboen - bezbojn
~ol	~l	topol – topl
~on	~on	avtohton - avtonton
~ov	~v	ednakov - ednakv
~uen	~jn	buen - bujn
~sten	~sn	mesten – mesn
~zhok	~shk	zhezhok – zheshk
~zok	~sk	blizok – blisk

Table 2: Alterations of non-verbal adjectives

	Masculine	Feminine	Neuter	Plural
No	/	~a	~o	~i
Yes	~iot	~ata	~oto	~ite
Distal	~iov	~ava	~ovo	~ive
Proximal	~ion	~ana	~ono	~ine

Table 3: Inflectional suffixes of Macedonian adjectives

The inflections are formed by adding the suffixes to the inflectional base (Table 3). The columns present the suffixes for the three genders in singular and the plural, which are identical for all genders. Rows correspond to definiteness. Similarly to nouns and pronouns, definiteness is expressed by the three suffixed articles: undetermined (yes), distal, and proximal. Distal and proximal definiteness are language specific and they do not exist in other Slavic languages (Stojanovska, 2019).

Many non-verbal adjectives are derived from nouns, such as: boen (боен), which is derived from the noun boj (бој / battle) and the verb boi (бои / to colour); vozen (возен), derived from the noun voz (воз / train) and the verb vozi (вози / to drive); soboren (соборен), derived from the noun sobor (собор / gathering, feast), and the verb sobori (собори / to knock down, shoot down, demolish). Their stems are identical: boen (боен), vozen (возен), soboren (соборен), but the inflections for the same morpho-syntactic description and the translations are different.

In the daily news and yet more often in the advertisements, the inflections of non-verbal and verbal adjectives are usually mixed. Of these two options, the former occurs because the spell-checking applications do not recognise them as incorrect, while the latter is usually due to illiteracy of people. For example, masculine, singular, definite and positive form of the adjective boen (боен) is either bojniot (бојниот / military, battle) or boeniot (боениот / painted, coloured, stained, dyed), which, if wrongly used, produce the collocations: bojniot dzid (бојниот ѕид / the military wall), instead of boeniot dzid (боениот ѕид / the painted wall), and boeniot otrov (боениот отров / the dyed poison), instead of bojniot otrov (бојниот отров / the military poison). The phrase boeniot otrov (боениот отров / the dyed poison) exists 6 times on the Web, compared to 282 correct collocations. Even more frequent is the collocation soboreniot hram (собореният храм / the overthrown temple), that apears 125 times instead of the correct soborniot hram (соборниот храм / the cathedral temple), which occurs more than 100000 times. Google Translate translates both: the incorrect soborniot avion (соборниот авион) and the correct soboreniot avion (собореният авион) as the downed plane. The translations for soboreniot hram and soborniot hram are: the overthrown temple and the cathedral, confirming that Google Translate recognises both forms of the adjective soboren (соборен).

Depending on the word they were derived from, more than 600 adjectives have an identical stem and two word forms for each grammatical category. In parallel with the morphological differences, both types of adjectives have different translations, depending on the category of the word they have been derived from. They are the subject matter of the research presented in this paper.

The paper proposes a solution intended to resolve inflectional ambiguity of non-verbal and verbal adjectives with the same headword and different meanings. Section 2 presents several examples of inflectional ambiguities and the proposed solutions for their disambiguation. Particular attention is paid to lexical taxonomies, which are the proposed approach for the resolution of inflectional ambiguity. Section 3 introduces the process of extracting the adjectives with inflectional ambiguity, as well as the hierarchical classifiers that enable the disambiguation. Section 4 is dedicated to created taxonomy. Section 5 summarises the introduced approach and announces further extensions and practical use of the project.

2. Inflectional ambiguity and lexical taxonomies

Inflectional ambiguity is not uniquely defined. Branco and Nunes (2012) introduce it at two independent layers, according to different substrings that qualify a given word form, as well as according to admissible affixes, the latter conveying more than one admissible value. The main goal of their project was disambiguation of Portuguese verbs, which can have identical third person with the infinitive verb form; identical forms for first and third person; as well as inconsistency between the inflected infinitive and the subjunctive future. All the three experiments implemented the machine learning based MFF algorithm supported by a verbal lemmatization tool (Branco and Nunes, 2012).

