sciendo
DOI 10.2478/jazcas-2025-0019

PHRASEMES AND COLLOCATIONS IN THE CORPUS – HOW TO FIND UNKNOWN VARIANTS

HANA SKOUMALOVÁ[1] – PŘEMYSL VÍTOVEC[2] – MILENA HNÁTKOVÁ[3]
[1]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0000-0002-3519-0233)
[2]Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0009-0004-2286-1706)
[3]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0000-0002-4790-9807)

SKOUMALOVÁ, Hana – VÍTOVEC, Přemysl – HNÁTKOVÁ, Milena: Phrasemes and Collocations in the Corpus – How to Find Unknown Variants. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 212 – 222.

Abstract: This paper addresses the identification and annotation of multiword expressions (MWEs) in Czech corpora, focusing on enhancing the search procedure through transformations of existing lexicon entries and the addition of new entries based on syntactic patterns. We discuss the limitations of current annotation systems and introduce a new, efficient annotation system that leverages a comprehensive MWE dictionary. Our methodology includes the use of syntactic patterns to identify new collocations, automatic transformations of known MWEs, and manual searches for creatively varied expressions. The results demonstrate significant improvements in the success rate of corpus annotation, with newly identified collocations and transformed MWEs contributing to a richer and more accurate linguistic resource.

Keywords: emultiword expressions, corpus annotation, syntactic patterns, lexicon transformations, Czech language

1 INTRODUCTION

Collocations and phrasemes are integral to language, studied within phraseology, but their annotation in corpora lags behind other types of annotation. The most comprehensive phraseologically annotated corpus of Czech is SYNv13 (Křen et al. 2024), with a size of about 6.5 billion positions. Another corpus with partial MWE annotation is PDT-C (Hajič et al. 2024), where MWE annotation is part of the annotation at the deep syntactic level. The MWEs were manually annotated and are mostly verb phrases contained in the Vallex dictionary (see Lopatková et al. 2016 and 2022). The size of the syntactically annotated part of PDT-C is about 2.25 mln, but MWEs are annotated only in its part (the original PDT) of about 675,000 words. There is also a pilot corpus resulting from the PARSEME project (Savary et al. 2023), which contains about 830,000 positions.

Another large Czech corpus is csTenTen from the TenTen corpus series (Jakubíček et al. 2013), which contains about 5.7 billion words. The corpus does not directly contain phraseological annotation, but it is possible to use so-called word sketches that reveal the collocation of individual words. Another tool associated with the TenTen corpora is a frequency-ordered list of n-grams, which actually represent MWEs.

When searching for phrasemes and collocations in the corpus, two approaches are possible. One is to use various statistical measures or sketches and n-grams, whereby the user is given a frequency list of the collocations found. These methods are useful for extracting information about individual words and their collocability. The other approach searches and annotates the corpus for collocations based on the dictionary, trying to find all variants of a certain, previously known collocation. This approach is suitable for phraseological research done on corpora. Unlike the first approach, it is possible to annotate (and later retrieve) e.g. proverbs or long sayings that would be difficult to find using statistical methods. Methods based on n-grams or word sketches cannot capture all possible word order variations, different inter-word distances and possible MWE transformations.

2 ANNOTATION OF CORPORA WITH PHRASEMES

In our work, we use an MWE dictionary. For corpus annotation we still use the now obsolete FRANTA system (see Kopřivová and Hnátková 2012). The disadvantages of this solution are (1) the specific format of the phraseological dictionary, which is only machine-readable, and (2) the insufficient speed of annotation. In response to these shortcomings, we are currently working on a new annotation system that works with data from the MWE database LEMUR (see Skoumalová et al. 2024) and implements a very efficient retrieval and annotation algorithm.

