

Quantitative Characteristics of the Contemporary Macedonian Language

Квантитативните карактеристики на современиот македонски јазик

Abstract: DEREBEY, Sinasi. *Quantitative Characteristics of the Contemporary Macedonian Language*. How many English words does the Macedonian language contain? What is the ratio between verbs and nouns? Are there more masculine or feminine nouns? Which areas of knowledge are covered best? Those and many similar questions can be answered using the Digital Dictionary of the Macedonian Language. The quantitative characteristics are the fingerprints revealing how a language supports the communication and exchange of knowledge. This information can be useful when translating from or into Macedonian.

Key words: *Contemporary Macedonian Language, Statistics, Internet Dictionary, www.makedonski.info, www.makedonisch.info*

Апстракт: ДЕРЕБЕЈ, Шинаси. *Квантитативните карактеристики на современиот македонски јазик*. Во текстот се настојува да се одговори на некои поважни прашања за квантитативните карактеристики на современиот македонски јазик, а одговорите на тие прашања, фактички, можат да се најдат и во Дигиталниот речник на македонскиот јазик кој е изработен од авторот на трудот. Статистичките податоци и добиените информации се несомнено драгоцен и мошне корисни при преведувањето од / на македонски јазик.

Клучни зборови: *Современ македонски јазик, статистика, дигитален речник, www.makedonski.info, www.makedonisch.info*

1. Introduction

The Digital Dictionary of the Macedonian Language (DRMJ) was created by the need of a good, modern Internet German-Macedonian dictionary to facilitate the integration of the Macedonian immigrants in Germany. Shortly after starting the work it became apparent that a good monolingual Internet Macedonian dictionary is a prerequisite for bilingual dictionaries. Currently there are only few printed authoritative dictionaries available: the dictionary by B. Koneski published in 1961-1966¹, the dictionary of the Institute for Macedonian Language (Tolkovnik, IMJ, I-V, 2003-2011) which is still incomplete², and the spelling dictionary by K. Koneski (1999)³.

At the beginning those printed dictionaries were used to create a data base (DB) of Macedonian words which later became DRMJ⁴. The DRMJ was integrated into the bilingual German-Macedonian dictionary, but both dictionaries are also separately usable.

The structure of the DRMJ has been designed according linguistic principles and comprises grammatical, semantic and pragmatic categories. The base unit is a lexeme consisting of one or more sememes. Each

lexeme is unique within the DB. It is distinguished by its grammatical category. A lexeme can belong only to one single grammatical category. This means that the DB is dealing with homonyms on the level of sememes.

Each sememe has a description, examples for its use, synonyms, antonyms, hyponyms and a hyperonym. It also has translations into English and Albanian. The English translation is used as ontology ("shared concepts")⁵. The Albanian translation was implemented to facilitate the communication between the Macedonian and Albanian community in Macedonia. The work on the Albanian translations is still going on and currently only about 30% of the lexemes have been translated.

One of the specifics of the Macedonian language is the considerable number of words coming from the Turkish language. They are remnants from the 500 years old Ottoman rule in the Balkans. The identification of those words was done manually using the Turkish Electronic Dictionary (TDK)⁶.

The semantic structure maps a semantic tree which has been defined arbitrarily, but is flexible and can be easily changed. It divides the sememes into a number of flexible

¹ КОНЕСКИ, Блаже: *Речник на македонскиот јазик I-III*. Скопје 1994.

² ВЕЛКОВСКА, Снежана et al. (eds.). *Толковен речник I-V*. Скопје 2003 – 2011.

³ КОНЕСКИ, Кирил. *Правописен речник на македонскиот литературен јазик*. Скопје 1999.

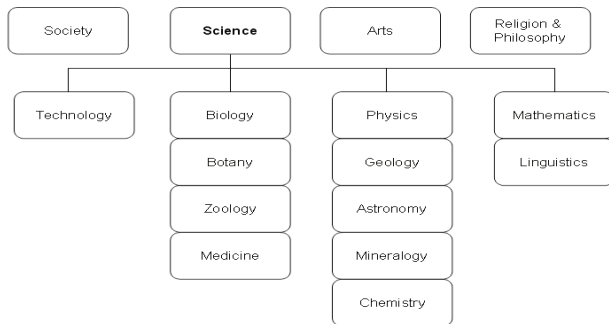
⁴ Дигитален речник на македонскиот јазик [online]. Available at: <www.makedonski.info>

⁵ HEYER, Gerhard – QUASTHOFF, Uwe – WITTIG, Thomas. *Text Mining: Wissensrohstoff Text*, W3L –Verlag. Herdecke Bochum 2008.

⁶ *Türk Dil Kurumu* [online]. Available at: <www.tdk.gov.tr>

semantic categories (Society, Science, Arts and Religion & Philosophy) and subcategories. This allows searching by those areas and can be also used for determining the context of a given text.

Here is an example of the semantic tree:



Total semantic categories: 45

The pragmatic structure comprises 16 categories (i.e. archaism, dialect, vulgar, colloquial).

The DB currently contains approx. 51.000 lexemes with 77.000 sememes and is permanently growing. By automatic analysis of texts new candidates for lexemes are extracted and inserted into the DB.

2. Methodology

All sememes can be selected by approx. 15.000 combinations of different criteria.

The most populous of the grammatical categories are:

- Nouns (female, male, neuter and plural only)
- Verbs (finite, infinite and finite + infinite)
- Adjectives
- Adverbs

Beside this, all sememes are internally marked if they are related to English (internationalisms). This is done semiautomatically by defined rules of similarity and in special cases manually corrected.