The inflection ambiguity of the Finish language encompasses two aspects: ambiguity of words with two decomposable readings and ambiguity due to homographic stem allomorphs (Järvikivi et al., 2009). The disambiguation is based on early segmentation of inflected words. Both aspects achieved similar results for unambiguous, partly or completely ambiguous inflected forms (Järvikivi et al., 2009).

Third explanation of inflectional ambiguity relates to the possibility of implementing several conjugation rules for the verbs in Arabic (Ismail et al, 2017). More detailed explanation and the disambiguation process are not presented in the paper.

The first two examples of inflectional ambiguity are not similar to ambiguity of Macedonian adjectives. They include lemmatisation (removing inflectional endings to return the lemma (Schütze, 2008)), or word recognition (selection of the correct lexical representation from a set of candidates (Segui and Grainger, 1990)). Our approach is related to morphological synthesis, i.e. determination of inflected word forms (Bickel and Nichols, 2005).

No matter the target result: headword or word form, the disambiguation is heavily dependent on the available contextual information. For the resolution of inflection ambiguity of Macedonian adjectives, such contextual information can be extracted from the adjective-noun collocations. The nouns collocating with the non-verbal and verbal adjectives are mutually disjunctive, defining the two branches of the hierarchical taxonomy that entirely resolve the ambiguity.

The first association of successful lexical taxonomies is WordNet (Miller, 1998). Nouns within WordNet are hierarchically organised by connecting the hyponyms are hypernyms via is-a relationship. Knowledge structure is convenient for resolving the inflectional ambiguity. Although ambitiously announced (Saveski and Trajkovski. 2010), the Macedonian language is still not included in WordNet, and even if it was, this semantically organised lexical database is far too massive for the problem. Nevertheless, it remains the greatest inspiration for the creation of our hierarchical taxonomy.

Another valuable lexical taxonomy was proposed by Burtăverde and De Raad (2019). This hierarchical structure was obtained by splitting the personality-descriptive Romanian adjectives using different levels of abstraction.

Taxonomy enrichment of Russian language has recently been efficiently done (Nikishina et al, 2020). Based on the defined set of potential hypernyms, this project had an intention to correctly classify new words that do not have any definition.

The capacity to efficiently cluster the words of the above mentioned projects was the major inspiration for the disambiguation of Macedonian adjectives. It was broken down into seven phases:

- Extraction of candidate non-verbal and verbal adjectives with different inflections
- Elimination of all the candidate adjectives that are not frequently used
- Elimination of the candidate adjectives that do not have full collocations for both types
- Determination of all the nouns belonging to mutually disjunctive sets of collocations
- Hierarchical classification of the extracted nouns
- Creation of the lexical taxonomy
- Labelling of the adjectives

They will be explained in more detail in the next section.

3. Disambiguation process

Candidate adjectives were extracted from non-commercial version of Macedonian lexicon MKLex, which can be downloaded from the CLASSLA CLARIN.SI repository (http://www.clarin.si/info/k-centre/). It consists of roughly 76000 headwords and more than 1300000 word forms presented as tab-separated triples: word form, headword, and annotation according to MULTEXT-East version 4. Since the development of MKLex, MSDs were upgraded and new dictionaries were published, unfortunately, none is available in a machine readable form. Therefore, the extraction was done by following these steps:

- Extraction of the pairs with two forms for definite masculine singular (358 pairs or 179 headwords);
- Extraction of the adjectives that have an identical or a similar root with the verbs from the lexicon (182 headwords);
- Extraction of the adjectives that have an identical or a similar root with the nouns from the lexicon (6 headwords);
- Extraction of nouns and verbs with an identical or similar root (420 headwords);
- Addition of the eligible adjectives that do not exist in the lexicon (17 headwords)
- Union of the five sets: in total, 634 headwords.

The most valuable resource for the next two steps was the digital dictionary DRMJ (http://drmj.edu). It presents all the words existing in the printed dictionaries, including many new words that were found in the dictionary embedded corpus of Macedonian literature. Each headword is accompanied with a ranking. The higher the value of the rank is, the lower is the frequency of word's occurrence in the embedded corpus. The most important feature of the digital dictionary is its linked structure. Namely, each headword is connected with a list of sentences from the corpus where it occurs in all the feasible word forms.