Both the dictionary of the FRANTA system (called FRANTALEX) and the dictionary represented by the LEMUR database are based on the Dictionary of Czech phraseology and idiomatics (SČFI, Čermák et al. 1983–2009), but are enriched with other phrasemes and collocations found in corpora (see Hnátková 2006). The dictionary contained in LEMUR is not only machine-readable but it is also suitable for human users (see Skoumalová et al. 2024). It also contains much more information about each entry so it is not only useful for finding collocations in the corpus. A final advantage of the new system is that it can annotate much faster than the previous one, which is mainly due to the fact that the dictionary is compiled into a machine-readable form before being used by the search engine.

FRANTALEX, which serves as a source of entries for the new system, contains about 56,000 entries. A large part of it has already been transferred to the new database, which contains about 26,000 entries, but the two numbers cannot be

straightforwardly compared – when the entries are transferred, some variants that were previously separate entries are merged into one entry.

When using either system for corpus annotation, we take care to search for different word-ordered and disjointed variants, or variants with changed lexical content, or fragments (see Kopřivová and Hnátková 2012), e.g.

(1) a. ***účel*** ***světí*** *v boji* ***prostředky***
purpose sanctifies in combat means
'the end justifies the means in combat'

b. ***účel*** *mediální propagandy* ***světí*** *jakékoliv* ***prostředky***
purpose of media propaganda sanctifies any means

c. ***účel*** *a případný úspěch v politice* ***světí*** *a často omlouvá* ***prostředky***
purpose and eventual success in politics sanctifies and often excuses means

d. *nepsal nic o* ***prostředcích,*** *které by* ***účel*** ***světil***
wrote nothing about means that.ACC would the purpose sanctified
'he didn't write anything about the means that would the purpose sanctify'

(2) ***vnímat*** ***jako*** ***hrozbu*** ⇒ ***brát*** ***jako*** ***hrozbu***
perceive as threat take as threat

(3) ***Kdo*** ***jinému*** ***jámu...***
Who.NOM else.DAT hole.ACC...
'[He] who [digs] another's hole [falls into it himself.]'

The word-order and discontinuous variants as well as fragments are described directly in the FRANTALEX dictionary and in the LEMUR database, respectively, and will not be dealt with in this article. We will assume that the newly identified and described MWEs can also occur in such variants.

3 METHODS OF SEARCHING FOR NEW (VARIANTS OF) MWES

However, we have more ambitious goals, namely to create additional variants during compilation – transformations of existing units.

In addition to working with existing units, we are also looking for candidates to be added to the dictionary. This search cannot be done during annotation, but special CQL queries are entered into the annotated corpus, the results are then sorted by frequency and candidates for addition to the MWE dictionary are manually selected.

A final way to search for unknown variants of known collocations is to search for variants that have been creatively varied by speakers of the language. These are various adaptations of proverbs, well-known quotes and sayings. Sometimes two such expressions are contaminated, either deliberately or through ignorance. Example 4 shows one such case.

(4) a. ***mlsný*** ***jazýček*** ***na vahách***
picky tongue/pointer on scales

b. ***mlsný*** ***jazýček***
picky tongue

c. ***jazýček*** ***na vahách***
tongue on scales
'pointer on scales'

The phraseme in 4. a. is a compound of the phrasemes in 4. b. and c. and was used to describe a small political party that could choose which way to lean, and therefore it could make demands.

3.1 New adepts for a dictionary

The basic way to enrich the dictionary with new entries is to search for new collocations based on syntactic patterns. In this way, by which we still enrich FRANTALEX and then transfer the found MWEs to LEMUR, mainly established compounds and terms are found. In the early days of dictionary building, we focused only on semantically idiomatic MWEs. However, we are currently expanding the dictionary to include statistically idiomatic expressions as well. Syntactic patterns such as Adj+Noun, Verb+Noun.ACC, Noun+Noun.GEN, etc. are useful for searching such expressions.

