Catching lexemes. The case of Estonian noun-based ambiforms

Geda Paulsen[1,2], Ene Vainik[1], Ahti Lohk[1], Maria Tuulik[1]

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia

[2] Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden

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

Abstract

The aim of this study is to test a statistic relying on corpus data, the distributional index (D-index): a statistical benchmark that helps lexicographers judge if a morphological form has been conventionalised to the degree of becoming an independent lexeme. Our focus is on the decategorisation type that originates from a case form of a noun and is directed to an adverb, adposition or adjective. The words or inflected forms corresponding to more than one word class interpretation are in this study termed ambiforms. The analysis compares the D-index levels of ambiforms categorised as nouns and another PoS. The results suggest that for the outcome to be most authentic, the noun-based ambiforms should be analysed without the decategorisation influence, i.e. the D-index analysis should be applied in the pre-PoS-disambiguation stage.

Keywords: form distribution; morphology; lexicography; language technology; Estonian

1. Introduction

An electronic dictionary striving to depict contemporary vocabulary needs to be updated constantly due to the changes that take place in the actual usage of language. Estonian lexicography is developing towards unification of lexical resources (dictionaries and term bases) into a central “super-dictionary”, the EKI Combined Dictionary (CombiDic), with the Ekilex dictionary writing system as its backbone, and lexicographic processes are moving towards a higher degree of automation. (About the recent developments regarding Estonian lexicographic resources, see Tavast et al., 2018; Tavast et al., 2020; Kallas et al., 2020.) Besides monitoring the most recent corpora for neologisms (Langemets et al., 2020), tracking and identifying the degree of grammaticalisation and lexicalisation of existent word forms are essential to attain an adequate overview of language development.

To be able to make a well-grounded decision about a new lexeme candidate, lexicographers need more fine-grained processing of corpus data than simple word frequencies (Paulsen et al., 2019). Blensenius & Martens (2019) argue for the use of word-form relative frequency information derived from existing corpora to improve dictionary content. When it comes to tracking morphological decomposition processes, Hay (2001) states that relative frequency is more elucidative than absolute frequency.

As a solution for capturing decategorising noun forms in Estonian, we suggest a specific statistic predicting a form's degree of salience: the distribution index (D-index). The D-index (DI) calculates the distributional value of nominal case forms as compared to the norm-based relative frequencies of the case forms (Vainik et al., 2021). The aim of this study is to ascertain whether the D-index enables one to detect forms emerging as potentially independent lexemes.

This article is the second report on our ongoing study of the D-index. In our earlier paper, we described the development of the index and tested it on a sample (N = 46) of Estonian noun-based ambiforms (words or inflected forms corresponding to more than one word class interpretation) in 11 (semantic) cases (Vainik et al., 2021). The results were compared to a control group of "ordinary" nouns (N = 26) with an abundant range of case forms displaying a regular distribution of case form frequencies. As a result of this study, we determined the threshold value of the distribution index as an indicator of heightened frequency.

In the present study, we tested the threshold value on a selection of noun-based declined forms that can be expected to be situated at some point in the decategorisation process. Our focus is hence particularly on morphology-based PoS change, i.e. the decategorisation type that originates from a case form of the noun and is directed to an adverb, adposition or adjective (for more about possible PoS combinations in Estonian, see Vainik et al., 2020)[1].

The data for the analysis of noun-based ambiforms were derived from the database of Estonian ambiforms, consisting of approx. 3,500 examples (see Vainik et al., 2020). We will calculate the D-indices of the selected noun-based ambiforms and consider the usability prospects of the distributional identification of case forms. Our main research questions are: Does the threshold of heightened frequency (Vainik, Paulsen & Lohk 2021) capture a form's movement to the status of an independent lexeme? Is it possible to establish other thresholds? What is the impact of corpus preprocessing on the results, i.e. automatic morphological tagging and PoS disambiguation, proceedings that are supported with data from the CombiDic? Can the D-index help to improve corpus tagging systems?

We will begin with a short overview of Estonian nominal morphology and the decategorisation processes related to case endings in Section 2. The methods and data used in the study – the D-index and its calculus, the corpus processing methods, and the data and data processing procedures – are explained in Section 3. Section 4 is devoted to the analysis and discussion of the DI levels of ambiforms with different lexicographic statuses and the effects of the principles of corpus annotation on DI calculations. Section 5 summarises and discusses the results.

[1] The operating of the D-index in practice is described in detail in Vainik, Lohk & Paulsen (2021, this issue).

2. The Estonian case system and inflectional decategorisation processes

In terms of their morphological behaviour, Estonian words can be divided into four main classes: (1) words that can be inflected for mood, time and person (verbs), (2) words that can be inflected for all cases (nominals), (3) words that have no grammatical case forms (some adverb types and some adpositions), and (4) words that have no inflectional forms (some adverb types and adpositions, conjunctions and interjections (Viitso 2003, 32). The Estonian nominals, i.e. nouns, adjectives, numerals and pronouns (and certain participles and infinitives) are inflected for number (singular (SG) and plural (PL)) and case. The semantic cases have functions similar to prefixes or suffixes in many other languages (ibid.). There are three grammatical cases – nominative (NOM), genitive (GEN) and partitive (PART) – and 11 semantic or adverbial cases: illative (ILL), inessive (INE), elative (ELA), allative (ALL), adessive (ADE), ablative (ABL), translative (TRA), terminative (TER), essive (ESS), abessive (ABE) and comitative (COM). (Ibid 32)

In Estonian, the decategorisation processes involving morphological forms are a considerable source of word-class fluidity: common nouns in a (usually semantic) case form may undergo PoS-shift into function words (mainly adverbs and postpositions). The development of nominal case forms into adverbs (or adpositions) is a characteristic feature of Estonian (Grünthal 2003; Karelson 2005; Habicht, Penjam & Prillop 2011). The adverbisation of Estonian nominal case forms can be seen as a type of lexical conversion (Kasik 2015: 40): a (more or less regular) word-formation process. An example of such a process is the adverb *tasuta* 'gratis, without fee', the abessive case form (expressing lack or absence of the noun it is attached to) of the noun *tasu* 'fee' (1):

(1) *tasu* 'reward, pay' > *tasu-ta* [reward-ABE] 'without reward, pay' > *tasuta* 'gratis'

The language internal forces behind morphosyntactic changes are in linguistics approached via two basically opposite notions: grammaticalisation and lexicalisation. While grammaticalisation reflects the development of a lexical item into a marker of a grammatical category (see e.g. Heine & Kuteva 2007, 34), lexicalisation involves a process that adds words with specific content-filled meanings to a language's lexicon (Brinton and Traugott 2005: 18). Both processes influence the natural changes in the lexicon that lexicographers need to observe to give an accurate description in a dictionary.