The creation of lexical taxonomy started with the pre-processing of candidate adjectives. It included two eliminations, followed by the creation of the taxonomy skeleton. The procedure included the following steps:

- Exclusion of the least frequent adjectives
- Exclusion of the adjectives without adjective–noun collocations for both threads
- Extraction of the adjective-noun collocations from DRMJ embedded corpus
- Creation of mini taxonomies for each adjective
- Joining mini taxonomies into a final taxonomy

First elimination was done by examining the ranking of the candidates. The adjectives with a rank above 35000 were removed. For the adjectives that were not found in DRMJ, the ranking of the noun or the verb they are derived from was also checked. High rankings did not necessarily mean that the adjective would not be included in the taxonomy. For example, the adjective broen (броен / non-verbal: number, numerical, numerous; verbal: counted, numbered) has a rank 1781 and almost no collocations for the verbal adjective: broeniot denar (брoениот денар / counted pennies), broeno kolichestvo (броено количество / counted quantity), broeni denovi (броени денови / numbered days, and broeni pari (броени пари / counted money). Conversely, the adjective vklopen (вклопен / non-verbal: timed, time switch; verbal: blended, embedded, assembled) with a rank 32007 is completely populated for both threads.

For each of the remaining adjectives, DRMJ embedded corpus was manually checked to extract the adjective-noun collocations for each gender and number. The adjectives without at least three out of the four possible combinations (masculine, feminine, neuter, plural) for both threads were excluded. The absence of corresponding collocations does not necessarily mean that they do not exist. Missing examples of common adjectives were searched on the Web. For example, the adjective loven (ловен / non-verbal: hunting; verbal: hunted) has only two non-verbal collocations in DRMJ corpus: lovniot trofej (ловниот трофеј / the hunting trophy) and lovna oprema (ловна опрема / hunting equipment), and no collocations for the verbal adjective. On the other hand, they are very frequent on the Web: lovno drushtvo (ловно друштво / hunting union), loveniot zajak (ловениот зајак / the hunted rabbit), lovena mechka (ловена мечка / hunted bear), lovenoto zhivotno (ловеното животно / the hunted animals), and loveni srni (ловени срни / hunted dears).

Since DRMJ is not available for crawling, interlinking of the adjectives and the corpus was manually done for each of the 634 candidate adjectives. They were stored in a large spreadsheet consisting of five columns. First column presents the headword of the adjective, the second and third correspond to all possible occurrences of nominal and verbal adjectives respectively, whereas the fourth column presents the headword of verbal adjective derived from the perfective pair of the headword with the same meaning, and the last column presents its occurrence. All the values were extracted from DRMJ. Although these five nominal adjectives are not frequent, they were kept for further processing, because the nouns in adjective-noun collocations of both verbal adjectives are identical. After these two stages, 96 headwords remained in the corpus (Table 4). Most of them are simultaneously nominal and verbal, and make their inflections implementing the rules presented in the Tables 2 and 3.

Derived from	Nouns and verbs	Verbs	Nouns
In total	60	34	2

Table 4: Distribution of ambiguous adjectives

Several ambiguous verbal adjectives are derived from a negated verb, which is formed by adding the particle ne (не / not) to the main verb. In the verbal adjectives, this particle becomes is transformed into a prefix: nezasiten (не засити → незаситен / greedy); and neodgovoren (не одговори → неодговорен / irresponsible).

The extraction process was done in parallel with the second elimination. All the possible adjective-noun collocations existing on the interpretation part of DRMJ and the existing collocations from the embedded corpus were stored in a spreadsheet together with their English translations and the superordinate and subordinate categories the corresponding nouns belong to. These categories are a combination of suggested categories within DRMJ, WordNet Domains Hierarchy suggested by Bentivogli et al. (2004) and English WordNet (Fellbaum, 2005), which is accessible from http://wordnetweb.princeton.edu/perl/webwn.

After the extraction, only 96 eligible adjectives remained in the corpus (see Appendix 1). They are presented with Macedonian Cyrillic script, their Latin transliteration, the ranking according to DRMJ, and the corresponding English translations of nonverbal and verbal adjective.