The search is performed by issuing a CQL query to find a certain sequence of tags, e.g. a query

```
1:[tag="A.*"] 2:[tag="NN.*"] & 1.c=2.c within <s/>
```

searches for an Adj-Noun sequence in the same case within a single sentence. Other similar queries are

`[tag="V.*"] [tag="NN..4.*"] within <s/>`, which searches for verbs with an object in the accusative;

`[tag="NN..[^2].*"] [tag="NN..2.*"] within <s/>`, which finds a noun modified by another noun in the genitive case;

`[tag="NN..[^2].*"] [tag="A...2.*"] [tag="NN..2.*"] within <s/>`, which finds a noun modified by an adjective and a noun in the genitive case.

We sort the results of each query by frequency and manually select new entries for the dictionary.

3.2 Identification of MWE transformations during annotation

Another way to search for variants that are not explicitly captured in the dictionary is to create regular transformations of (mainly) verb constructions. For FRANTA, these transformations are created automatically for single phrases and then manually added to the dictionary. For this reason, there are only a limited number of them in FRANTALEX. For a system using LEMUR, they are created automatically when the dictionary is compiled.

The simplest transformations are passivization, nominalization and adjectivization. In these transformations, the base word or its form is changed, and the valency frame may also change, which affects other words in the phraseme. For these diatheses we follow the rules formulated in Rosen and Skoumalová (2018). For example, the saying *hodit flintu do žita* lit. 'to throw rifle into rye(field)', 'to throw in the towel' is found in all moods, tenses and persons in texts, but it is also possible to create the periphrastic passive *flinta je hozena do žita* 'the towel is thrown in' or the reflexive passive *flinta se hodí do žita*. Since the verb *hodit* 'to throw' is transitive in this construction (and has an object in the accusative case), the transformation for the periphrastic passive is as follows:

1. The object in the accusative changes its case to the nominative and becomes the subject of the construction.
2. The rest of the construction is unchanged.
3. The verb can only be in the passive participle form.

For the reflexive passive, similar rules apply:

1. The object in the accusative changes its case to the nominative and becomes the subject of the construction.
2. The reflexive particle *se* is added to the construction.
3. The rest of the construction does not change.
4. The verb can only be in active forms.

In the algorithm described above, we do not mention the subject of the original construction, because we only work with verb constructions in their basic (dictionary) form, which is the infinitive.

For both kinds of passive, it holds that they cannot be formed from reflexive verbs, so that, for example, in the saying *bojovat/prát se/zahrát si pro čest a slávu* 'to fight/play for honor and glory', only the verb *bojovat* 'to fight' can undergo passive diathesis.

Another possible transformation is nominalization; in our example it would be *hození flinty do žita* 'throwing a rifle into rye', or the adjectivization *flinta hozená do žita* 'rifle thrown into rye'. Some nominalizations and adjectivizations of verbs are

word-formationally regular and are captured in the morphological dictionary (see Štěpánková et al. 2020). Other, irregular derivations are retrieved using the Derinet system (Ševčíková and Žabokrtský 2014). From the Derinet network, we retrieve not only nouns derived (according to the traditional view of word formation) from verbs, but we also retrieve words that did not arise by traditional derivation (e.g. *práce* 'work' as a derivative of *pracovat* 'to work'). We can also look for derivations of nouns that fill other positions in the phrase, e.g. diminutives, or feminine nouns. In this way, automatic transformations yield additional variants of the phrases in the dictionary, e.g. *hození flinty do žita* 'the throwing of a rifle into the rye', *flinta hozená do žita* 'a rifle thrown into the rye', or *házející flintu do žita* '[sb] throwing a rifle into the rye', or possibly *ministryně financí* 'female-minister of finance' or *zdravotní sestřička* lit. 'medical little sister', 'nurse'. A partial listing of derivations made during dictionary compilation is shown in Fig. 1.

```
... Processing hodit flintu do žita

Scope of "zahodit:zahodit:V flintu:flinta:NNFS4 do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]" = CLAUSE
Derived PASSIVE: "flintu:flinta:NNFS1 být:VB-S[aux,ignore] zahodit:zahodit:Vs do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived PASSIVE: "flintu:flinta:NNFS1 být:Vp.S[aux,ignore] zahodit:zahodit:Vs do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived REFLPASS: "flintu:flinta:NNFS1 se:P7--4[ignore] zahodit:zahodit:VB-S do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived REFLPASS: "flintu:flinta:NNFS1 se:P7--4[ignore] zahodit:zahodit:Vp.S do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived REFLPASS: "flintu:flinta:NNFS1 se:P7--4[ignore] být:VB-S[aux,ignore] zahodit:zahodit:Vf do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NOMVERB: "zahodit:zahození:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NEGNOMVERB: "zahodit:nezahození:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NOMDER: "zahodit:zához:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NOMRESPASS: "zahodit:zahozenost:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NEGNOMRESPASS: "zahodit:nezahozenost:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived ADJRESPASS: "flintu:flinta:NNFS2 zahodit:zahozený:A[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived ADJRESACTT: "zahodit:zahodivší:A flintu:flinta:NNFS4 do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
```

