The Lexicographer's Voice: Word Classes in the Digital Era

Geda Paulsen, Ene Vainik, Maria Tuulik and Ahti Lohk

Institute of the Estonian Language, Estonia

E-mail: geda.paulsen@eki.ee, ene.vainik@eki.ee, maria.tuulik@eki.ee, ahti.lohk@eki.ee

Abstract

The present study examines the role of word classes in contemporary lexicography using examples from Estonian. Since Estonian is a morphologically rich language, the results may be extendable to other languages with abundant morphology. Two research questions are examined: i) What are the problems and practices of lexicographers when determining word classes? and ii) What are the needs and expectations of lexicographers for a possible digital tool that would facilitate word class identification? The results of a metalexicographic survey carried out among 23 Estonian lexicographers show the relevance of word classes as a categorial frame in their lexicographic work. There is a need to improve or reconsider the (theoretical and technical) factors influencing the process of PoS tagging. A reliable software application (provisionally a PoS evaluator) easing the decision making process would be welcome. According to the ideas suggested by the respondents, the solution would be an improved morphological and syntactic parsing system with respect to the present solutions, and a corpus-driven application presenting statistics with regard to the morphosyntactic distribution of an ambiguous word with access to the data source.

Keywords: lexicography; word classes; metalexicographic survey; Estonian

1. Introduction

The challenge of modern lexicography is to create digital tools which would be able to present meaningful and reliable generalizations over a large amount of raw data in a way that meets the specific needs of lexicographers, including inter alia the procedures of word class categorization[1]. Although several studies indicate that lexical categories are far from clear-cut or self-apparent (see, e.g., Mark, 2015; Croft, 2001; Culicover 1999), grouping lexemes is vital for the information that dictionaries provide. The task of PoS markup has not disappeared in the digital era of lexicography, when stand-alone dictionaries are increasingly replaced by unified and standardized databases. On the contrary – integrated all-purpose root databases comprise all kinds of information for as many lexemes as possible. The data models of such databases typically include a PoS unit (e.g. Tavast et al., 2018).

The aim of this study is to clarify the role of word class categorization in contemporary lexicographic work in Estonian. It comprises the first step of a project that aims to develop a corpus-driven solution, tailored to the needs of lexicographers. We believe that the results can be extended to other languages with abundant morphology when

[1] Throughout the study, we use the notions word class and part of speech (PoS) synonymously.

planning e-lexicographic projects with similar goals.

In order to establish the extent of the open class problem for lexicographic work, the present practices, and the needs and expectations for language technology, we have conducted and carried out a metalexicographic survey with questions such as: Are word classes even a necessary and useful concept in modern lexicography? How challenging is word-class categorization for Estonian lexicographers? What are their actual expectations for the possibilities of the digital era in that respect?

First, we provide a background to the study in Section 2 presenting a short summary of the general traits of Estonian along with its most recent word class systematization and the treatment of word classes in some Estonian dictionaries (subsection 2.1). In subsection 2.2 we explain the setup of the semi-structured interviews with lexicographers. Section 3 focuses on the delineation and analysis of the data concerning our first research question, i.e. the problems lexicographers experience in connection with word classes. The second main question of this study, the solutions for aiding the lexicographer in categorization and presentation of word classes, is addressed in Section 4. The results are then summarized in Section 5.

2. Background and details of the study

2.1 Estonian and its word class system

Estonian is a Finno-Ugric language spoken by about 1 million people in the Estonian Republic and abroad. Although Uralic languages are considered agglutinating, Estonian morphosyntax is generally more fusional and analytic than that of the northern branch of Finnic languages (Finnish, Karelian, Veps etc.), which are characterized by a high degree of allomorphy and grammatical syncretism (see Viitso 2007, Remes 2009). Estonian can be described as a morphologically rich language: words inflect (nouns and adjectives for number and case; (finite) verbs for mood, tense, person and number; adjectives and adverbs for degrees of comparison) and are subject to agreement. There are approximately 100 native derivational suffixes, and new words are productively formed by compounding (Kerge, 2016: 3228). Nouns and adjectives decline for 14 morphological cases; nominative, genitive and partitive are traditionally considered grammatical cases and the remaining 11 cases are held to be semantic. For instance, spatial relationships are expressed by inner (illative, inessive, elative) and outer (allative, adessive, ablative) locative cases, besides adpositions. Verbs have, in addition to the abovementioned finite conjugational forms, infinitival, converbal and participial forms. A typical Estonian adjective normally agrees with its head noun in case and number (see (1)); a verb agrees with its subject in person and number (2):

(1) *kirju-de-st* *koer-te-st*
piebald-PL-ELA dog-PL-ELA

(2) *te* *jook-si-te*
you run-PST-2PL

The common categorization of word classes involves two main types: content words (typically nouns, verbs, adjectives, and adverbs) and function words (adpositions, pronouns, conjunctions, etc.). The criteria of categorization are generally based on morphosyntactic properties, but cross-linguistically these can only be identified semantically (Haspelmath, 2001). There are diverse approaches to the Estonian word classes, related to different language varieties and different methodological perspectives: see Kaalep et al. (2000) for contemporary written language and automatic morphological tagging, Habicht et al. (2011) for old written language, Lindström et al. (2006) for dialectal language and Hennoste (2002) for (contemporary) spoken language. The latest general word class system for Estonian proposed by Erelt (2017: 58–61) divides Estonian words into four main classes based on syntactic and semantic criteria[2]:

1. autonomous content words (verbs, nouns, adjectives, numerals and adverbs) that occur independently in a phrase and convey their denotative meaning obvious without context,
2. autonomous functional or substitution words (pronouns, proadverbs)
3. non-autonomous functional words or auxiliaries (auxiliary verbs, affixal adverbs, adpositions, conjunctions),
4. syntactically independent pragmatic words or particles (modal adverbs, interjections).

