What belongs in a dictionary? The Example of Negation in Czech

Dominika Kovarikova, Lucie Chlumska & Vaclav Cvrcek

Keywords: *negation*, *lexicography*, *grammatical category*, *frequency*, *lemmatization*.

Abstract

In this paper, the authors try to answer the basic lexicographical question: how do we know whether a particular word is a mere word form, or a new lexeme that should thus be assigned an individual entry in a dictionary? The issue of negation in Czech (namely negative forms of nouns, adjectives, adverbs and verbs) serves them as a perfect example. They introduce two criteria for the choice of dictionary entries, the frequency criterion and the grammatical category criterion, and show how the negation of the parts of speech examined differs and what the implications are for lexicographers.

1. Introduction

There is no doubt that a reliable monolingual dictionary in any language serves as a crucial source of knowledge for both native speakers and foreigners. When compiling a new dictionary of a language, there are at least two possible approaches. The first, still very popular among some lexicographers, makes use of existing dictionaries and their lists of entries, no matter how obsolete they may sometimes be, and simply modifies them. The other method – which we prefer – starts from scratch, i.e. uses a large contemporary corpus to compile a new index of words, which are frequent enough to be included in a dictionary (depending on its intended size and purpose, of course).

2. Czech Dictionaries

In Czech, there are two large monolingual dictionaries that aspire to comprise the Czech language as a whole; the first, *Prirucni slovnik jazyka ceskeho* (PSJC), with 250 000 entries, was published in 1935-57, and the newer *Slovnik spisovneho jazyka ceskeho* (SSJC), with 192 000 entries, in 1960-71. The lexicographical tradition was interrupted in the 70s and between 1978 and 2003 only several slightly revised and abridged versions based on the SSJC were published. It is therefore apparent that a new comprehensive dictionary of contemporary Czech is more than necessary. Preliminary work on this dictionary has already started at the Czech Academy of Sciences. Surprisingly enough, it is not entirely corpus-based, although a large corpus of contemporary Czech has been available since 2000. Although it would be interesting to discuss the pros and cons of the CAS approach, in this paper we would like to focus on the one question every lexicographer must ask: When writing a dictionary, what criteria help us decide what to include? In other words, how do we know whether a particular word is a mere word form, or a new lexeme that requires its own individual entry in the dictionary? The issue of negation and negative forms in Czech will serve us as a perfect example. [1]

3. Corpus Data

The SYN corpus used for our research was created at the Institute of the Czech National Corpus and currently comprises 1.3 billion words (tokens). The vast majority of the texts in

the SYN corpus belong to the category of newspapers and magazines. Although it is not balanced and representative in the traditional sense, it is the largest corpus of contemporary Czech available and thus provides the best overview of current Czech. Moreover, the issue under study, i.e. negation of nouns, verbs adjectives and adverbs, is almost inert to the deviations caused by non-representativeness of corpora. The SYN corpus is lemmatized (error rate app. 2 %) and morphologically tagged (error rate app. 4 %), which has both advantages and disadvantages for lexicographers. With Czech being a highly flective language (i.e. the average Czech lexeme has app. 13 different word forms (Cvrček (forthcoming)), lemmatization may save a lexicographer many hours of work.

On the other hand, relying entirely on lemmatization may be very risky, since it is not entirely consistent, esp. for negative adjective and adverbial lemmas. For example, the lemma *šťastný* (*happy*) includes both the affirmative and negative forms (*šťastný* and its derivates as well as *nešťastný* and its derivates), whereas negative lemma *nešťastný* (*unhappy*) includes solely rare morphological variants *nešťasten/nešťastna/nešťastno* (*unhappy*). The lemma *slyšící* (*hearing, non-deaf*) comprises both *slyšící* and *neslyšící* (*deaf*), although the second is a proper term (*the hearing-impaired*) and should have its own lemma. Negative verbs are usually listed under an affirmative lemma and negative nouns have their own negative lemma. However, it would be advisable to adopt a unified approach based on corpus evidence.

4. Frequency Criterion

Our basic hypothesis is that a reliable dictionary should contain the core of a language, e.g. the most frequent and thus most important vocabulary. The frequency threshold of 250 occurrences was derived from the recent estimate of the size of a language core based on observation of the hapax-type ratio in the process of building a corpus. (ibid.) As we add texts to the corpus (up to a few million tokens), the hapax-type ratio decreases from its initial and maximal value (=1) to the local minimum. This is because the majority of added tokens are new instances of words already present in the corpus. After that, a turning point occurs beyond which the number of hapaxes grows at a faster pace than the number of non-hapax types. This turning point delimits the core of each language with regard to its structure and type. The average size of a core for Czech lemmas is 86,157. (ibid.) In order to cover all possible core elements, in our experiment we included all noun, adjective, verbal or adverbial lemmas which are in the 90,000 most frequent lemmas in the SYN corpus (corresponding to a frequency of 250 occurrences and higher).