In our synchronic study of inflectional forms that stand out statistically from the regular frequency patterns, certain grammaticalisation paths of nouns as content words to a function-word usage are observable (> adjective; > adverb; > adposition). There is, however, also the question of a morphological form becoming an independent lexical item, an autonomous dictionary entry[2]. Since the aim of this study is not to give a theoretical explanation of the particular changes behind the (miscellaneous) group of noun-based ambiforms, or to define the stages of grammaticalisation paths, we use the umbrella term *decategorisation* to refer to categorical changes in nominal ambiforms[3].

3. Methods and data

3.1 The distribution index and its formula

The question lexicographers face when analysing a form separating from its lemma is basically: How frequent is frequent enough to establish the form as an independent lexeme? This question is clearly relative: just as the absolute frequencies of lexemes vary, particular forms can also be expected to display different (relative) frequencies. We propose a statistical measure of such relative frequency – the distribution index (DI) – which indicates whether the frequency of a word form fits its normal distribution as a noun form or deviates from it.

The idea behind such an index lies in the assumption that proper nouns tend to have constant distributions along with the case forms (combinations of number and case, e.g. plural elative and singular abessive) in the corpora. If such a constant normal distribution holds, it is possible to predict the frequencies of word forms based on their lemma frequencies. The very idea of the DI is to compare the actual (observed) frequency of a case form in a corpus with its expected frequency. The values of expected and observed frequency should be equal or close as long as the studied form follows the normal distribution. If there is a considerable difference between the values of expected and observed frequencies, one can conclude that the distribution is abnormal.

The hypothesis of constant distribution of word forms was controlled for in a study where the distribution data of case forms from two annotated corpora (the Balanced Corpus of Estonian[4] and the Morphologically Disambiguated Corpus[5]) were compared (Vainik et al., 2021). The distribution of all of the case forms (i.e. 29 combinations of number and case) demonstrated very steady proportions in both corpora (r = 0.999; StDev 0.000). We established these constant proportions of case forms as norms and used them as the basis for calculating the distribution indices (ibid.; Vainik et al., Paulsen 2021).

[2] For a discussion on such forms and their lexicographic status, see Paulsen et al. (2020).

[3] Note that decategorisation of morphological forms is also observable in languages without extensive case morphology, e.g. the plural form of nouns in Swedish (e.g. *blomma* 'flower' > *blommor* 'flowers', see Blensenius & Martens 2019).

[4] https://www.cl.ut.ee/korpused/grammatikakorpus/

[5] https://www.cl.ut.ee/korpused/morfliides/

The DI is calculated according to the following formula:

$$\mathrm{DI} = (Z - X \times Y) \; / \; X$$

Z = the observed frequency of the word form
Y = the norm of that particular case form (taken from a table of such norms)
X = the frequency of the lemma.

The expected frequency of a word form is calculated as a product of the frequency of the lemma X and the norm of that particular case form (Y). The result of the comparison should be normalised, i.e. the subtraction divided by the frequency of the lemma.

The values of the DI can (theoretically) vary from nearly −1 to 1. Values close to zero indicate normal distribution, and negative values indicate that the word form is underrepresented compared to its expected frequency. Values above zero indicate that the word form occurs more frequently than expected by the norm. On a few occasions, the value can be as high as 0.9, which indicates that the frequency of the lemma and the frequency of case forms are very close: the word occurs mostly in a certain case form. This is a situation far from the normal distribution and such cases can be classified as autonomous or emancipated word forms. These words lack the normal paradigm and can be labelled as uninflected.

In an empirical study that compared the DI of normal case forms and ambiforms, we were able to establish a tentative threshold of DI = 0.130. Values equal to or greater than this clearly show abnormal distributions (Vainik et al., 2021). Values higher than about zero but lower than the threshold show moderate deviation from the normal distribution. Overall, four intervals/ranges can be defined for the stages of DI values (ibid.):

underrepresentation:	−1	...	−0.5
normal distribution:	−0.04	...	0.04
moderate overrepresentation:	0.05	...	0.129
critical overrepresentation:	0.13	...	1

The advantage of the DI is that its values do not depend on the size of the corpus or the position of the lemma or word form in a list of frequencies. The index shows only whether the frequency of the word form follows the normal distribution as a case form in the selected corpus. As a benefit, the behaviour of both rare and frequent word forms can be measured on the same scale of relative frequency. As a result, the DI has the potential to function as a useful heuristic in certain stages of lexicographic work, i.e. when the status of a lexeme as a headword is estimated.

3.2 The corpus and its automatic processing

The study of the distributional index of nominal ambiforms is based on the largest corpus of contemporary Estonian, the Estonian National Corpus 2019, with 1.8 billion tokens[6]. The ENC2019 is lemmatised, tagged and disambiguated with the EstNLTKv.1.6 toolkit (Laur et al., 2020). The EstNLTK[7] is a natural language toolkit targeted explicitly for the Estonian language. The structure of the toolkit is written in the Python programming language and executes basic NLP tasks: tokenisation, morphological analysis (MA), lemmatisation, named entity recognition, etc. (Orasmaa et al., 2016: 2460).

In the case of a morphologically rich language such as Estonian, where different forms may have identical phonological shapes[8], the role of morphological disambiguation (rule-based, probabilistic or neural) is significant for frequency results. The result of the DI analysis hence directly reflects the outcome of the MA analysis, which in the case of Estonian starts with morphological segmentation and proceeds to PoS annotation. The current MA proceedings are based on the Vabamorf analyser, which combines rule-based and statistical models. Its lemmatisation system is mainly a dictionary-based approach, also featuring the Hidden Markov Model for disambiguation of ambiguous output. The problem with this approach is the lack of accuracy and precision with rare words that are not covered by the rules. (see Milintsevich & Sirts, 2020: 158—159.) Particularly problematic is the analysis of grammaticalised and lexicalised words or forms when the morphological tagging of lemmas and PoS is based on an unrenewed dictionary (Koppel, 2020: 59).

The Vabamorf lexicon is incorporated into the EstNLTK toolbox via the Vabamorf morphological analyser. The common ancestor of the Vabamorf lexicon and the morphological database of the Estonian language (MAB) is Ülle Viks's *A Concise Morphological Dictionary of Estonian* (1992). The inflectional patterns of Estonian words are centralised into MAB, which serves all datasets (including the CombiDic) in the dictionary writing system Ekilex[9], the centre to which the databases of the Institute of Estonian language are aggregated (Koppel et al., 2019; Kallas et al., 2020; Tavast et al., 2020).