Fig. 1. Partial list of transformations of the saying *hodit flintu do žita*

We can see that during transformations, overgeneration occurs – we will probably never encounter *nezahozenost flinty* 'rifle's un-thrown-ness' in the corpus, but for this reason, overgeneration is not a problem. A problem can arise if, for example, a diminutive has a different meaning than the base word. For example, *stará panna* 'old maid' and *stará panenka* 'old doll'. We must solve these cases individually and prevent such diminutives from being generated and used.

3.3 Searching for unknown variants in an annotated corpus

As mentioned above, the authors of the texts often creatively modify well-known proverbs, sayings and quotations and their identification in the corpus is difficult. For these purposes, there is no choice but to pick out a possible phraseme and enter CQL queries that might reveal its variations. We illustrate this search with the idiom *vlk se nažral a koza zůstala celá* lit. 'the wolf has eaten and the goat has remained whole', 'an order was formally filled, but practically nothing changed'. If

we enter a CQL query (5) that searches for the lemmas *koza* 'goat' and *nažrat* 'eat' within 10 positions of each other within a single sentence, we get the occurrences shown in Example 6.

(5) `(meet [lemma="koza" & col_lemma=""] [lemma="nažrat"] -5 5)`

(6) a. *..., aby* ***se*** *konkurzní hyeny* ***nažraly*** ***a koza zůstala celá***
..., so that bankruptcy hyenas ate and goat remained whole

b. ***vlk*** *poznání* ***se nažere*** ***a*** *klipová* ***koza*** *mečí do éteru dál*
wolf of knowledge eats and clip goat keeps bleating into ether

c. ***Vlk se nažral*** ***a kozy*** ***zůstala*** *půlka.*
Wolf has eaten and of goat remained half.
'The wolf has eaten and a half of the goat remained.'

d. *... dát* ***nažrat*** ***vlkovi,*** *aby* ***koza*** *přitom* ***zůstala celá.***
... give eat.INF wolf.DAT so that goat at the same time remained whole
'... to let the wolf eat so that the goat remained whole at the same time'

e. ***Koza se nažere,*** ***vlk zůstane celej,*** *já mám po starostech,...*
Goat eats, wolf remains whole, I have no troubles

It is clear that all of these findings refer to the original saying, but none of them has been identified as an occurrence of it. The variation may consist in an altered lexical setting (6. a. and b.), in a modification of meaning (6. c.), in a change of modality with the corresponding change of case (6. d.), or in a complete reversal of meaning (6. e.). If we modify the CQL query to include two other words from the original saying, we will get additional variations.

The question is whether we should even try to describe and find these variants when annotating. For those that preserve the semantics of the original saying, we need to modify the constraints in the dictionary to allow other lexical settings, or to allow a fragment to suffice for identification. Where the semantics differs, we need to consider a new entry in the dictionary (if the new phraseme is frequent enough), which will be linked to the original entry by a super-lemma.