4. Creation of lexical taxonomy

The creation of mini taxonomies for each adjective is divisive (or top-down), starting from the root node toward the leaf nodes (Roux, 2018). The organisation of noun categories is hierarchical, with a root node divided into superordinate nodes, each a parent of at least one node at subordinate layer. Subordinate categories become superordinates if they can further be divided into subtler categories. Initially, the idea was to create the lexical taxonomy by combining the extracted clusters, but the creation of mini taxonomies and their merging was more convenient. It started with the adjective vlezen (влезен / non-verbal: entry, input; verbal: entered), which is alphabetically the first adjective that determines at least 10 distinct categories. It is the seed lexical taxonomy. Collocation nouns of verbal adjective vlezen (влезен / entered), in which the inflectional base is identical with the headword are Living beings, which are direct descendants of the root, so they belong to layer B (Figure 1).

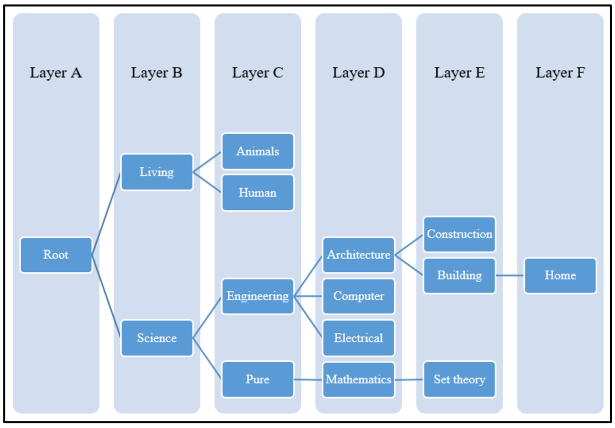

Figure 1: Seed lexical tree

Each layer is marked with a letter, staring with the root node. The nodes within one layer are marked with integers, starting from the topmost node, which is 1. Each noun belonging to one cluster is labelled with the pair: letter of the layer, and number of the node within the layer. Living beings are divided into two clusters: Animal, which unites animal species (cats, dogs, horses, etc.) and Human (boys, girls, men, women, etc.). These clusters are terminal, so they are nodes belonging to leaf categories: C1 for animals, C2 for human. If some adjectives collocate with a specific part of the terminal cluster, it can further be divided.

The nouns collocating with the nominal adjective vlezen (влезен / entry, input), in which the inflectional base is formed by dropping the rightmost vowel are disjunctive with the clusters Animal and Human. They are part of the superordinate category Science. Science is divided into two subordinate layers: Engineering and Pure science. Engineering is a superordinate category for Architecture, Computer and Electrical Engineering. Architecture is further divided into Constructions, where the nouns from the first collocation of vlezna porta (влезна порта / entry gate) belong. The second embraces the nouns related to Home, where vlezna porta (влезна порта / entry doorway) belongs. Collocations: vlezna porta (влезна порта / input port) and vlezni uredi (влезни уреди / input devices) are terms belonging to Computer engineering, while vlezniot prikluchok (влезниот приклучок / the input switch) is related to Electrical engineering. Pure sciences have one subordinate category: Mathematics, where the nouns such as vlezno mnozhestvo (влезно множество / input set) belong. To distinguish Set theory from other mathematical branches, Mathematics can further be extended.

In the lexical tree of adjective vlezen, all the labels belong to the leaf nodes. The leaves from the layer C (C1 for animals, C2 for human) collocate with the verbal adjective. The nouns of the leaves from layer D: D2 for Computer engineering, D3 for Electrical engineering and D4 for Mathematics collocate with the nominal adjective, together with the nouns from layer E: E1 for Constructions and E2 for Home. This is the initial stage of the lexical taxonomy. The same strategy was implemented for all the adjectives in the corpus. The taxonomy creation continues according to the algorithm presented on Figure 2.

```
Lexical taxonomy creation:
While list of adjectives is not empty
    Get the lexical taxonomy of the new adjective
    Merge the taxonomy of new adjective with the existing
    Remove the adjective from the list
endwhile

Merge:
If superordinate category exists
    While list of subordinates is not empty
        Get the name of the top subordinate category
        If new name is alphabetically prior to existing one
            While list of subordinates is not empty
                Increment the numerical value of the labels
                Go to next layer
            endwhile
            Go to next subordinate
        endif
        Go to next subordinate
    endwhile
endif
```

Figure 2. Pseudocode of taxonomy creation

Merging of the seed lexical taxonomy with the mini taxonomy of the adjective boen (боен / non-verbal: military; verbal: coloured, dyed, painted, stained) starts with the nodes from layer B. The new superordinate category Objects, which is a parent of Physical objects is alphabetically between Living organisms (node B1), and Sciences (before the addition, node B2). After adding the superordinate category Objects, the numerical value of the label for Science is incremented: Science (node B3).