The markings for the semantic and pragmatic categories have been entered manually, similar to the entries in the printed dictionaries (i.e. archaic word, offensive, jovial).

The software for the statistics was written to extract the needed information and is easily adaptable⁷.

3. Statistics

Here is a glance at the current statistics of grammatical categories:

Grammar	Lexemes	Derivations	Sememes
Verb infinite	3598	4767	6365
Verb finite and infinite	1101	8	1813
Verb finite	7782	53	16713
Verb	12481	4828	24891
Adverb	1644	1518	2062
Noun female	9453	3649	13515
Noun male	11167	409	15332

Nouns neuter	4809	5822	6189
Nouns plural only	191	12	218
Noun	25620	9892	35254
Adjective	10127	2010	13118
Number	137	9	161
Pronoun	107	3	203
Exclamation	179	18	235
Conjunction	78	2	164
Particle	76	5	132
Preposition	75	2	312
Modal word	18	5	25
Prefix	128	1	194
Abbreviation	36	2	37
Compound word	38	2	39
Name (toponym)	126	6	131
Total	50870	18303	76958
Total, including derivations	69173		95261

Table 1

Derivations are lexemes which are directly connected to other lexemes, have the same meaning, but vary in small aspects (i.e. male/female).

In order to simplify the calculations we will concentrate only on lexemes of the four main categories (noun, adjective, verb and adverb).

Grammar	Lexemes
Nouns	25620
Adjectives	10127
Verbs	12481
Adverbs	1644
Total	49872

Table 2

The comparable statistics of lexemes related to English:

Grammar	Lexemes
Nouns	3277
Adjectives	640
Verbs	543
Adverbs	26
Total	4486

Table 3

Lexemes related to Turkish:

Grammar	Lexemes
Nouns	1717
Adjectives	190
Verbs	187
Adverbs	62
Total	2156

Table 4

⁷ JOHNSON, Keith. *Quantitative Methods in Linguistics*. Oxford 2008.

As a comparison here are numbers for the German lexemes derived from the German-Macedonian dictionary⁸ and the monolingual German dictionary⁹.

Grammar	Lexemes
Nouns	21794
Adjectives	6328
Verbs	7031
Adverbs	963
Total	36116

Table 5

Here is also an overview of the leading lexemes by semantic categories:

Medicine	1730
Economy	1380
Botany	1069
Gastronomy	1030
Religion	939
Zoology	878
Agriculture	599
Architecture & Building	594
Music & dancing	589

Table 6

4. Results

The first striking result is the ratio between the nouns and verbs (N/V) in the Macedonian language which is 2.05. In German this ratio is 3.10. This means that the Macedonian language has disproportionately high number of verbs in comparison to German.

The current statistics indicates 9.0 percent English words (internationalisms) and 4.32 percent Turkish words in the Macedonian language.

A look at the sememes shows that Macedonian is tilted toward male gender. The ratio between male and female nouns is 1.13.

The list of the semantic categories can be also seen as priorities of the Macedonian way of living, concerned mostly with health, food, religion and music & dancing!

5. Interpretation

The most striking fact is the relative dominance of the verbs. At the level of sememes this is even stronger (N/V=1.42). The relative dominance of the verbs is easily explainable. The first explanation lies in the existence of finite and infinite verbs. Many verbs appear in both forms,

but as different verbs. Using multiple prefixes creates numerous variations of each verb, which leads to a richness when expressing actions. But on the other side this also shows the lack of nouns. This can be seen from the table of the English and Turkish words. Most of the foreign words are nouns (English N/V=6.30, Turkish N/V=9.18).

The lesson for the translators dealing with Macedonian is that they should be familiar with the verbs and be ready to circumscribe the lack of nouns.

6. Conclusion

Until now, the classification and characterization of languages has been done based on a single characteristic or on a group of selected words. With the advance of computer based dictionaries we have the possibility to compare languages on a quantitative level. This is the first attempt and we hope there will follow more soon.

Creating a dictionary is a permanent process. It requires mechanisms for identifying new lexemes and new sememes. But at the end it becomes a treasury for analysing the language and identifying its strengths and weaknesses. It also gives us a tool for analysing the influences from other languages.

The Macedonian dictionary will be further developed, especially by improving the automatic analysis of text and creation of different corpora.

SUMMARY: *Quantitative Characteristics of the Contemporary Macedonian Language*. The huge, pioneer project of the Digital Dictionary of the Macedonian Language (Digitalen rečnik na makedonskiot jazik) has come about as a by-product of the effort to create a high-quality, state-of-the-art online German-Macedonian Dictionary, intended for Macedonian immigrants in Germany. Yet, the corpus that had been collected exceeded all original expectations. It has grown to become an independent digital dictionary of Macedonian, one that continues in the tradition of printed academic dictionaries of standard Macedonian. This tradition includes Koneski's dictionary (1961, 1996) and the extensive monolingual Dictionary of the Macedonian Language (Tolkoven rečnik na makedonskiot jazik), compiled at the Institute for Macedonian language "Krste Misirkov" at Ss. Cyril and Methodius University in Skopje from 2003 to 2014. The structure of the Digital Dictionary of the Macedonian Language fully reflects all linguistic principles regarding grammar, semantics, and pragmatics.

Sinasi Derebey
Deutsch-Makedonische Gesellschaft e.V. Berlin
Germany
derebey@sam97.de

⁸ *Deutsch-Makedonisches Wörterbuch* [online]. Available at <www.makedonisch.info>

⁹ *Digitales Wörterbuch der Deutschen Sprache* [online]. Available at <www.dwds.de>