4 RESULTS OF INDIVIDUAL METHODS

All three methods mentioned in the previous section were really used, although the third method (manual search for variants) was used only to a limited extent. However, the first two methods significantly improved the success rate of corpus annotation. Following is an overview by each method.

4.1 Newly added phrasemes and collocations

The new collocations added by the syntactic patterns search were used in the annotation of a testing corpus of 130 million words, NEWTON2023, a corpus of journalism acquired in 2024, which was annotated by the FRANTA system. Counting the types, the new collocations represent 7.57% of the annotated collocations and for occurrences (tokens) they represent 16.35%. The following table compares the frequency of new collocations with the original ones.

	original types	new types	original tokens	new tokens
Adj-Noun	5,321	2,208	627,272	457,497
Noun-Adj.GEN-Noun.GEN	12	291	133	6,946
Noun-Noun.GEN	2,273	756	56,512	77,284
Verb-Noun.ACC	5,479	362	252,498	12,702

Tab. 1. Comparison of the frequency of new collocations with original collocations

We can see that some syntactic patterns yielded a large number of collocations identified in the corpus, although the newly found types (i.e., individual collocations) were not as numerous. However, these were the most frequent established expressions such as *životní prostředí* 'environment', *hlavní město* 'capital city', or *mistrovství světa* 'world championship', *růst cen* 'price rise', *ministr financí* 'finance minister', etc.

4.2 Collocations and phrasemes identified using transformations

This method has not yet been used for the annotation of any published corpus, we are still testing it. We annotated the same test corpus of 130 million words with a method using a compiled dictionary from LEMUR with automatic transformations, and we found that 5,335 transformations were applied out of 765,518 generated, which is about 0.7% of the proposed transformations. However, there are some very frequent ones among them, such as *zvýšení daně* 'tax increase', which has a higher frequency than the basic form *zvýšit daň* 'to increase tax' (i.p.m. 13.73 versus 5.78), or *odchod do důchodu* 'retirement' (i.p.m. 9.55) versus *jít do důchodu* 'to retire' (i.p.m. 4.91). The distribution of transformations in the corpus by type is shown in Tab. 2.

Type	Occurrences	% of collocations
Without transformations	3,397,366	97,56
PASSIVE (participle ending with *-n/-t*)	6,082	0,17
REFLPASS	12,100	0,35
NOMVERB (nominalization of verb – *-ní/-tí*)	35,694	1,03
NEGNOMVERB (negation of the above)	603	0,02
NOMDER (derived noun – *hrát* 'play' – *herec* 'actor')	17,409	0,50

NOMRESPASS (pass. result – *-nost/-tost*)	4	0
NEGNOMRESPASS (negation)	0	0
NOMPOTIMP (possibility – *-telnost*)	19	0
NEGNOMPOTIMP (negation)	1	0
NOMRESACTL (act result – *-lost*)	3	0
NEGNOMRESACTL (negation)	0	0
NPDER (diminutive, feminine... – *-yně/-ček/-čka*)	7,485	0,21
ADJPOTIMP (possibility – *-telný*)	37	0
ADJPROC (active adj. – *-ící*)	2,933	0,08
ADJRESACTL (act. pres. result – *-lý*)	96	0
ADJRESACTT (act. past result – *-vší*)	1	0
ADJRESPASS (passive result – *-ný/-tý*)	2,339	0,07
TOTAL	3,482,172	100

Tab. 2. Frequency of transformations in the corpus

We can see that some transformations have very low representation in the texts. For example, NOMRESPASS denotes derived nouns expressing a resulting state after some action, ending in *-ost*, e.g. *zajištěnost dodávek* 'supply assurance', *sehranost komedie* 'comedy enactment', etc. On the other hand, derived nouns ending in *-ní/-tí* (NOMVERB) represent the most numerous group among the transformations.