Word class can be seen as a link between grammar and lexis, providing a hint to a word's general meaning, its paradigmatic (morphological) behaviour and sentential function. It also gives the non-native user an idea about the uses of a particular word in the language and what patterns of grammar it should follow. In the Estonian lexicographic tradition, word class information is part of the description of a word's lexical behaviour, but PoS tagging is somewhat sporadic and this task is (implicitly) assigned to certain dictionaries, not all. Tagging all words with a PoS label is complicated, as since it is being a morphologically rich language Estonian is characterized by a tendency where inflected word forms may shift their lexical categorial status in respect to the base word.

There is a tradition of presenting PoS information in some of the general monolingual dictionaries (e.g. *The Explanatory Dictionary of Estonian* (EKSS, 2009), *The Dictionary of Estonian* (DicEst)[3] and *The Dictionary of Estonian Word Families* aim at systematic PoS markup) and in particular, in learner's dictionaries (e.g. *Collocations*

[2] Morphologically, the Estonian words fall into inflected (verbs, nouns and adjectives) and uninflected words (particles).

[3] EKSS and DicEst cover word class information over the Estonian lexis most comprehensively, however, even these dictionaries do not tag all headwords with word class information.

Dictionary (ECD)[4], and *The Basic Estonian Dictionary* (BED)) as well as bilingual dictionaries. As a rule, PoS is not marked in orthographic[5], onomastic, terminological, dialect, or etymological dictionaries. For instance, *The Dictionary of Standard Estonian* (DSE) marks word class traditionally only in certain exceptional cases relevant for advisory purposes. The question of word classes has become more topical along with the development of the Ekilex database and dictionary writing system, an integrated lexical resource, where PoS belongs to the structure of every lexical entry (Tavast et al., 2018). In the most recent output of the lexicographic resources, the language portal Sõnaveeb[6] ('Wordweb, 2019'), the explicit marking of word classes is an ultimate goal.

Regarding the technical side of word class marking, there are two parallel sets of labels for word classes in the Estonian lexicographic tradition – one using loanwords of international origin (e.g. *substantiiv* ´noun´) and the other using coined Estonian terms (e.g. *nimisõna* 'lit. name word'), which are more transparent. The two sets of terms basically address different users: the international terms are for experts and the transparent ones for learners. Most dictionaries use abbreviations of international terms (*v* for verbs, *adv* for adverb etc.), but the language portal Sõnaveeb and BED use non-abbreviated native Estonian terms.

2.2 Details of the study

2.2.1 Methods

To clarify the opinions and experience of professional dictionary-makers about the role of word class in their everyday work, we conducted a metalexicographic survey in the form of semi-structured oral interviews containing both open-ended questions and opinion ratings on a Likert-type scale. The target group were lexicographers working with the Estonian language in monolingual or multilingual dictionaries; the participants were informed beforehand about the general topic of the survey (the challenges of parts of speech categorization in their lexicographic practices). The interviews were carried out by three interviewers in February and March 2019 and lasted approximately 30 minutes each. All the conversations took place privately at the lexicographers' work environment. The conversations were taped, transcribed and analysed content-wise (searching for qualitatively different opinions). The numerical data were subjected to simple scoring.

[4] In the ECD, displaying the collocational behaviour of the 10 000 most frequent words in Estonian, PoS tagging has crucial relevance as the grouping of collocates is based on their word class affiliation.

[5] The paradigmatic affiliation of the entries is traditionally indicated by inflectional types (*muuttüübid*) marked by numerical indices.

[6] https://sonaveeb.ee/ (25.5.2019).

2.2.2 Respondents

Altogether 23 lexicographers (F=21, M=2) participated in the survey. For comparison: in the cross-European survey of lexicographic practices only 8 of the Estonian lexicographers participated, which – in the context of that study – was a rather high rate (Kallas et al., 2019: 7). We suppose that the collegial one-to-one setting and oral form of the survey facilitated participating in our study. The majority of our respondents were current employees of the Institute of the Estonian Language, the institution producing and publishing most of the academic dictionaries in Estonia. Only a few respondents were from Tartu University or some other institution.

The lexicographers' work experience varied from 0.5 to 48 years, averaging at 18 years. More than 10 (48%) of them had worked in this field for at least 20 years. The European lexicographer's average work experience is approximately the same, with a slightly smaller proportion of professionals (35.6%) having more than 20 years´ experience. It was pointed out in the cross-European survey that the profession of a lexicographer tends to be a lifelong one (Kallas et al., 2019: 8).

The experience of our respondents was also impressive in terms of content; and altogether 38 different dictionaries were mentioned as their past or current projects. The variety of dictionaries included both monolingual and bilingual dictionaries, both standard Estonian and dialects, both descriptive and prescriptive ones, among others. Altogether 22% of the respondents had some experience with general monolingual dictionaries (such as EKSS, DicEst, the *Dictionaries of Standard Estonian*, the *Basic Estonian Dictionary*, the *Dictionary of Word Families*, the ECD etc.); 26% had worked with specific monolingual dictionaries, such as the *Dictionary of Estonian Dialects*, *Estonian Etymological Dictionary*, *Low German Loanwords in Estonian*, etc). The smallest proportion (4%) had only worked on bilingual dictionaries, such as Estonian-French, Estonian-Finnish and Finnish-Estonian, Russian-Estonian, German-Estonian, etc. Almost half (47%) of our respondents had experience with multiple types of dictionaries.

Considering the length and range of the working experience, it is quite remarkable that most of the interviewees (83%) had worked or were currently working with an electronic dictionary writing system (mostly the institute's own in-house software EELex). The percentage of lexicographers using corpora (and specific tools for corpus search, such as Sketch Engine) was 74%. While many lexicographers mentioned using the Estonian National Corpus, some other more specific corpora were also mentioned, such as the Corpus of Old Literary Estonian, the Educational Corpus of Estonian etc. Some of the lexicographers reported using a corpus that had been specifically designed for compiling the dictionary at hand. In comparison with European lexicographers, on average, our respondents' use of IT resources and tools was 10% higher (see Kallas et al., 2019).