The primary difference between the Vabamorf lexicon and the MAB is the emphasis

[6] The ENC2019 corpus contains texts collected from various domains. It consists of the Estonian Reference Corpus (texts from the 1990s until 2008 compiled by Tartu University), the Estonian Web (2013, 2017 and 2019), Estonian Wikipedia (2017 and 2019) and Estonian DOAJ (2020). The last data were crawled at the beginning of the year 2020. The ENC2019 is accessible via the Sketch Engine interface (Kilgarriff et al., 2004) at www.sketchengine.eu/ (accessed 24 March 2021).

[7] The EstNLTK toolkit is available at https://github.com/estnltk/estnltk.

[8] An example of form homonymy between nominal and verbal forms in Estonian: *viis* 'five' vs. *viis* 'brought'.

[9] https://ekilex.eki.ee/ (accessed 2 April 2021)

on either formalised morphological rules or lexicographic information. Both systems use both dictionaries and rules, although Vabamorf is focused on rules and the MAB compiles dictionary information that contains morphological paradigms in many different languages, along with frequencies and pronunciations. Moreover, the MAB does not separately perform morphological analysis.

The EstNLTK is used to parse the Estonian National Corpora (the data source in this study) and "Vabamorf is the EstNLTK's brain, heart and liver" (Indrek Hein, personal communication), meaning frequency values of forms are derived from Vabamorf. Updates to Vabamorf's lexicon are made on a daily basis and are immediately available for developers to use in the analyser and for broader use when the creator of Vabamorf, Heiki-Jaan Kaalep, officially updates Vabamorf (this information is based on personal communication with the EKI software developer Indrek Hein).

The Vabamorf analyser gives the rate of correct analyses for at least 97% of the words in texts and produces a list of analyses without the correct analysis for approx. 0.4% of words (Kaalep & Vaino, 2001). Veskis & Liba (2010) report the average accuracy of the morphological disambiguator in the standard 10-fold cross-validation test on the Morphologically Disambiguated Corpus as 96.23%. However, as Jakubíček (2021) points out, PoS tagging (a task depending directly on the morphological analysis) is an NLP task that is poorly evaluated, and its accuracy is conventionally reported on the token level[10], which only gives about 50% sentence accuracy.

In addition to the Vabamorf toolkit, neural models of morphological tagging and disambiguation are currently under development for Estonian. These models, trained on the Universal Dependencies (UD) corpus, have already achieved significant results (see e.g. Tkachenko & Sirts, 2018); however, they are not available for users yet (Kairit Sirts, personal communication). A comparison of DI results based on different morphological analyses would be an interesting task for future research.

3.3 The data and procedures

The data for the analysis of the statistical distribution of ambiforms, i.e. word forms with ambiguous status in respect to their qualification as dictionary headwords and/or their PoS affiliation, derive from a database of approx. 3,500 such ambiguous lexical items[11]. The database is organised into ambitypes according to particular PoS combinations (Vainik et al., 2020). This study focuses on the semantic case forms of nouns (see examples in (2a)); the singular partitive is included because of its participation in a semiproductive construction of "parametric words" (Sahkai, 2008: 173–174; see (2b)):

[10] See https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)

[11] At the moment, the database of ambiforms is available at https://drive.google.com/file/d/1ZEchvhupJ_1qS48nFTzSAmkKE_vUsmBJ/view

(2a)	*pilves* [cloud-INE]	'cloudy; stoned'	adverb/adjective
	kõrval [ear-ADE]	'next to'	adverb/adposition
	huvides [interest-PL-INE]	'in the interest of (smb.)'	adposition-like case form
	hetkeks [moment-TRA]	'for a moment'	adverb-like case form
	plussmärgiga [plus.sign-COM]	'positive'	adjective-like case form
(2b)	*mõõtu* [size-PART]	'size of (something)'	adposition-like case form

The selected test set comprises 965 ambiforms (i.e. roughly one-third of the registered records in our database). The number of possible interpretations of those forms is 2,021 in our initial data table, because each ambiform is associated with at least two PoS affiliations. The data table of ambiforms and their possible interpretations (in terms of PoS and case form) was provided with data on the frequencies of the actual occurrences of their different interpretations in the corpus (ENC2019). The frequency data of a word form and its potential lemma were needed as source data for calculating the DI values.

To generate the summary data table of the DI values for the selected ambiforms, we created an application written in the Python programming language. The input data table (MS Excel) consists of three columns: the first column contains the ambiform, the second the part-of-speech symbol, and the third indicates the morphological form (number + case), if applicable. For each input data triplet (ambiform, part-of-speech and morphological form), an automated HTTP request was made to the text corpus ENC2019 via the Sketch Engine[12] platform. In the DI calculation, we relied on normal distribution rates of the word form and the DI formula. The obtained statistical information, calculated DI and input data were written to a new Excel file.

The results table displays the values of the DI formula components: the absolute frequency of the assumed lemma of the ambiform (X), the frequency of the particular ambiform (Z) and the norm value (Y) for the particular case form of the input ambiform. A label indicating the DI interval was attached to the table, too. The summary table also provides information about the results of automatic morphological analysis in terms of which lemmas in which forms were recognised in each particular case. This additional information provides insights into whether an ambiform has just a single interpretation or if there are possibly several interpretations available: a factor affecting the outcome of DI calculations (see section 4.1 below).

The main data table was further provided with information about the current lexicographic statuses of the ambiforms in the CombiDic (and its underlying database Ekilex[13]), involving three options:

[12] https://www.sketchengine.eu/

[13] We thank Arvi Tavast for conducting the query on the Ekilex database.

- an ambiform is not included in the dictionary, yet. For this group, we use the label **"Candidates"** in the analysis in Section 4.
- an ambiform is included as a headword but the entry gives no information about its PoS. This group is labelled as **"Underspecified"**.
- an ambiform is included in the CombiDic as a headword and provided with PoS label(s) other than noun, i.e. the decategorisation process has been completed and the form has been approved as an autonomous lexeme. This group is called **"PoS-tagged"**.

4. Applying the D-index to noun-based ambiforms with different tagging statuses in EstNLTK and the CombiDic. The results and influencing factors

The automatic analysis of the ENC2019 corpus reveals that there is not necessarily any correspondence between the lexicographic lexicon (MAB) and the basis for the morphological analysis of EstNLTK, the Vabamorf lexicon (see the description of the interrelations between the different lexicographic and corpus analysing devices in section 3.2). When a case form of a noun has been reinterpreted as an indeclinable word (an adverb, adposition or indeclinable adjective) in the Vabamorf lexicon, the corpus tagging system is forced to "decide" whether to tag a running word in the corpus as a noun or as another part of speech. The result is that if a word form has risen to the status of a dictionary headword (e.g. *kõrval* [ear-ADE] 'next to'), the statistics on its occurrences in a text corpus will be split, too. The discrepancy in PoS-tagging between the CombiDic and the Vabamorf lexicon may be caused by differences in the lists of indeclinable words or the lexicon for the ambiforms with dynamic lexicographic status has not been updated.