5 CONCLUSIONS

In our paper, we have shown three methods that can be used to extend the MWE lexicon and/or improve the success rate of corpus annotation with MWEs. We tried three methods: we retrieved potential collocations according to a syntactic pattern, we used transformations of known collocations and phrasemes, and we tried retrieval of lexically varied and fragmentary variants.

The first method seems to be the most beneficial in terms of the number of subsequently annotated collocations. However, it has a limitation in that it only finds collocations that occur in the canonical word order in the texts and are not split by other words.

The second and third methods do not yield as many newly annotated variants of collocations, but they open up new possibilities in research on the variation of phrasemes and collocations. On the one hand, there are possibilities to investigate what transformations are possible for collocations and how often speakers use them, and on the other hand, it is also possible to investigate the creative variation of established collocations and phrasemes.

In future work, we will develop all three methods of dictionary enrichment and corpus annotation and use them to annotate other corpora.

ACKNOWLEDGEMENTS

The research has been supported by the Czech Science Foundation under the project GA CR 24-10254S, by the Technology Agency of the Czech Republic under the project TQ01000177, and by Ministry of Education, Youth and Sports under the Large Research, Development and Innovation Infrastructures, LM2023044.

References

Čermák, F., et al. (1983–2009). Slovník české frazeologie a idiomatiky 1–4. Praha: Academia/Leda.

Hajič, J., et al. (2024). Prague Dependency Treebank – Consolidated 2.0 (PDT-C 2.0). Data/software, LINDAT-CLARIAH-CZ. Accessible at: http://hdl.handle.net/11234/1-5813.

Hnátková, M. (2006). Typy a povaha komponentů neslovesných frazémů z hlediska lexikálního obsazení. In: F. Čermák – M. Šulc (eds.): Kolokace, Nakladatelství Lidové noviny/Ústav Českého národního korpusu, Praha, pp. 142–167.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013). The TenTen Corpus Family. In 7th International Corpus Linguistics Conference CL 2013, pp. 125–127. Lancaster.

Kopřivová, M., and Hnátková, M. (2012). From Dictionary to Corpus. In Phraseology in Dictionaries and Corpora, pp. 155–168. Maribor.

Křen, M., Cvrček, V., Čapka, T., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Skoumalová, H., Šindlerová, J., and Škrabal, M. (2024). Korpus SYN, v13 from 27/12/2024. Ústav Českého národního korpusu FF UK, Praha. Accessible at: https://www.korpus.cz.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2016). Valenční slovník českých sloves VALLEX. Praha: Karolinum.

Lopatková, M., Kettnerová, V., Mírovský, J., Vernerová, A., Bejček, E., and Žabokrtský, Z. (2022). VALLEX 4.5. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague. Accessible at: http://hdl.handle.net/11234/1-4756.

Rosen, A., and Skoumalová, H. (2018). No way to have your say out of the frame: specifying valency of multi-word expressions. Prace filologiczne (LXXII), pp. 301–320.

Savary, A., et al. (2023). PARSEME corpora annotated for verbal multiword expressions (version 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Accessible at: http://hdl.handle.net/11372/LRT-5124.

Ševčíková, M., and Žabokrtský, Z. (2014). Word-Formation Network for Czech. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 1087–1093. Reykjavík.

Skoumalová, H., Kopřivová, M., Petkevič, V., Jelínek, T., Rosen, A., Vondřička, P., and Hnátková, M. (2024). Lemur: A lexicon of Czech multiword expressions. In: V. Giouli – V. Barbu Mititelu (eds.): Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives. Language Science Press, Berlin, pp. 1–37.

Štěpánková, B., Mikulová, M., and Hajič, J. (2020). The MorfFlex Dictionary of Czech as a Source of Linguistic Data. In Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, pp. 387–392. Democritus University of Thrace, Thrace, Greece.