In conclusion, the interviewed lexicographers have been quite flexible to adjust to the rapid changes in the field of the lexicographers' workflow while the institutions have

provided good technological support. This is the background for the lexicographers' expectations for IT resources and tools: a long experience as a lexicographer and a certain degree of familiarity with the affordances that the electronic era, in principle, could provide make the lexicographers wish for even better tools and technological support.

3. The challenge(s) of word classes in lexicographic work

3.1 Can we manage without word classes?

As a point of departure, we focus on the general role of word classes in our respondents' everyday tasks. The interviewees were encouraged to reflect on the necessity of PoS markup and on whom it benefits. The respondents generally considered the task of PoS markup rather important and beneficial. All the interviewed lexicographers agreed that the PoS information is presented for "the user". Some lexicographers emphasised the role of PoS for a regular user (pupils, language learners, teachers) and took into account their restricted ability to cope with the overwhelming lexicographic information, while the rest took the perspective of expert users (linguists, lexicographers) and provide information as detailed as possible, "because it is in the interests of the researchers". Lexicographers saw themselves among the potentially beneficiary parties of the PoS information. If a dictionary has already been PoS-tagged, a professional has analysed the material, and the earlier work of the colleagues needs to be (re)valued.

The respondents ranked the necessity of the word class information in the dictionary they currently worked with on a 5-point scale. As the response rates in Figure 1 show, the results varied from "very necessary" to "somewhat unnecessary", while none of the lexicographers rated it "completely unnecessary":

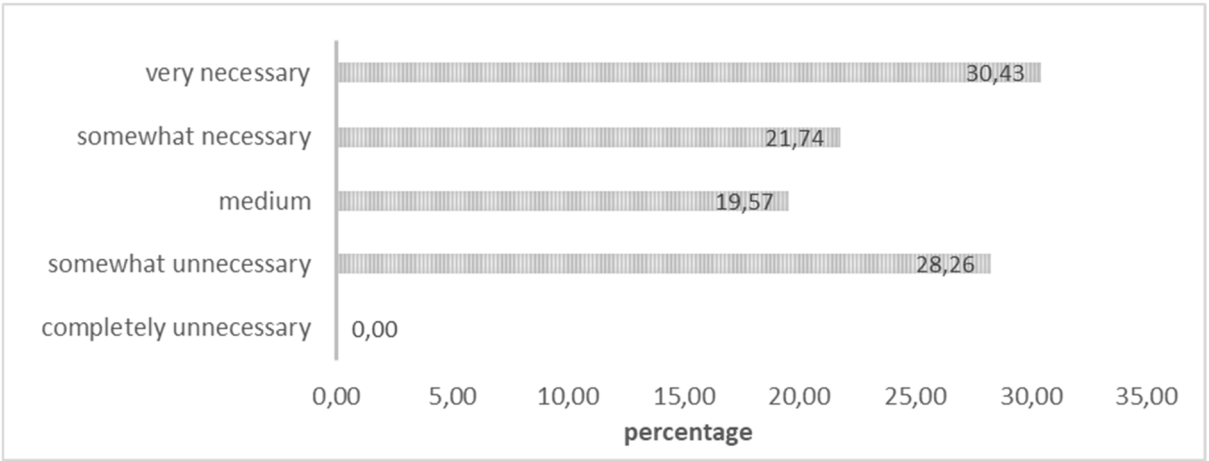

Figure 1: The necessity scale.

There are two peaks in the diagram: 30% of the respondents claimed that PoS marking is very necessary and 26% considered it somewhat unnecessary. This polarization of opinions can be explained by variation in the representation of word classes (mandatory or occasional) in different dictionaries, motivated by the assumed needs of the target groups. As mentioned in Section 2.1, the PoS tagging is also determined by tradition and a certain division of labour. Although the necessity rating of PoS information was primarily based on the respondents' ongoing projects, the previous experience seemed to influence the assessments. In addition, word classes seem to provide a general logical structure for organisation of the material (they "help to categorize the material in one's mind").

The lexicographers' reflections reveal the relevance of word classes in the systematization of the material and a need for as clear criteria of classification as possible. Since word class categorisation involves different linguistic levels, the balance may be swaying towards one or another of them. Some of the respondents emphasised the relevancy of semantics in the specification of a word's categorial affiliation: "I definitely take meaning into consideration, because semantic features define the essence of the word". Other respondents base the judgment on syntactic properties, considering a word's typical function in a sentence. Problems arise from the classical description of, for example, the noun as the argument of a clause – whenever a noun tends to occur in another function, for instance as an adverbial, its syntactic properties (and thus the word class attributes) will change. The respondents consider morphology to be the main source of the word forms departing from a paradigm, seen as a special characteristic of morphologically rich languages. Although the resulting ambiguous cases complicate information retrieval in databases, they also reveal ongoing lexical changes. Yet the respondents engaged in ascertainment of word class boundaries on a daily basis would still prefer an "ideal" situation where every word has a definite word class label.

3.2 Different dictionaries – different challenges

To a great extent, the problems faced depend on the properties related to the dictionary type the lexicographer is working on – its object of description, purpose, target group and other factors.