In the analysis below, we take advantage of the mismatches in these databases and focus on the noun-based ambiforms from two general angles: (1) cases where the morphological analyser does not tag the ambiforms already decategorised in the MAB with a PoS other than S, and (2) cases where the ambiforms lack a PoS tag but have the status of a headword in lexicographic practice (i.e. in the CombiDic and, accordingly, also in the MAB), and those ambiforms that have no dictionary headword status. Discrepancies in the Vabamorf lexicon and the CombiDic offer the opportunity to study the effect of official decategorisation (interpretations of an ambiform as a noun vs. multiple PoS) on the DI of noun-based ambiforms. In the following, we examine the noun-based ambiforms from two perspectives: corpus processing analysis (4.1) and lexicographic treatment (4.2).

4.1 The impact of morphological analysis and PoS disambiguation on the D-index

In this section, we focus on a set of clearly decategorised ambiforms that are marked as indeclinable headwords in the CombiDic (N = 192), i.e. all of these forms have headword status confirmed with a PoS other than a noun. Some examples of the "PoS-tagged" ambiforms with their DI-values are presented in (3):

(3)	*tasuta*	(DI 0.69)	[fee-ABE]	'free of charge'	(adverb)
	kraesse	(DI 0.24)	[collar-ILL]	'upon smb'	(adverb, adposition)
	süles	(DI 0.37)	[lap-INE]	'in arms'	(adverb, adposition)
	käpas	(DI 0.04)	[paw-INE]	'mastered'	(adverb)

An interesting subset of this group is 51 ambiforms that are still analysed only as case forms of nouns by EstNLTK without alternative interpretations. These ambiforms can be accounted for as the best examples of nouns in the process of decategorisation (a process completed for these forms in the CombiDic and not started yet in the Vabamorf lexicon). The DI analysis of this group allows us to test the previously established threshold value (0.130) of distinctly independent lexemes (see Vainik et al., 2021): the fully decategorised case forms should demonstrate DI values clearly above the threshold. We refer to this small group of ambiforms with discrepant PoS statuses as "Noun".

As a comparison set, we present a group of "PoS-tagged" ambiforms with split PoS analyses (N = 141) to reveal the effects on DI values caused by decategorisation and splitting of the interpretations (nouns and some alternative PoS). These forms may be tagged with several PoS tags by EstNLTK, e.g. *lambist* (noun or adjective) [lamp-ELA] 'randomly', *asjata* (noun, adverb or adjective) [thing-AB] 'pointless'), or the forms may have alternative interpretations as case forms of the same noun due to homonymy (e.g. the forms *mõõtu* [size-ADT] and [size-PART] 'size of' coincide). Therefore, the number of calculated DI is larger (178) than the number of ambiforms in this group (141). We use the label "Noun+" to refer to this group. The DI values in the second group should be lower on average, i.e. fewer items should exceed the threshold of heightened frequency.

The DI results of the two "PoS-tagged" ambiform groups "Noun" and "Noun+" are depicted as a box plot graph in Figure 1. Table 1 presents the descriptive statistics about the compared groups.

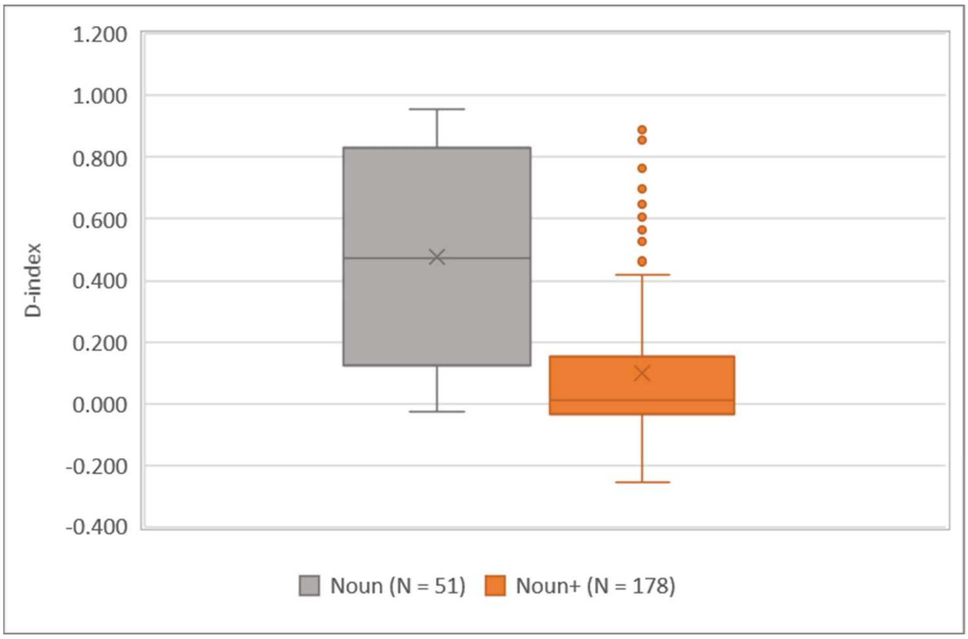

Figure 1: The variation of DI values of the noun-based ambiforms tagged as case forms of nouns (“Noun”) and as other PoS in addition to nouns by EstNLTK (“Noun+”)

	“Noun”	**“Noun+”**
N	51	178
Max	0.958	0.889
Min	−0.026	−0.256
Median	0.461	0.013
Ave	0.465	0.098
StDev	0.349	0.224

Table 1: Descriptive statistics of “Noun” and “Noun+”

Regarding the threshold level (0.130) established in our previous research (Vainik et al., 2021) distinguishing the forms with critically higher levels of relative salience from those following normal distribution rates or from those overrepresented moderately, the results of the respective samples (“Noun” and “Noun+”) show distinct tendencies. The median of “Noun” (0.461) is 35 times higher than the median of the “Noun+” group, and the average value of “Noun” (0.465) exceeds the average of “Noun+” by a factor of 4.7. Outside the boxes, the “Noun+” group shows a noticeably larger variation array, as well as extreme outliers over the upper quartiles as “abnormal” cases in respect to the limitations set by the whiskers. The variability of “Noun” is restrained by the limits of whiskers.

Figures 2 and 3 below present the DI of both groups as dot charts in a descending order. We have highlighted the values closest to the threshold on both diagrams.