The new category B2 has its own children node in the layer C: Physical objects. This node becomes the new C3 node. This addition causes an increase of the numerical values of all the nodes after the first child of the former category B2: Engineering (node C4) and Pure (node C5). Since the node C3 (Physical objects) has no children, the procedure continues with the new Science, which is a node at layer C.

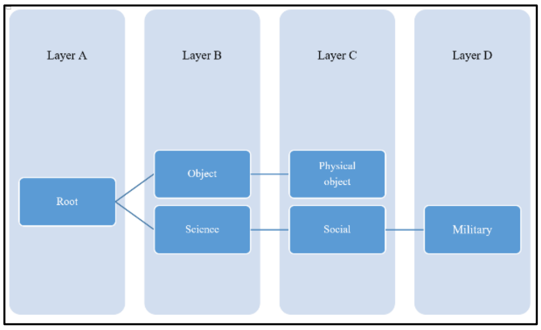

Figure 3: Mini taxonomy of adjective боен (boen)

Non-verbal adjective-noun collocations define the new scientific subordinate category Social as a child category of Science (Bentivogli et al., 2004). Its subordinate category is Military (Figure 4.). Both new nodes do not affect the previous alphabetic ordering of sciences, thus the node will be labelled as C6: Social and its subordinate nodes continue the previous numbering at all the descendant layers, in this case only the new subordinate at layer D, which becomes D5: Military.

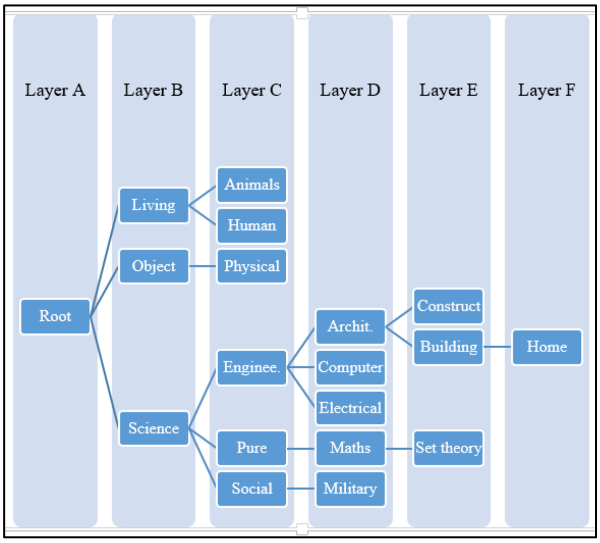

Figure 4: Lexical taxonomy after adding the first adjective

By continuing the procedure, the final lexical taxonomy was created. It embraces 138 meanings (and consequently, English translations) of the adjectives belonging to non-verbal thread, and 118 meanings from verbal thread. All the meanings from both threads are mutually disjunctive, proving that the division was worthwhile.

4.1 Nouns that collocate with non-verbal adjectives

Non-verbal thread introduced these 105 clusters: abstract, actions, activity, administration, analysis, anatomy, animal, approach, architecture, authorization, beauty, birds, body, bomb, character, chemistry, church, civil engineering, computing, consequences, construction, data, dermatology, disease, document, documentation, economy, effort, electrical engineering, emotion, emotion, event, exam, examples, finance, finance, food, furniture, gastronomy, genetics, geometry, goods, grammar, house, human, image, institution, instrument, justice, law, letter, line, material, mathematics, measures, medicine, military, music, nature, paper made, part, path, payment, person, pet, philosophy, physical object, place, plan, post, price, profession, religion, reply, results, river, road, scene, science, season, senses, sentence, signs, smooth material, soil, solution, speech, states, technology, temperature, text, theory, thing, time, topology, traffic, transport, transport means, travel, view, water, weapon, weather, words, and work.