The lexicographers working with bilingual dictionaries generally use a database of Estonian that contains the most frequent words with word classes already defined (the Estonian-X dictionary[7]). The respondents belonging to this group generally assessed word class categorisation as not a too complicated task. The interviewees working with general and specific monolingual dictionaries (see the distribution of dictionaries the respondents work with in Section 2.2) are not that unanimous. For the most part, the dictionaries in the general monolingual group require explicit formulation of the word

[7] The database is available at http://exsa.eki.ee/exsalogin.cgi (25.5.2019).

class, except for the DSE. The lexicographers experiencing particularly challenging problems with PoS tagging work either with the DicEst, ECD (general dictionaries), or the dictionary of old written Estonian (a specific dictionary). The impression of interviewers is that the lexicographers compiling the dictionaries that provide a systematic markup of word classes manifest a particularly deep sense of responsibility, as their work results will be source material for other lexicographers. In other dictionaries labelled as specific in our study, the word class category is generally not a prominent issue, even though it may be a topic puzzling the lexicographer in the background.

For instance, in compiling an etymological dictionary the word class category is not the primary concern, as the main focus is on the origin of a word stem and words with the same stem are gathered in a same entry. Hence, the derivative relations appear as most important; in case there are doubts regarding a word's root form, the most plausible variant is preferred and a word class suggestion often appears in the definition of the word. The lexicographers compiling the dictionary of old written Estonian handle specific problems such as different stages of lexicalization-grammaticalization compared to contemporary language and a rather limited availability of linguistic sources, which complicates the determination of the developmental stages of a word and hence also the determination of the word class. The compilation of dialect dictionaries involves analogical tasks compared to the etymological and old written language ones, but the work has its own logic, since the object of description basically originates in colloquial spoken language and data collections based on fieldwork.

The lexicographers were invited to assess the challenge of the task of word class categorisation among the other tasks they perform in their everyday occupation with dictionary compilation on a 5-point scale: "easy", "pretty easy", "medium", "challenging", and "very challenging". The assessments followed the normal distribution with the peak of 56% at the point "medium" and 22% on both "pretty easy" and "challenging". None of the respondents used the extremes of the scale. The results reflect the factors related to the somewhat different challenges of the compilers of different types of dictionaries and those related to the ambiguity of certain forms, which will be discussed in more detail below.

3.3 The natural flux of word classes in Estonian and its implications for lexicographic work

A characteristic feature of Estonian is that the inflected word forms tend to move from their basic lexical categorial status to another. For instance, the boundaries between adpositions, nouns and adverbs in Estonian are considered to be rather fuzzy (see, for example, Grünthal, 2003), and there are always words and word forms in a transition stage, appearing both as standard nouns and as part of more or less fixed expressions with more abstract meanings (see, for example, Paulsen, 2018, 2019). The natural flux of words from one word class to another – detectable in changes in their

syntactic/pragmatic function and, occasionally, in a shift of meaning – was reflected in the reasoning of our interviewees, too.

Lexicographers are trained to recognize words by their word class membership and in most instances it is not a critical concern. In less self-evident cases, for instance when the actual usage of the word (or its form) in the corpus shows idiosyncratic tendencies, the lexicographer may be in a difficult position. The respondents were encouraged to bring up examples of some particularly striking cases. Almost all of them could think of such examples, the total number of tokens being 145. The number of different examples (types) was 127. The average number of critical examples per person was six, which falls well within the limits of one's short-term memory. In reality, some of the respondents had prepared for the interview and brought up more examples; four respondents declared that they either did not have any considerable problems with word classes or they could not remember the exact problems.

Figure 2 presents the distribution of the examples across the "classical" word classes. The thicker the line in the figure, the higher the proportion of the examples falling between the two categories located at the ends of the line. The noun sits in the centre of the diagram because it has the highest proportion of "overlapping" cases: altogether 35% of the total number of examples enjoyed "dubious" membership with this category (and with adverbs, adpositions and adjectives, respectively). Adjectives appeared as the second most "slippery" word class, with 26% of the total number of critical examples.

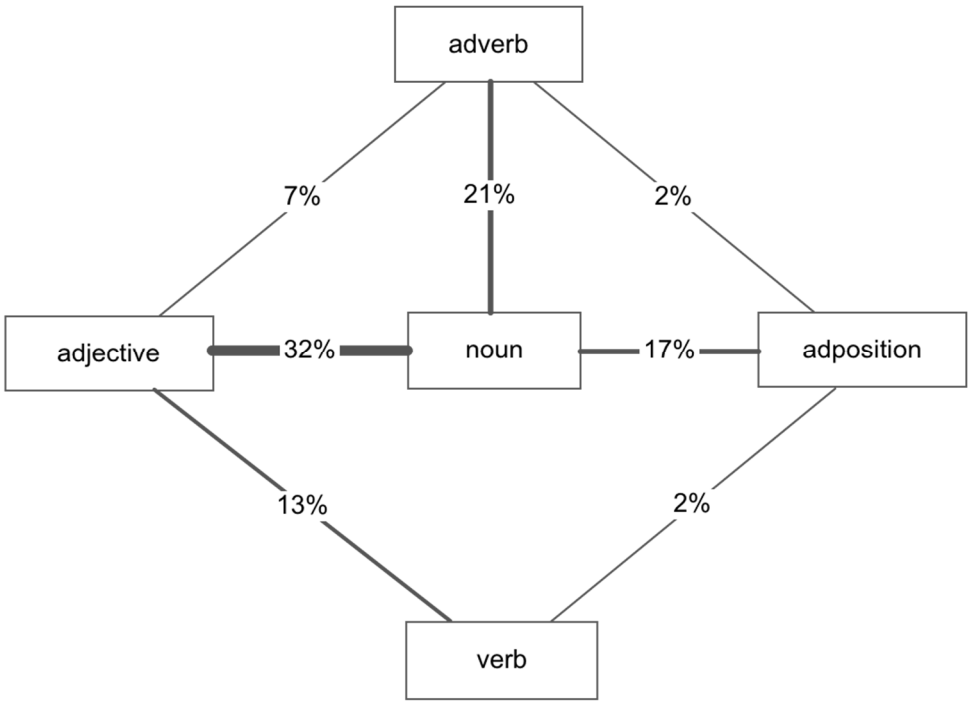

Figure 2: Distribution of the critical examples along their word class membership.