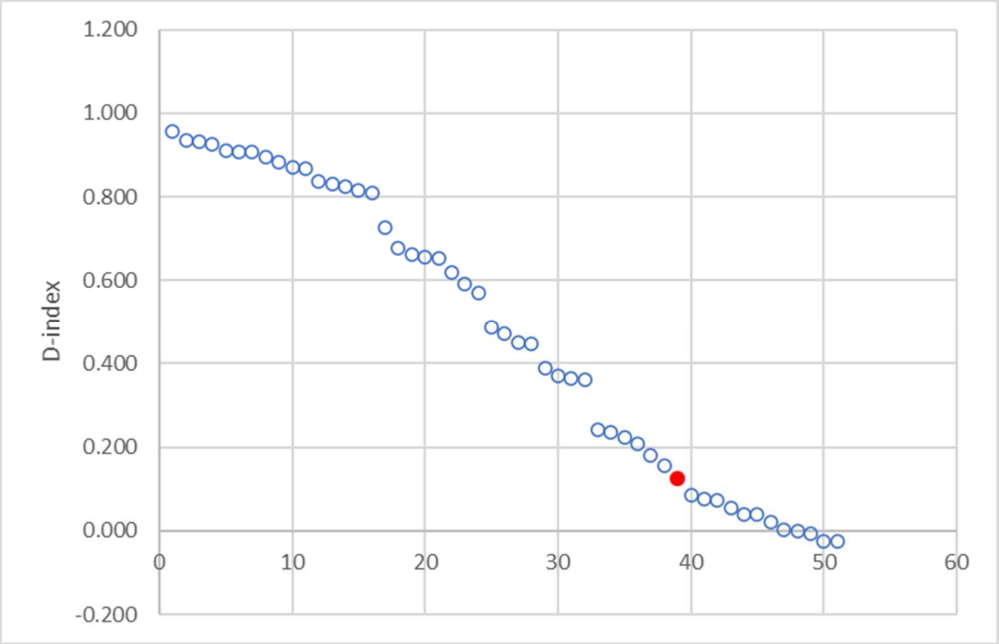

Figure 2: The DI values of the group “Noun” (the 51 “PoS-tagged” ambiforms identified only as case forms of nouns by the EstNLTK morphological analyser)

In Figure 2, we have highlighted the value 0.123 (for word form *lademes* [stratum-INE] ‘loads of’) as closest to the threshold (0.130). It appears that 75% of the ambiforms in the group “Noun” have indices above the threshold level. This result meets our expectation that the threshold value reveals most of the fully decategorised ambiforms.

The ambiforms with the highest DI values appear to be mostly compounds (see (4)), but there are also forms of some simple words (see (5)).

(4)	*otseloodis*	(DI 0.95)	[straight.level-INE]	‘in a straight line’
	eesotsas	(DI 0.93)	[front.end-INE]	‘leading’
	erandkorras	(DI 0.93)	[exception.time-INE]	‘as an exception’
	üldjuhul	(DI 0.93)	[general.incident-ADE]	‘in general’
	eestvõtmisel	(DI 0.91)	[front.taking-ADE]	‘on the initiative’
	südametäiega	(DI 0.9)	[heart.whole-COM]	‘angrily’
	teosammul	(DI 0.89)	[snail.step-ADE]	‘at a snail’s pace’
	esirinnas	(DI 0.87)	[forefront-INE]	‘in the front lines’
	ahvikiirusel	(DI 0.87)	[monkey.speed-ADE]	‘lightning fast’
(5)	*vahendusel*	(DI 0.9)	[medium-ADE]	‘via’
	hetkel	(DI 0.62)	[moment-ADE]	‘at the moment’

baasil	(DI 0.45)	[basis-ADE]	‘on the basis’
õnneks	(DI 0.57)	[luck-TRA]	‘luckily’
süles	(DI 0.37)	[lap-INE]	‘on sb.’s lap’
hoolega	(DI 0.36)	[care-COM]	‘with care’

However, relative frequency is not a clearly cogent factor leading to the status of a dictionary headword with PoS tags: 25% of the ambiforms with discrepant PoS statuses in the group “Noun” display DI that are below the threshold level:

(6)	*esirinnast*	(DI −0.03)	[forefront-ABL]	‘from the front lines’
	hääles	(DI −0.02)	[sound-INE]	‘in tune’
	käpas	(DI 0.04)	[paw-INE]	‘mastered’
	krunnis	(DI 0.05)	[bun-INE]	‘in a bun’
	mõõdus	(DI 0.07)	[size-INE]	‘size’
	mängukorras	(DI 0.08)	[play.condition-INE]	‘in playing condition’
	südamest	(DI 0.09)	[heart-ELA]	‘wholeheartedly’

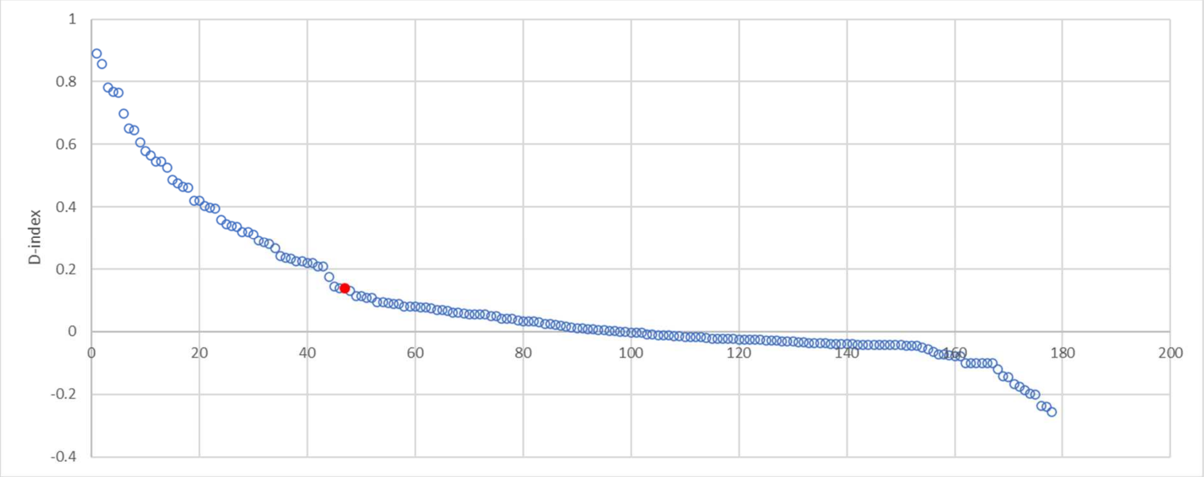

Figure 3: The DI values of the group “Noun+” (the 178 interpretations of the 141 “PoS-tagged” ambiforms labelled with several PoS tags both in the CombiDic and by the EstNLTK morphological analyser)

In Figure 3, we have highlighted the value 0.137, indicating the ambiform *mõõtu* [size-PART] ‘size of’ as closest to the tentative threshold (0.130). Its position indicates clearly that most of the ambiforms in this group have indices below the threshold; only 27.4% of the ambiforms exceed the level of the threshold. This finding confirms the hypothesis that fewer ambiforms in the “Noun+” group exceed the threshold than in “Noun”. Interestingly, the majority of ambiforms in this group (72.6 %) are below the threshold, indicating that split interpretations tend to follow a distribution that is normal or even below normal.