4.2 Nouns that collocate with verbal adjectives

The verbal thread that unites participle adjectives introduce 58 clusters, almost half of clusters for the non-verbal thread, mainly because the agents are either human beings or animals. They are: abstract, animal, article, chemistry, clothes, company, construction, drink, duty, economy, effort, facts, finance, flammable, food, garden, gastronomy, goods, group of people, human, image, industry, inflammable, inheritance, justice, law, life, living organism, lock, medicine, money, movement, object, obligation, part of animal, part of body, people, philosophy, physical object, price, profession, property, quantity, question, religion, senses, shoes, sound, space, task, technology, territory, thing, vegetables, vehicle, wire, words, and yarn.

While the adjectives belonging to both threads are disjunctive, the clusters of nouns they collocate with them intersect. In total, both threads define 137 clusters, 27 belonging to both: abstract, animal, chemistry, construction, economy, effort, emotion, finance, finance, food, gastronomy, goods, human, image, justice, law, medicine, part of body, philosophy, physical object, price, profession, religion, senses, technology, thing, and words. The maximum depth of the taxonomy is 7, and it was reached by the clusters related to both threads.

5. Conclusions and further work

Macedonian adjectives are specific due to their inflectional ambiguity. Depending on their etymology and derivation, they have two inflectional bases. The inflectional base of verbal adjectives coincides with the headword, while non-verbal adjectives are morphonologically altered. Unfortunately, these simple rules are not obeyed by online published news, and particularly not in advertisements and social media.

Main resource for extraction of inflectionally ambiguous adjectives was MKLex, a lexicon that was created more than 15 years ago with NooJ (Silberztein, 2005). MKLex was annotated with MULTEXT-East version 4 classifying the ambiguous adjectives as qualificative, although most of them are also participles.

Within the pilot project presented in this paper, a new approach for their disambiguation has been proposed. It suggests a division of all the adjectives into two different threads depending on the inflectional base. The first thread embraces the non-verbal adjectives, and the second are those that are derived from verbs. Although obvious, such distinction is not made in the new dictionary of Macedonian language, but it can be found in the digital dictionary, which explicitly points to the verb the adjective was derived from.

So far, lexical taxonomy has not been practically evaluated. It was manually checked with several incorrect inflections on the Web, and the intersection of adjective-noun collocations with wrong adjective inflection and adjective-noun collocation from the taxonomy was always empty, proving the correctness of the approach. This optimistic finding is the main motivation for further work.

Recently, two valuable text collections of Macedonian language have been released: comparable corpus collection consisting of Wikipedia dumps that were crawled in 2020 (Ljubešić et al, 2021b) and Macedonian web corpus MaCoCu-mk 1.0, which was built by dynamic crawling of ".mk" and ".мкд" internet top-level domains in 2021 (Bañón et al., 2022). These large corpora will be exhaustively researched in the following several months, in order to examine the availability of the selected inflectionally ambiguous adjectives and their collocations, to examine whether the adjectives that lacked adjective-noun collocations for some genders can be added to the existing collection and to discover new adjectives that were not discovered so far.

Unlike DRMJ corpus, these two large corpora are publicly available, so they will enable semiautomatic and automatic processing of available collocations, and well as successful evaluation of the created lexical taxonomy.

If powered with the forthcoming Macedonian WordNet, it will permanently resolve the inflectional ambiguity due to different etymology, different derivational morphological rules and different semantic properties of those adjectives that should be presented with two headwords, one belonging to qualificative, the second to participle adjectives.

Once created and accepted, this lexical taxonomy will facilitate the correct inflection of ambiguous adjectives in the online published news that are not proofread. It can also become a valuable resource for foreign language learners, because collocations are crucial for acquiring native-like fluency (Basal, 2019).

6. Acknowledgements

This paper is based upon work from action CA18209: European network for Web-centred linguistic data science supported by COST (European Cooperation in Science and Technology). It was partially financed by the Faculty of Computer Science and Engineering.

7. Bibliographical References

Bañón, M. et al. (2022). Macedonian web corpus MaCoCu-mk 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1512

Basal, A. (2019). Learning collocations: Effects of online tools on teaching English adjective-noun collocations. *British Journal of Educational Technology,* 50(1), pp. 342-356.

Bentivogli, L., et al. (2004). Revising the wordnet domains hierarchy: semantics, coverage and balancing. *Proceedings of the workshop on multilingual linguistic resources*, pp. 94-101.

Bickel, B., and Nichols, J. (2005). Inflectional synthesis of the verb. *The world atlas of language structures*, pp. 94-97.