As a generalisation, we can say that the lexemes the most complicated to define either i) occur in two or more lexical classes keeping the same base form (as examples (3) and (7) below), ii) are in a transition phase from one word class to another (e.g. the inflected forms developing new functions, see examples 4–6) or iii) only appear in certain inflected forms (8). From a nominal perspective, the main source of the word class shift lies in the semantic cases, especially the locative case forms of nouns that develop autonomous uses both semantically and syntactically (see example (4)). The main source of categorial shift for verbs are infinitives/converbs (5), participles (6), and nominalisations (*ela-mine* [live-NOM] 'habitation').

(3) S → ADJ
Koer poiss roni-b puu otsa
dog boy climb-3SG tree.GEN tip.ILL
'The naughty boy is climbing the tree'

(4) S → ADV
Mu-l on tema-st kõri-ni
I-ADE be.3SG he-ELA throat-TERM
'I have had enough of him / I'm fed up with him'

(5) S → ADP
Pole riigi huvi-des makse alandada
do.not state.GEN interest-CONV tax.PL.PART reduce
'it is not in the interest of the state to lower the taxes'

(6) V → ADJ
Ta and-i-s töö-le hävita-v-a hinnangu
he give-PAST-3SG work-ALL destroy-PTCP.GEN judgement.GEN
'He gave the work a devastating assessment'

(7) ADV → ADJ
a. *Õpetaja on alati abivalmis.*
teacher be.3SG always help.ready
'The teacher is always helpful'
b. *Mõned on abivalmi-m-ad kui teise-d*
some be-3PL help-ready-COMP-PL than other-PL
'Some are more helpful than others'

(8) ?
Üritus toimus noor-te eestvõtte-l
event happen-PAST-3SG youth-PL.GEN front.grasp-ADE
'The event took place on youth initiative'

How do lexicographers solve the puzzle of classifying ambiguous words? The reported strategies were lookup in other dictionaries, checking the grammars, consulting relevant research, using syntactic tests, and looking at the distribution of the given word in a corpus. If these measures do not suffice, they turn to colleagues – after discussions and consideration of different possible perspectives, the team of lexicographers may decide the PoS markup collectively, by voting.

As for IT-solutions, the lexicographers mainly use the corpus query software Sketch Engine (Kilgarriff et al., 2014), particularly word sketches and concordance queries. Only a few of them were aware of the Sketch Engine function "Lempos" showing the distribution of a lemma in certain positions. Some respondents use automatic morphological analysis[8] to get an idea about possible alternative interpretations of the word. In short, only a fraction of the respondents use an IT solution in the decision-making process of PoS markup. The lexicographers confirmed repeatedly that they search for information, evidence and opinions, but the final decision about the PoS markup is up to them. There is no automatic PoS markup in the current practice of lexicographers, and they thought it would be almost impossible to use one, mostly because of the questionable reliability – "I will trust only myself as a researcher".

3.4 Multiple or zero tagging? Practical implications

The PoS categorisation and markup is not only of theoretical interest, as it has numerous practical implications on lexicographic reasoning. The lexicographer has to take a quick stand on the forms undergoing grammaticalisation or lexicalisation, fix the base word class in case a word belongs to different word classes within one morphological paradigm (like adjectives and nouns in Estonian) and consider the diverging opinions expressed in the linguistic literature. The approach to the word class affects the whole structure of a dictionary starting with the number and the organisation of entries. The PoS categorisation problem can, for instance, be solved by presenting the questionable form as a subheadword instead of a separate independent headword; the PoS tag of the subheadword can then be omitted (this is the solution used in the EKSS). This is a way to present the items that are (yet) not fully lexicalised or grammaticalized, indicating an ongoing change. However, as mentioned in Section 2.1, the development of the database and dictionary writing system Ekilex and the integrated language portal Sõnaveeb ('Wordweb') set completely new demands for the lexicographer, as the goal is to provide the PoS information for every lexical entry.

Most lexicographers agreed that it is acceptable and even inevitable that some headwords have two PoS tags, e.g. *haige* 'ill' (ADJ) and 'patient' (S). An argument for this was that the dictionary should reflect the actual usage: If the words tend to be used in different kinds of constructions typical of different PoS, then the dictionary must display it. It is also expected to facilitate comprehension for language learners by

[8] artur.eki.ee/morf (25.5.2019).

explicitly tagging the two possible ways of usage instead of having the users study the examples and make their own inferences.

Some respondents were more dubious about multiple tagging of the same headword and stressed that it can be accepted only in cases when the meaning of different parts of speech is "exactly the same". The typical example of such a case is *all* 'under, below', used either independently (as an adverb) or as part of a phrase (as a postposition). It was argued instead that they should be presented as separate headwords except when the cases are semantically strictly identical. Again, the motivation behind the one-to-one relationship in the description was "user needs".

The lexicographers agreed that the degree of specification of the PoS markup depends on the type and purpose of the dictionary. There was an opinion that everyone would benefit from at least one reliable source (a kind of "master dictionary") assembling the word class information. Ideas differed on how to deal with ambiguous words. Some lexicographers trust that every word can be classified, even if it seems difficult at the beginning. Others are less idealistic, and propose that sometimes it would be practical to present the questionable form not as a fully independent headword but as a subheadword without a special PoS tag (like the solution in EKSS discussed above). There are also lexical items other than words (idioms, multiword expressions, phrasal verbs) that could hardly be tagged for PoS. It was pointed out that there are other possibilities to demonstrate the usage of the word, such as by presenting examples. There was an agreement that in the vague cases a word cannot be classified properly without context. Another practical question concerning the corpus data was "What is the sufficient degree of frequency?".