There are, however, some ambiforms with exceptionally high levels of DI in this category (see (7)). There are two explanations for the outstanding DI despite the multiplicity of PoS interpretations: these are either the dominating forms of lemmas with very low corpus frequency (e.g. the descriptive state adverbs *kössis*, *norus*, *kronkus* and *jõllis*: less than 1000), or clearly highly frequent forms from lemmas with high frequency in all forms (e.g. *näiteks* < *näide*, *tasuta* < *tasu* and *täiega* < *täis*).

(7)	*kössis*	(DI 0.89)	[slumped-INE]	'slumped over'
	jommis	(DI 0.86)	[drunk-INE]	'drunk'
	norus	(DI 0.78)	[somberness-INE]	'sombre'
	näiteks	(DI 0.77)	[example-TRA]	'for example'
	tasuta	(DI 0.69)	[charge-ABE]	'free of charge'
	täiega	(DI 0.65)	[full-COM]	'fully'
	kronksus	(DI 0.6)	[curled-INE]	'curled up'
	eos	(DI 0.56)	[seedling-INE]	'at the start'
	jõllis	(DI 0.55)	[bulging-INE]	'bug-eyed'

As a result of the comparison of ambiforms tagged only as case forms of nouns and the ambiforms tagged with more PoS tags than nouns by the EstNLTK morphological analyser, we can conclude that the multiplicity of PoS interpretations (also including homonyms and homographs) generally reduces the DI levels. All in all, the effect of ambiguity followed by the split PoS marking has a considerable effect on the DI of an ambiform and diminishes its reliability as a statistic of relative frequency.

In the following analysis, we will use the set of ambiforms marked as dictionary entries in the CombiDic but interpreted solely as nouns by the EstNLTK (N = 51) as a standard of the DI variation of the good candidates for decategorisation into indeclinable words.

4.2 The impact of the lexicographic status of ambiforms on their D-index

In the following analysis, we will examine the DI variation in two groups of ambiforms based on their lexicographic status. These groups will be set against an external comparison basis, the "Noun" group, representing the ambiforms tagged as case forms of nouns only (see the previous section).

The first group – "Candidates" – consists of 465 ambiforms that are not headwords in the CombiDic at all. These ambiforms originate from different sources, for instance the forms collected during the compilation of the Estonian Collocations Dictionary (2019; see Vainik et al. 2020 for the sources of the database of ambiforms), and can be seen as a possible reserve of new headwords. The question is, do the DI results indicate those ambiforms' critical relative salience and mark them as candidates for entries in the CombiDic? These ambiforms have 516 interpretations by the EstNLTK in our data table, due to form homonymy.

The second group – “Underspecified” – includes the 190 ambiforms in our noun-based ambiform selection that are headwords in the CombiDic but not tagged for PoS. These lexemes are present in the CombiDic in such an underspecified manner as a result of the aggregation processes of the superdictionary (CombiDic) from dictionaries in different formats. Some of these entries were originally subheadwords to main headwords in the Explanatory Dictionary of Estonian (2009); as a way to deal with the decategorising forms of a donor word, the subheadwords had no PoS tags. During the integration process with the CombiDic, all sub-headwords were automatically upgraded to headwords. The PoS-tagging situation of PoS-less headwords constantly changes when the dictionary is updated by lexicographers. These ambiforms have 399 interpretations in the EstNLTK analysis, 206 as case forms of nouns.

The DI variation of the headword candidates and the underspecified headwords in comparison to the set of ambiforms tagged as case forms of nouns by the EstNLTK (see Section 4.1) is presented in Figure 4. The descriptive statistics are given in Table 2.

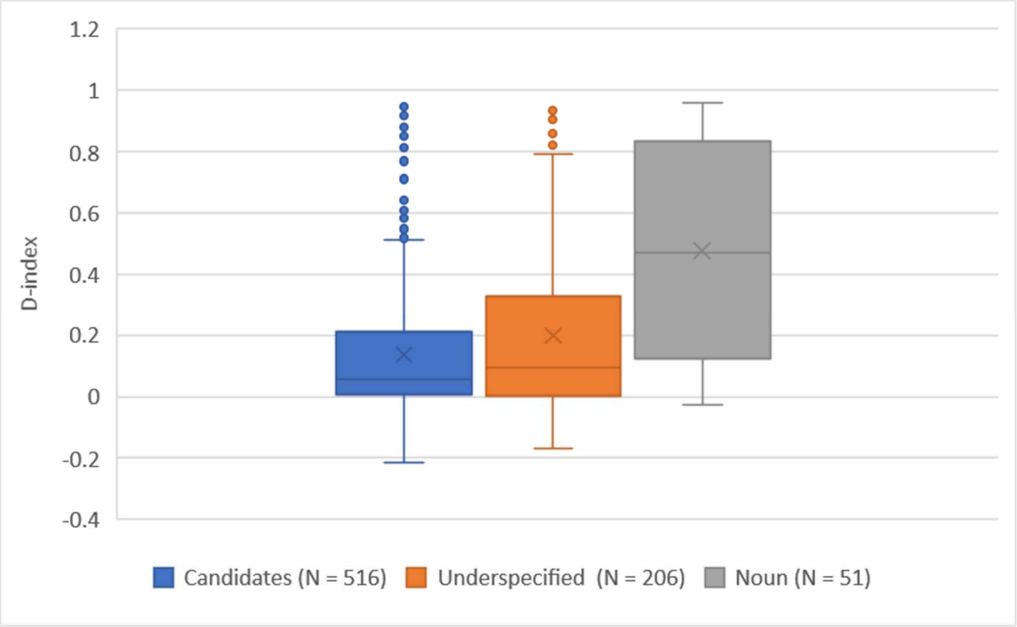

Figure 4: Variance of the DI among two sets of ambiforms: headword candidates and underspecified headwords without PoS tags compared to the ambiforms tagged as case forms of nouns by the EstNLTK

	Candidates	**Underspecified**	**Noun**
N	516	206	51
Max	0.964	0.951	0.958
Min	−0.216	−0.170	−0.026
Median	0.056	0.095	0.471
Ave	0.136	0.198	0.477
StDev	0.208	0.256	0.342

Table 2: Descriptive statistics of headword candidates, underspecified headwords without PoS tags, and the ambiforms tagged as case forms of nouns by EstNLTK morphological analysis

The data in Table 3 reveals that the maximum levels of DI are similar in all three sets, indicating that there are good candidates for decategorisation in each set, regardless of the current lexicographic status of the ambiforms. The average and median are considerably lower in the "Underspecified" group, the ambiforms in headword status without PoS tags, and the lowest in the case of "Candidates". This indicates that the lexicographic status, on average, follows the trend characterised by the relative salience of the word forms.