5. Grammatical Category Criterion

The frequency criterion is a necessary presumption, but it is not sufficient for distinguishing between a word form and an individual lexeme. For example, on what grounds do we decide whether a negative word (with a frequency above 250) is a mere word form of an affirmative lexeme or whether it is an individual lexeme which deserves a separate entry in the dictionary? Usually, the criterion of grammatical category proves to be useful: if particular word forms can be considered variants of a word within a particular grammatical category (number, tense, case, comparison for adjectives etc.), they should be listed under a single entry in a dictionary. For example, the English word forms such as *bigger*, *dogs*, *jumped* are listed under the lemmas *big*, *dog* and *jump*. However, what to do with negative words in Czech? In other words, is negation a grammatical category in Czech? Traditionally, it is not considered to be; linguists usually describe it within syntax (verbal negation) or not at all

(adjective negation etc.), which results in differences and inconsistencies both in lemmatization and in existing dictionaries.

Let us first look at three distinctive features of a grammatical category as described in the latest Grammar of Contemporary Czech (Cvrček et al. (2010)). First, the meaning conveyed by the grammatical means in question is general, stable and roughly the same for all words concerned. Second, the meaning is expressed by a small number of formal constituents, and third, the meaning applies to the whole group of words (e.g. part of speech). The results will show whether these criteria apply to negation in Czech.

Another possible criterion – collocation – to distinguish between word forms and individual lexemes has not proved to be precise enough, since collocations differ not only for different lexemes, but also for different word forms, as Sinclair (2004) has pointed out.

6. Negation in Czech

To fully understand this issue, it is also necessary to describe the system of negation in our language, since it is different from other languages (e.g. English). In Czech, negation is formally created a) with the prefix ***ne-***, and b) less frequently with a particle ***ne*** placed in front of a word in a sentence (syntactical negation). There are other possibilities as well, but these are very rare. The prefix ***ne-*** can be added to a noun, verb, adjective or adverb, turning them into a negative word, e.g. *šťastný* > ***ne****šťastný* (*happy* > *unhappy*), *spát* > ***ne****spat* (*sleep* > *not to sleep*), *zdravě* > ***ne****zdravě* (*in a healthy way* > *in an unhealthy way*). Sometimes, there is a strong shift in meaning or even semantic field, e.g. *mocný* (*powerful*) v. ***ne****mocný* (*ill*). From obvious reasons, this type of negation is a difficult task for lexicographers. Should all the negative words (e.g. words created by adding a prefix ***ne-***) be assigned an individual entry in dictionary? Moreover, even though the grammatical means is identical (prefix ***ne-***), are there any differences between the parts of speech under examination?

7. Research Results

The first question we would like to answer is whether negation in Czech can be considered a grammatical category. Chart 1 shows how many per cent of nouns, adjectives, verbs and adverbs have both the affirmative (unmarked) and the negative (marked) form compared to the prototypical grammatical adjectival category of number (singular being considered an unmarked form and plural a marked form).