4. Expectations for solutions

The second main research question of this study concerns the possibilities for facilitating the PoS-categorization task in lexicographic workflow. The lexicographers were asked about solutions they could think of when dealing with complicated cases of word class identification.

4.1 Could we just change the classification?

The system of PoS marking in a language holds as a part of the general agreement about the linguistic categories and no single lexicographer nor group of lexicographers can easily change it. The lexicographer must adopt the existing system and find reasonable practical solutions. Would the word class system need an adjustment into a more suitable one? This question has two possible answers: The classification can either be generalised and schematised into more heterogeneous groups, thus increasing the average number of class members, or the system can be elaborated by increasing the number of classes and creating specific labels for the classes of "ambiguous" cases with a more homogeneous class membership as a result.

The respondents presumed that the system could in principle be changed if it would match the actual usage and become more comprehensible for the user. They pointed out that dictionary-wise the word class labels vary anyway: Some dictionaries distinguish between prepositions and postpositions, while others use the more comprising term adposition; some dictionaries mark just adjectives, while others tag also its subclass of indeclinable adjectives etc. The EKSS uses 17 different tags for word classes because in addition to the traditional labels (noun, adjective, adverb, etc.) some specific ones have been created, such as abstract noun, diminutive, proper noun, (adjective-like) participle, actor noun, and action noun.

The attitudes towards potential changes differ notably. The lexicographers oriented to the needs of regular users (particularly learners) prefer a simple and elegant PoS markup: just a few word classes with transparent native terms. The respondents focusing on the needs of expert users are against losing the attained level of granularity and seek for continuous enhancement. They prefer the present system of PoS labels and are ready to welcome a more precise and detailed system, if justified. Some respondents are aware of the heightened need of precision for natural language processing applications and are therefore in favour of finer granularity. They admit, however, that not every detail known to the lexicographer or to the "system" needs to be presented to the regular user. In the case of an e-dictionary, an adjustable interface conforming to the needs of different users could be a solution.

Some of the lexicographers mentioned that a good system of PoS markup could be a hierarchical one containing both more general classes and the more specific ones (subclasses as well as subclasses of subclasses). Such a system would remind one of the general prototype model of human categorisation with its basic, superordinate and subordinate levels of knowledge (see Rosch et al., 1976). The present system of PoS in Estonian follows, in some respects, such a hierarchical model: The words are divided into inflected vs uninflected words, content vs function words, and further into specific classes with their specific combinations of meaning, form and function (see Section 2.1).

4.2 Visions of a PoS evaluator

The lexicographers were encouraged to share their conception of an ideal IT tool that would help them solve the ambiguous cases of word class affiliation. They expressed certain scepticism and even reluctance towards this idea – mostly because the respondents got the impression that they were expected to present a fully conceived technical solution in detail. However, some of them were disappointed with lexicographical IT tools in general. They shared a suspicion that no perfect tool would be possible, and envisaged themselves correcting the mistakes made by an automatic system. The respondents who were more aware of the technical nuances pointed out that no system can work better than the underlying automatic tagging (both morphological and syntactic) of corpora, and thus any result relying on the same tagged corpus would present results similar to those of Sketch Engine.

The ideas the lexicographers came up with can be divided into (structurally) simple and complicated ones. The relatively simple, not particularly corpus-driven solutions make helpful information easily available and facilitate the exchange of information among lexicographers:

1) **A database of ambiguous cases**, collecting the earlier (also divergent) lexicographic judgments with eventual reference to the corpus data the definitions rely on. The result would be like a "master dictionary", where the lexicographers can test their intuition or find analogical cases to base their judgments on. Such a solution requires a group of experts charting all ambiguous cases and making justified decisions about their PoS. Although the data can be updated and changed by the lexicographers themselves, it would be an off-line solution by nature – it would not refresh automatically when the corpus data are updated (illustrative examples can be added by lexicographers). The examples gathered in this study (see Section 3.3) can serve as a starting point for such a database, and these cases can also be used for extraction of similar cases from large corpora.

2) **A lexicalisation-grammaticalisation scale.** A word (form) should match a set of explicit criteria in order to get a certain PoS tag. The (grammatical, distributional) criteria would be included as a module in the lexicographers' workbench (EELex or now Ekilex). The problem with this solution is that it differs only a little from the lexicographers´ current task, saving their time and energy only by making the criteria easily accessible. The solution relies on the "classical" understanding of category membership (the necessary and sufficient conditions), and it is unclear whether it would produce sufficient solutions for the ambiguous cases that share the criteria of many classes or lack some necessary condition of the main class.

3) **A set of smart syntactic tests to "try out" the PoS membership.** Lexicographers use "testing it mentally" in their everyday practice. For example, if an ambiguous participial form is agrammatical in the comparative form or in a phrase with the intensifier *väga* 'very', there is a question of a verb form rather than of an adjective. This kind of test could help the lexicographer to make a proper decision about the PoS. Such a solution requires a group of experts to refine the system of adequate tests. The task of PoS evaluation would be facilitated, but the decision relies on the lexicographer's grammaticality judgment of composite phrases.

The main idea of a more advanced PoS evaluator is a corpus-driven tool that searches the corpus and presents the (up-to-date) statistics of the morphosyntactic distribution of an ambiguous form on the lexicographer's desktop. The behaviour of a questionable word would be compared to the corresponding profiles of the typical members of different PoS and the percentage of overlap and discrepancy would be revealed. The

prototypical PoS profiles (in terms of syntax, morphology, semantics) should first be established in the corpus data. The respondents prefer a visualised output with an indication of the dominant PoS profile and the degree of predominance. The tool should generalise over the results, suggest qualitative distinctions, and provide access to the original data the statistics is based on (concordances with a gateway to the context). The raw material should be presented according to its relevance, showing explicitly which criteria of the particular PoS are satisfied and which are not. Statistics about the presence of a semantic shift would also be welcome. Basically, the tool should be similar to Sketch Engine but even more advanced and reliable.