Bonchanoski, M., and Zdravkova, K. (2018). Learning syntactic tagging of Macedonian language. *Computer Science and Information Systems*, 15(3), pp. 799-820.

Branco, A., and Nunes, F. (2012). Verb analysis in a highly inflective language with an MFF algorithm. In *International Conference on Computational Processing of the Portuguese Language*, pp. 1-11.

Burtăverde, V., and De Raad, B. (2019). Taxonomy and structure of the Romanian personality lexicon. *International Journal of Psychology*, 54(3), 377-387.

Erjavec, T. (2017). MULTEXT-East. In *Handbook of Linguistic Annotation*, pp. 441-462

Fellbaum, C. (2005). WordNet and wordnets. In: *Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics*, Oxford: Elsevier, pp. 665-670.

Ismail, S., Maraoui, H., Haddar, K., and Romary, L. (2017). ALIF editor for generating Arabic normalized lexicons. In *2017 8th International Conference on Information and Communication Systems (ICICS)*, pp. 70-75.

Järvikivi, J., Pyykkönen, P., and Niemi, J. (2009). Exploiting degrees of inflectional ambiguity: Stem form and the time course of morphological processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1), 221

Körtvélyessy, L. (2016). Word-formation in Slavic languages. Poznan Studies in Contemporary Linguistics, 52(3), pp. 455-501.

Laudanna, A., Badecker, W., and Caramazza, A. (1992). Processing inflectional and derivational morphology. *Journal of Memory and Language*, 31(3), pp 333-348.

Ljubešić, N., et al. (2021a). The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Macedonian 1.1.

Ljubešić, N., et al., (2021b), Comparable corpora of South-Slavic Wikipedias CLASSLA-Wikipedia 1.0, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1427.

Miller, G. (1995). WordNet: A Lexical Database for English, *Communications of the ACM* 38(11): pp. 39-41.

Nikishina, I., Logacheva, V., Panchenko, A. and Loukachevitch, N. (2020). *RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian language*. arXiv preprint arXiv:2005.11176.

Roux, M. (2018). A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, *35*(2), pp. 345-366.

Saveski, M., and Trajkovski, I. (2010). Automatic construction of wordnets by using machine translation and language modeling. In 13th *Multiconference Information Society*, Ljubljana, Slovenia.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). Introduction to information retrieval (Vol. 39), pp. 234-265. Cambridge: Cambridge University Press.

Segui, J., and Grainger, J. (1990). Priming word recognition with orthographic neighbors: Effects of relative prime-target frequency. *Journal of Experimental Psychology: Human Perception and Performance,* 16(1), 65.

Silberztein, M. (2005). NooJ: a linguistic annotation system for corpus processing. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pp. 10-11.

Stojanovska, B. (2019). The Definite Article in the Macedonian Language. DEIXIS IN LANGUAGE, 22.

Zdravkova, K., and Petrovski, A. (2007). Derivation of Macedonian verbal adjectives. In *Proceedings of international conference recent advances in natural language processing" (RANLP'07)*, pp. 661-665.

Appendix: Inflectionally ambiguous adjectives (part 1)