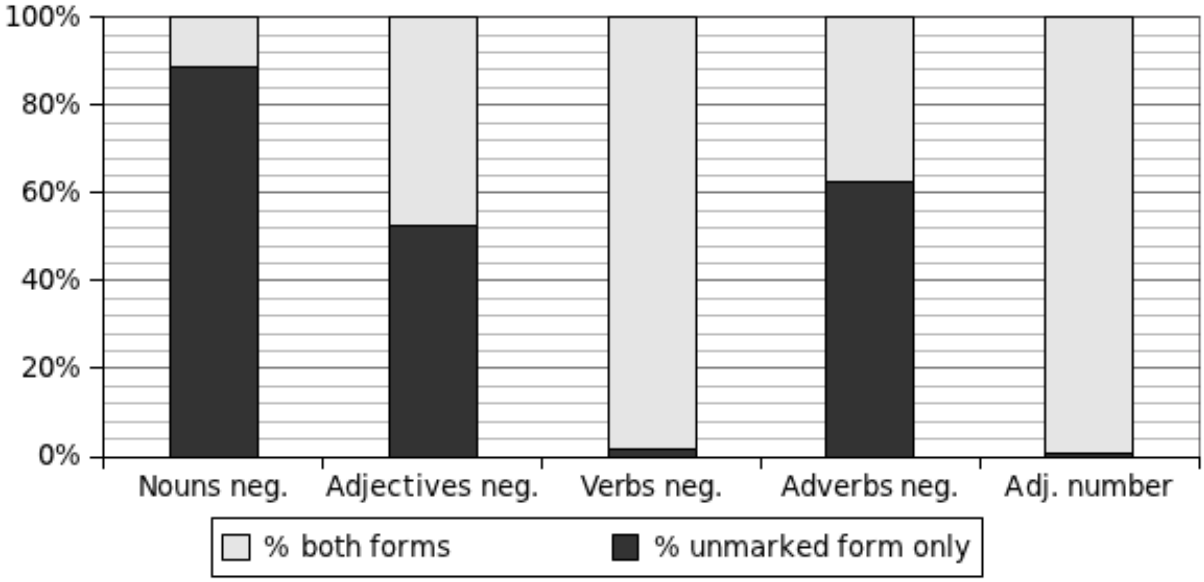

Chart 1. Is negation a grammatical category? Comparison of the percentage of negative forms (within nouns, adjectives, verbs and adverbs) and the prototypical category of number (adjectives).

As the fifth column suggests, the prototypical grammatical category of number covers 99 % of adjective lemmas (with the proper frequency), i.e. 99 % of adjective lemmas occur both in singular and plural form. As regards verbal negation, the coverage is almost identical. Nouns, on the other hand, do not follow the same pattern: only 12 % of noun lemmas occur in the corpus in negative form. Therefore, noun negation should most probably be described within lexicology, not grammar. Adjective and adverbial types of negation (48 % and 37 % respectively) come somewhere in between: they are to a certain extent regular, but cannot be compared with verbal negation.

As far as the above-mentioned criteria of grammatical category are concerned, the first two – the same meaning as well as a limited number of grammatical constituents (the prefix ***ne-***) – are met in all cases. The third condition – the full coverage within a word group – is definitely met in the case of verbal negation; we will therefore consider negation a grammatical category for this part of speech and suggest not including negative verbs in a dictionary as individual entries. When it comes to adjectives and adverbs, the situation is more complex and negation cannot be considered a grammatical category here.