In relation to the "Noun" sample, the "Candidates" and "Underspecified" groups stand out for showing similar tendencies. These two sets have more tightly grouped DI values: the median results of these sets (0.056 and 0.095) are considerably lower than that of the comparison basis of "Noun" (0.471). Moreover, the average DI of the two analysed groups is 3.5 and 2.4 times lower than that of "Noun". The range of variation outside the box of 50% of the data, however, is much wider in the "Candidates" and "Underspecified" groups than in "Noun"; the extreme outliers over the upper quartiles show "abnormal" cases in these two groups.

The DI values of the headword candidates with no CombiDic headword tags are displayed in a dot chart in Figure 5:

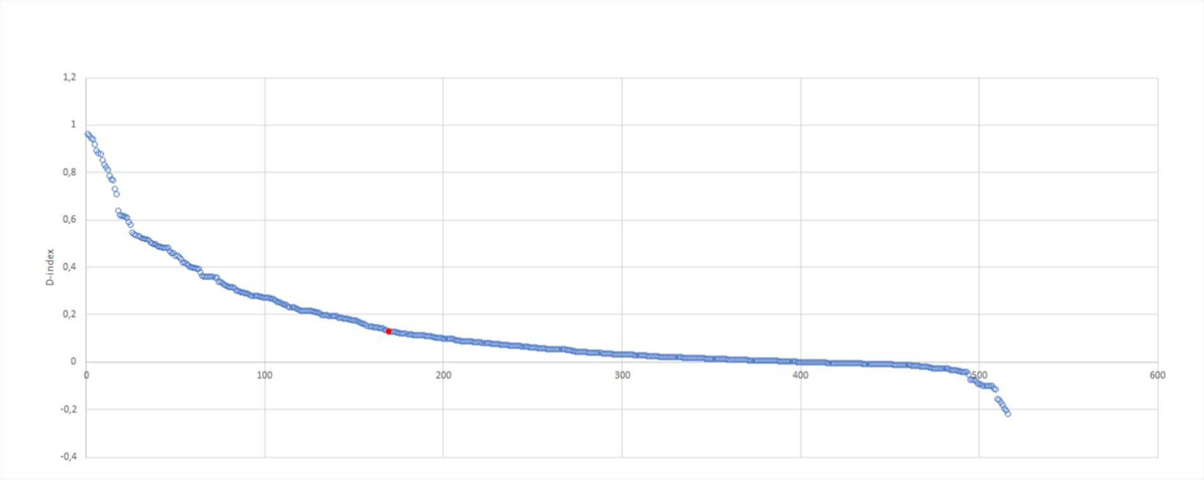

Figure 5: Descending values of the “Candidates” for dictionary headwords

This is a large set of ambiforms (N = 516). The value closest to the threshold (0.129 for the word form *keskmesse* [midpoint-ILL] ‘to the centre’) is highlighted. Only 33% of the ambiforms in this selection exceed the threshold (0.130) and truly qualify as candidates for headwords based on their morphological distribution statistics. Overall, this group shows particularly broad variation, from extremely high DI values (0.964) to negative values down to −2.16, indicating underrepresentation in relation to the expected frequency. At the top of the list are several compound ambiforms (see 8), but there are also non-compound words with exceptionally high DI (9):

(8)	*tikutulega*	(DI 0.96)	[match.light-COM]	‘scrupulously’
	ajajooksul	(DI 0.94)	[time.run-ADE]	‘over time’
	äravahetamiseni	(DI 0.89)	[away.exchange-TER]	‘interchangeable’
	reaalajas	(DI 0.88)	[real.time-INE]	‘in real time’
	vastutasuks	(DI 0.87)	[for.pay-TRA]	‘in return’

(9)	*alustuseks*	(DI 0.95)	[commencement-TRA]	‘for a start’
	nõrkemiseni	(DI 0.92)	[exhaustion-TER]	‘to exhaustion’
	maksvusele	(DI 0.82)	[validity-ALL]	‘validated’

The “Underspecified” ambiforms show a smoother decline in Figure 6. The value closest to the tentative threshold (0.129 for the ambiform *võtmes* [key-INE] ‘à la’) is highlighted. Compared to the “Candidates”, this group has more ambiforms over the threshold: 45% of the calculated DI values. These 93 case forms are good candidates for decategorisation as indeclinable words.

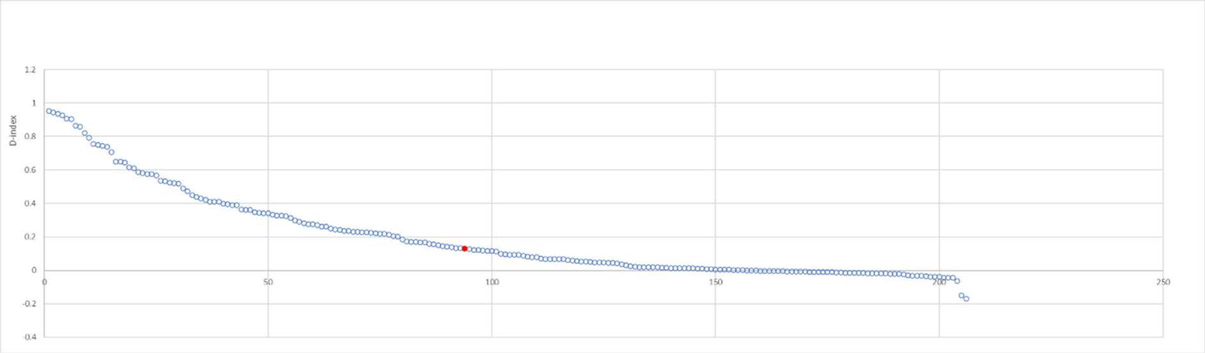

Figure 6: The Descending DI values of the "Underspecified" CombiDic headwords without PoS tags

Similarly to the previous group, "Candidates", the ambiforms with the highest DI are mostly compounds (see (10) and (11)). The ambiforms with DI levels indicating abnormal distributions in the form of underrepresentation (see (12)) are all provided with the comment "used only in negations" in the CombiDic. The reason for that is the emphatic suffix *-gi/-ki* after the case endings, often adding a sense of negation to the stem.