Cyrilic	Latin	Ranking	Translation of nonverbal adjective	Translation of verbal adjective
боен	boen	4932	military, battle	painted, coloured, stained, dyed
броен	broen	1781	number, numerical, numerous	counted, numbered
буден	buden	2000	awake, watchful, vigilant	awaken
варен	varen	6124	limestone, lime	boiled
вграден	vgraden	7091	built-in	implanted, embedded, installed
верен	veren	1512	faithful, devoted	engaged
виден	viden	28385	prominent, noticeable, visible	seen
вклопен	vklopen	32007	timed, time switch	blended, embedded, assembled
вкусен	vkusen	7570	tasty, tasteful, delicious	tasted
влезен	vlezen	188	entry, input	entered
возен	vozen	9779	driving	driven
вратен	vraten	6399	neck	returned, repaid
гасен	gasen	35739	gas	extinguished
гледан	gledan	6653	view	groomed, viewed
горен	goren	1624	upper, higher	burned
граден	graden	5720	chest	built
грешен	greshen	5622	sinful	erroneous
димен	dimen	27006	smoke, smoking	smoked
договорен	dogovoren	6274	contractual	agreed
дрвен	drven	990	wood	wooden
дробен	droben	9429	tiny, small, little	minced, chopped
забавен	zabaven	4640	entertaining	slow
заборавен	zavoraven	2689	forgetful	forgotten
завршен	zavrshen	1678	final	completed
заглавен	zaglaven	13856	initial	stuck
задоволен	zadovolen	1247	content	fulfilled
задушен	zadushen	11557	part of memorial service	stuffy, silenced
заклучен	zakluchen	8166	final	locked
заложен	zalozhen	32151	security	pawned, pledged
залуден	zaluden	9075	fruitless, futile, vain, wasted	insane, mad, spoony
занесен	zanesen	3790	exhilarating, enchanting	absent-minded
заобиколен	zaobikolen	12033	detour	indirect, surrounded
запален	zapalen	2327	combustible	inflamed
заразен	zarazen	18869	infectious, catching	infected
заслужен	zasluzhen	6199	deserving	justified
заштитен	zashtiten	3602	protective	protected
земен	zemen	8716	earthy	taken
извршен	izvrshen	2889	executive, effective	executed, finished
излезен	izlezen	169	exiting	output
искусен	iskusen	5242	experienced, skilful	tried
исправен	ispraven	4747	correct	upright, straight
исцрпен	iscrpen	7610	exhaustive, comprehensive	exhausted
јаден	jaden	20211	pitiable, angry	eaten
книжен	knizhen	10571	paper	registered
кожен	kozhen	4073	skin	leather
ладен	laden	2586	cool, cold	cooled, chilled
ловен	loven	34882	hunting	hunted
матен	maten	2578	obscure, unclear, dull	stirred
мачен	machen	2481	difficult, suffering	forced, tormented
набавен	nabaven	35836	purchase	purchased
нагазен	nagazen	23924	stepping	stepped
нагорен	nagoren	13771	rising, steep, upward	burned

Appendix: Inflectionally ambiguous adjectives (continued)

Cyrilic	Latin	Ranking	Translation of nonverbal adjective	Translation of verbal adjective
нареден	nareden	1659	next	arranged, ordered, lined up
наследен	nasleden	8182	heritance, hereditary	inherited
научен	nauchen	1365	scientific	learned
нацртан	nacrtan	2995	descriptive	drawn, made up
незаситен	nezasiten	27589	greedy	unsaturated
неизмерен	neizmeren	14721	immeasurable	unmeasured
неодговорен	neodgovoren	17393	irresponsible	unanswered
неуреден	neureden	22040	untidy	disorderly, messy
носен	nosen	22247	nasal	worn
обиколен	obikolen	23361	bypass	bypassed, surrounded
одговорен	odgovoren	2189	responsible	answered
одделен	oddelen	1963	separate, different, individual	separated
отсечен	otsechen	11882	decisive	cut off
пазарен	pazaren	4969	market	negotiated, purchased
платен	platen	4942	salary, buying	paid
погоден	pogoden	2807	suitable	hit, affected, agreed
погубен	poguben	11778	deadly	killed, executed
поздравен	pozdraven	29809	welcoming	welcomed
поправен	popraven	11868	correctional	corrected
попречен	poprechen	14116	crosswise	disabled
поразен	porazen	8529	disastrous, devastating	defeated
потврден	potvrden	11257	confirmed	proven
потресен	potresen	6958	shocking	shocked, worried
потрошен	potroshen	10454	expendable	spent
пофален	pofalen	18247	lauding, commendable	praised
правен	praven	1229	legal	made, prepared, completed
преден	preden	2208	frontal	spun, spinning
преселен	preselen	17541	migratory	moved, relocated
пресечен	presechen	9173	intersection	cut off
пријавен	prijaven	35015	reported	registered
присвоен	prosvoen	26513	possessive	seized
речен	prechen	10142	river	said
роден	roden	769	fruitful, native	born, talented
следен	sleden	431	next, following	pursued, stalked
сложен	slozhen	1407	united	complex
соборен	soboren	15432	cathedral	overthrown
составен	sostaven	2158	composite, compound	joined, composed
среден	sreden	1437	middle, medium, average	ordered
товарен	tovaren	7030	transport	loaded
точен	tochen	3288	accurate, correct	draft, pour
украсен	ukrasen	6653	decorative	decorated
употребен	upotreben	7617	practiced	used
уреден	ureden	5775	tidy, orderly	arranged