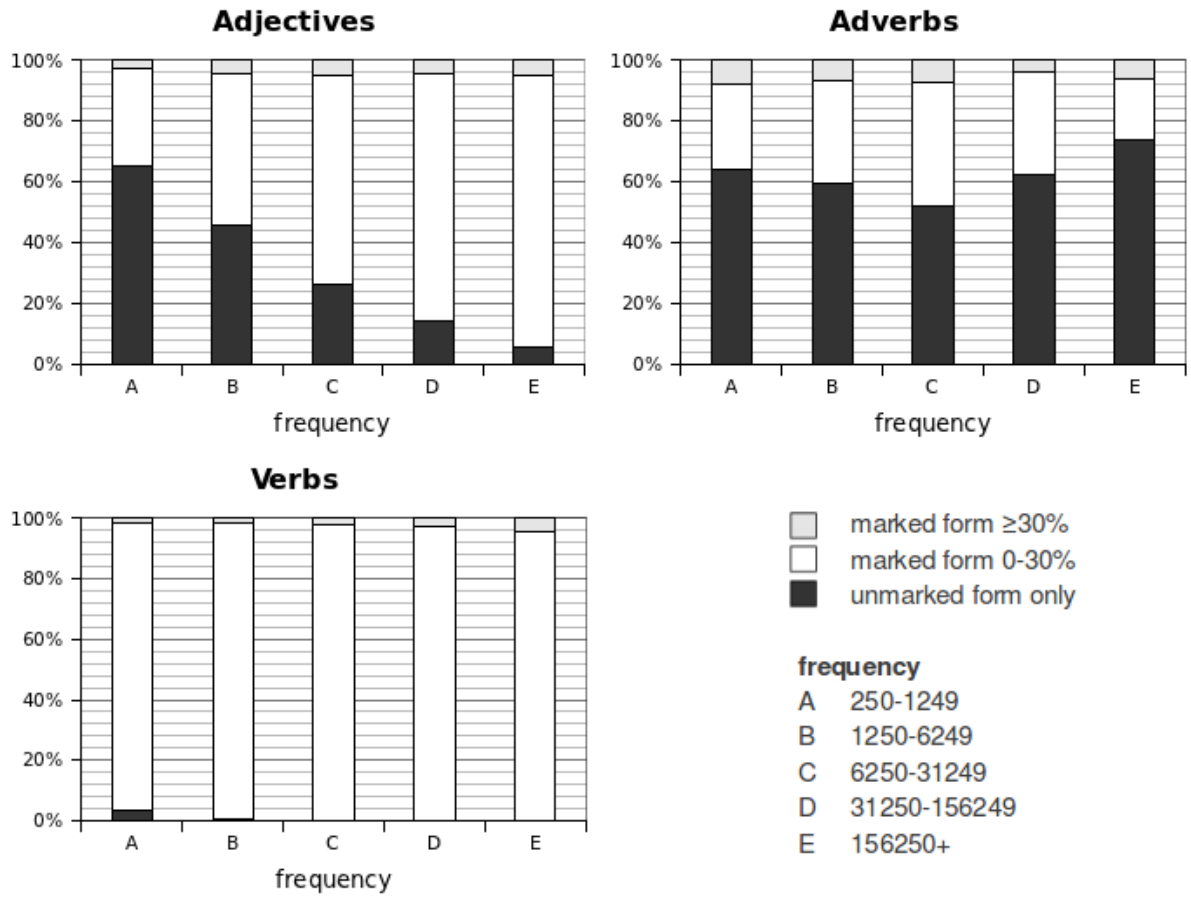

Chart 2. Negation in different frequency levels.

Chart 2 summarizes the coverage of marked and unmarked forms in five different frequency levels (the number of lemmas in each level is different; for the sake of comparison, the percentage is shown). Nouns are not included due to different lemmatization. The dark area refers to lemmas present in the corpus only in the affirmative form; grey and white areas cover lemmas with documented negative forms (grey stands for lemmas with more than 30 % of negative forms). Chart 2 shows how the number of negative forms changes with the increasing frequency of a lemma. These results suggest that the more frequent the adjectival or adverbial lemma[2], the higher the coverage of negative forms. This means that in such cases even adjectival and adverbial negation can be considered a grammatical category.

To summarize the results, we suggest the following approach to negation in Czech. Where negation is a proper grammatical category (verbs), the negative forms should not be assigned an individual entry in the dictionary. [3] Both of the dictionaries of Czech applied this rule intuitively and listed only affirmative forms of verbs. Negation of nouns belongs to lexicology; the negative forms should therefore be listed separately. As far as adjectives and adverbs are concerned, we suggest the following approach: based on their coverage, the frequent shift of meaning of their negative forms, and in consideration of both lexicographic tradition the convenience of dictionary users, we would recommend including both affirmative and negative forms (provided that they are frequent enough) in a dictionary as separate entries. Moreover, in an electronic dictionary (likely to be the most prevalent form of dictionary in future), it is more convenient to list affirmative and negative forms individually and interconnect them via referencing tools.

8. Further Implications

The results of this research have clearly shown that the negation in Czech has different characteristics for different parts of speech. This has implications not only for lexicographers, but also for Czech lemmatization in corpora; the lemmatization should consistently reflect these differences.

This study may also be useful when considering the possibilities of an electronic dictionary. The main advantages of such a dictionary – almost unlimited size, interconnectivity of entries, easy referencing both within the dictionary and to a corpus – can also be used to describe negation, not only in Czech, with all its aspects.

Notes

[1] In the above mentioned dictionaries, negation is treated very inconsistently. Some negative adjectives and adverbs are listed separately, some are mentioned within an affirmative lexeme. Some nouns do have an individual negative entry, some do not. Verbs are only mentioned in the affirmative form.

[2] The special pattern of adverbs is due to the high frequency of deictic adverbs, which do not have a negative form.

[3] The question is whether to apply the same rules to lemmatization as well. Many linguists argue that lemma in a corpus is not the same as a lexeme in a dictionary. If we consider lemma a mere functional unit designed to facilitate searching in a corpus, lemmatization should be first and foremost consistent: either to list all negative forms of all parts of speech under the affirmative lemma, or list both affirmative and negative forms separately. On the other hand, should lemma refer to a linguistic unit, the lemmatization ought to reflect the findings of this research (separate lemmas for nouns and preferably for adjectives and adverbs as well, affirmative lemmas only for verbs).

References

Cvrček, V. Forthcoming. 'How Large is the Core of Language?' In *Proceedings from the sixth international corpus linguistics conference 2011*. Birmingham.

Cvrček, V. et al. 2010. *Mluvnice současné češtiny*. Praha: Nakladatelství Karolinum.

Czech National Corpus - SYN. Institute of the Czech National Corpus, Prague. 10. 9. 2011, Accessible at: <http://www.korpus.cz>.

Kováříková, D. 2011. 'Gramatická kategorie negace.' *Korpusová lingvistika Praha* 2011.2. Výzkum a výstavba korpusů. NLN, Praha.

Příruční slovník jazyka českého. Praha 1935–1957.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. 2004. *Trust the Text. Language, corpus and discourse.* London: Routledge.

Slovník spisovného jazyka českého. Praha – 1. edition 1960–1971, 2. edition 1989.