There were different ideas about the scope of the task that the PoS evaluator should perform. There is no need for such an application for the typical "well-behaving" word forms. The respondents imagine an application providing a desktop window where one can insert the search term and receive its statistics and tendencies related to a PoS. Presuming such a rather narrow task, the other steps of the lexicographic workflow would remain the same. Some of the lexicographers came up with a broader view of the task: The tool would analyse the corpus for "suspicious" word forms (e.g. nouns that appear mostly or only in locative case forms), create a list of the potential new headwords and then analyse them in detail, according to the lexicographer's choice. Such an automated procedure would draw the lexicographers' attention to certain changes in usage that would otherwise remain unnoticed.

The respondents would prefer to have the tool as a module in their habitual work environment, either as part of their workbench (EElex, Ekilex) or as part of the corpus searching tool (e.g. Sketch Engine). Some of the lexicographers envisaged that the PoS evaluator would be useful not only for lexicographers, but also for the general public. In that case an application with a simplified interface is needed – the information served to a language learner should not be too abundant or complicated.

How to arrive at such a system is a task for the future. The aspects of knowledge, mentioned in relation with the "simple solutions" (a database of critical cases, a scale of explicit criteria, a set of discriminative tests), will be useful sources of information also when striving for an automated PoS markup.

5. Conclusion

This study aims to grasp the lexicographers' experiences and visions regarding word class categorisation and to relate these ideas to the changed paradigm of modern lexicographic work. PoS categorisation is a topical issue in Estonian lexicography, as the current trend is to avoid omitting tags as well as the multiplicity of PoS markup. The ultimate aim is to provide a word class tag for every dictionary entry in the main database (regardless of whether the end-product contains or displays the PoS tags). This trend is dictated by the data model of the Ekilex database and dictionary writing system and the design of its main output, the language portal Sõnaveeb.

The results of the survey indicate that in lexicography word classes provide a categorial frame that is in the background, even if PoS tagging is not an explicit task in the dictionary a lexicographer works with. Changing the word class label of a word is a long-term process and the changes are not made easily; the lexicographer has to take into account the fact that every decision may add new boundaries and ambiguous spots. Is it necessary to take a more flexible approach to lexical category membership (see also Smith 2015)? What if all words cannot be PoS-tagged?

The first research question our study focuses on is the problems and practices of lexicographers dealing with PoS categorisation problems. Three issues were pointed out as the linguistically most problematic: the lexemes that i) occur in two or more lexical classes keeping the same base form, ii) are in a transition phase from one word class to another, or iii) only appear in certain inflected forms. Morphology was considered the main reason for the word forms departing from a paradigm. Regarding the possible reformation of the current Estonian word class system, opinions diverged: Considering the needs of regular users, a more general system was seen as preferable, but for the expert users a more fine-grained system was preferred. As an "applied approach" to word classes, the idea of a flexible display (applicable to a lexicographic root-database and dictionary writing system like Ekilex) emerged, taking into account both the needs of dictionary users and those of the experts.

The main concern is how to make well-grounded decisions based on the deluge of linguistic material. All in all, the lexicographers consider numerous aspects of their work but are also open to innovative solutions if they see the advantages. The respondents actually working with word class identification expressed a need to improve the factors influencing the process of PoS tagging, but also a certain scepticism towards an "ideal machine" that would be able to solve the categorisation issues characteristic of natural languages.

This leads us to the second focus of this study: the expectations lexicographers have with regard to modern technology-related solutions. We can conclude that despite a grain of scepticism, the lexicographers would welcome a reliable software solution to ease the decision-making process. In general, there is indeed a need for an improved morphological and syntactic parsing system, as well as for detection of changes in words' semantic behaviour, and the latter is perhaps the most difficult to achieve. The solution would be a corpus-driven application presenting the statistics over the morphosyntactic distribution of an ambiguous word with access to the data source.

All in all, the lexicographers share an acute sense of responsibility related to the PoS judgment. They show remarkably high levels of empathy by having in mind both the regular user (or its conception), when making their proposals, e.g., an application of a possible technological tool with a simplified interface, and colleagues, when conceptualising the applied database assembling the judgments on difficult phenomena. Moreover, the potential PoS evaluator was considered useful not only for lexicographers

but also for the general public.

6. Acknowledgements

This work was supported by the Estonian Research Council grant PSG227.

7. Abbreviations

3 = third person; ALL = allative case; COMP = comparative; CONV = converbal; GEN = genitive case; ILL = illative case; PART = partitive case; PTCP = participle; PL = plural; SG = singular; TERM = terminative; TRA = translative.

8. References

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective.* Oxford: Oxford University Press.

Culicover, P. W. (1999). *Syntactic Nuts.* Oxford University Press: Oxford.

Erelt, M. (2017). Sissejuhatus süntaksisse [Introduction to syntax]. In M. Erelt & H. Metslang (eds.) *Eesti keele süntaks.* Tartu: Tartu Ülikooli Kirjastus, pp. 537–564.

Habicht, K., Penjam, P. & Prillop, K. (2011). Sõnaliik kui rakenduslik probleem: sõnaliikide märgendamine vana kirjakeele korpuses ['Parts of speech as a functional and linguistic problem: annotation of parts of speech in the corpus of Old Written Estonian']. *Estonian Papers in Applied Linguistics*, 7, pp. 19–41. https://doi.org/10.5128/ERYa7.02

Haspelmath, M. (2001). Word classes and parts of speech. In P. B. Baltes & N. J. Smelser (eds.) *International encyclopedia of the social and behavioral sciences*, pp. 16538–16545. Amsterdam: Pergamon.

Hennoste, T. (2002). Suulise kõne uurimine ja sõnaliigi probleemid. In R. Pajusalu, I. Tragel, T. Hennoste & H. Õim (eds.) *Teoreetiline keeleteadus Eestis*, pp. 56–73. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Tartu: Tartu Ülikool.

Kaalep, H.-J., Muischnek, K., Rääbis, A. & Habicht, K. (2000). Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? ['Do the available morphological descriptions of Estonian work on a real text?']. *Keel ja Kirjandus*, 9, pp. 623–633.

Kallas, J., Koeva, S., Kosem, I., Langemets, M. & Tiberius, C. (2019). *Lexicographic practices in Europe: a survey of user needs.* Accessed at: https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf (08 June 2019)

Kerge, K. (2016). Word-formation in the individual European languages: Estonian. In P. O. Müller, I. Ohnheiser, S. Olsen & F. Rainer (eds.) *Word-Formation. An International Handbook of the Languages of Europe.* Vol. 5. Handbooks of Linguistics and Communication Science 40. Berlin, New York: De Gruyter, pp. 3228−3259. https://doi.org/10.1515/9783110424942-009

Lindström, L., Bakhoff, L., Kalvik, M.-L., Klaus, A., Läänemets, R., Mets, M., Niit,

E., Pajusalu, K., Teras, P., Uiboaed, K., Veismann, A. & Velsker, E. (2006). Sõnaliigituse küsimusi eesti murrete korpuse põhjal. In E. Niit (ed.) *Keele ehe. Tartu Ülikooli eesti keele õppetooli toimetised* 30. Tartu: Tartu Ülikool, pp. 154–167.

Paulsen, G. (2018). Manner and adverb: Fuzzy categorial boundaries in collocations. *Estonian Papers in Applied Linguistics,* 14, pp. 117−135. doi:10.5128/ERYa14.07

Paulsen, G. (2019). Sõnaliigipiiridest kollokatsioonide vaatenurgast: erikäändelised noomenadverbid [Word class boundaries and collocations: The Estonian nominal adverbs in special cases]. *Estonian Papers in Applied Linguistics*, 15, pp. 121−137. doi.org/10.5128/ERYa15.07.

Remes, H. (2009). *Muodot kontrastissa. Suomen ja Viron vertailevaa taivutusmorfologiaa. Acta Universitatis Ouluensis Humaniora B 90.* Oulu, Oulun yliopisto.

Rosch, E., Mervis, C. B., Gray, W., Jason, D. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, pp. 382–439.

Smith, M. C. (2015). Word categories. In J. R. Taylor (ed.) *The Oxford Handbook of the Word.* OUP Oxford: Kindle Edition.

Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17–21 July 2018.* Ljubljana University Press, Faculty of Arts, pp. 749−761.

Viitso, T.-R. (2003). Rise and Development of the Estonian Language. In M. Erelt (ed.) *Estonian Language* (Linguistica Uralica Supplementary Series 1.). Tallinn: Estonian Academy Publishers, 130–230.

Dictionaries:

BED: *Eesti keele põhisõnavara sõnastik* [*The Basic Estonian Dictionary*]. (2014). Kallas, J., Tiits, M., Tuulik, M. (eds.); Jürviste, M., Koppel, K., Tuulik, M. (compilers). Tallinn: Eesti Keele Sihtasutus. https://sonaveeb.ee

DicEst: *Eesti keele sõnaraamat* [*The Dictionary of Estonian*]. (2019). Langemets, M., Tiits, M., Uibo, U., Valdre, T. & Voll, P. (eds.); Kuusik, K., Kuusk, K., Langemets, M., Tiits, M., Uibo, U., Valdre, T. & Voll, P. (compilers). Institute of the Estonian Language. http://www.sonaveeb.ee

DSE: *Eesti õigekeelsussõnaraamat* [*The Dictionary of Standard Estonian*]. (2018). Raadik, M., Erelt, T., Leemets, T. & Mäearu, S. Tallinn: Eesti Keele Sihtasutus. http://www.eki.ee/dict/qs/

ECD: *Eesti keele naabersõnad* [*The Estonian Collocations Dictionary*]. (2019). Kallas, J., Koppel, K., Paulsen, G. & Tuulik, M. Institute of the Estonian Language. http://www.sonaveeb.ee

Eesti keele sõnapered. Tänapäeva eesti keele sõnavara struktuurianalüüs [*The*

Dictionary of Estonian Word Families. A structural analysis of the contemporary Estonian lexis]. Vol. I-II. Vare, S. (2012). Tallinn: Eesti Keele Sihtasutus.

EKSS: *Eesti keele seletav sõnaraamat* I–VI [*The Explanatory Dictionary of Estonian*]. "Eesti kirjakeele seletussõnaraamatu" 2., täiendatud ja parandatud trükk. Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll (Toim.). Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus, 2009. http://www.eki.ee/dict/ekss/

Sõnaveeb [Wordweb]. (2019). The Language Portal of the Institute of the Estonian Language. Hein, I., Kallas, J., Koppel, K., Langemets, M., Männiko, K., Nurk, T., Viks, Ü., Laubre, M., Ukkivi, R., Tavast, A., Lastovets, S. & Rautam, S. (eds.).

Corpora:

Kallas, J., & Koppel, K. (2018, March 26). *Eesti keele ühendkorpus 2017* [*The Estonian National Corpus*]. Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-071E7L

Corpus of Old Written Estonian. Prillop, Külli. (2013, January 9). *Vana kirjakeele korpus.* [*Corpus of Old Written Estonian*]. Center of Estonian Language Resources. https://doi.org/10.15155/1-00-0000-0000-0000-00075l

Kallas, J. & Koppel, K. (2018, April 23). *Estonian Corpus for Learners 2018 (etSkELL).* Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-07335l.