(10)	*üldjoontes*	(DI 0.95)	[common.feature-PL-INE]	'generally'
	esmapilgul	(DI 0.94)	[first.glance-ADE]	'at first glance'
	täismahus	(DI 0.93)	[full.capacity-INE]	'in full'
	lõppkokkuvõttes	(DI 0.9)	[end.conclusion-INE]	'in conclusion'
	eestvedamisel	(DI 0.86)	[front.leading-ADE]	'led by'
	tavamõistes	(DI 0.86)	[ordinary.sense-INE]	'colloquially'
	imeväel	(DI 0.78)	[miracle.power-ADE]	'miraculously'
	noaotsaga	(DI 0.76)	[knife.edge-COM]	'in a pinch'

(11)	*kamaluga*	(DI 0.93)	[cupped hands-COM]	'abundantly'
	mahitusel	(DI 0.9)	[encouragement-ADE]	'with the connivance of sb.'
	kuhjaga	(DI 0.61)	[pile-INE]	'heaped'
	kuubis	(DI 0.57)	[cube-INE]	'cubed'
	moel	(DI 0.57)	[way-ADE]	'in a way'
	sõnul	(DI 0.53)	[word-ADE]	'according to'

(12)	*varjugi*	(DI –0.15)	[shadow-PART-EMPH]	'(not) in the slightest'
	viluvarjugi	(DI –0.17)	[shade.shadow-PART-EMPH]	'(not) in the slightest'
	piiskagi	(DI –0.06)	[drop-PART-EMPH]	'not a drop'

4.3 Implications of morphological and lexicographic PoS tagging status on DI values

An examination of the impact of the morphological analyser on the DI results in Section 4.1 suggests that the most relevant and reliable results of the DI derive from the analysis of ambiforms that are processed as case forms of nouns without splitting the PoS interpretations into noun and additional categories. This suggests that for a realistic outline of the distributional analysis of an ambiform, all of its PoS-readings should be reverted to the noun if possible.

The influence of the headword-labelling situation of ambiforms on their DI levels examined in Section 4.2 raises the question of the relation of lexicographic treatment and ambiforms. We can ask if the DI exposes the lexicographic status of ambiforms, i.e. can the DI predict which word forms are headwords in the combined dictionary? According to our results, the answer is no: the DI variation of ambiforms that are headword candidates (not headwords in the CombiDic) and underspecified ambiforms (headwords without PoS tags) does not show significant differences.

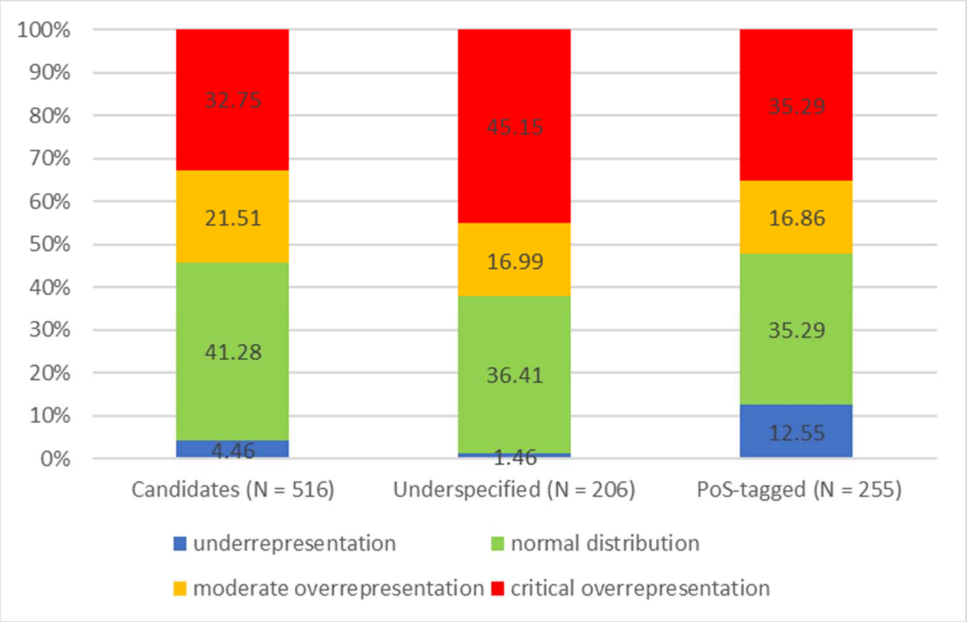

Figure 7: The division of DI results in three data sets: headword candidates, underspecified headwords and PoS-tagged headwords in the CombiDic

The results of the analysis in Sections 4.1–4.2 are summarised in Figure 7. The diagram visualises the division of DI results according to the four degrees of DI values in four data proportions: underrepresentation, normal distribution, moderate overrepresentation, and critical overrepresentation. The three columns represent the examined data from the perspective of their lexicographic status:

- “Candidates” – the ambiforms without headword status in the CombiDic
- “Underspecified” – the ambiforms with headword status but no PoS tags in the

CombiDic

- "PoS-tagged" – the ambiforms with PoS tags other than noun in the CombiDic (this column unites the data analysed in Section 4.1: the case forms of nouns in the EstNLTK morphological analysis ("Noun") and the ambiforms with split PoS analyses ("Noun+")

The proportion of critical and moderate overrepresentation is the highest and the underrepresentation the lowest in the group of underspecified ambiforms, which might indicate why these ambiforms have been given headword status in the CombiDic, although not PoS yet. The headword candidate group has a slightly smaller proportion of critical overrepresentation forms, but the highest proportion of moderate overrepresentation. The group with the expected highest proportion of critical and moderate overrepresentation, the PoS-tagged ambiforms, do not stand out in this respect; surprisingly, this group shows the largest underrepresentation level. It should be noted here that the headword inclusion in the CombiDic has not been related to the statistical distribution of the form so far. For further discussion about the reasons for including word forms with lower-than-normal distribution levels, see Vainik et al. (2021).

After the examination of the ambiform groups with different statuses in morphological analysis and lexicographic practice, we can ask if it is possible to specify any further thresholds in the relatively large area of the critical overrepresentation between the DI values 0.13—1.0. The analysis of the four groups of ambiforms (cf. Figures 3—6) reveals a gap in the line graphs around the value 0.62—0.63. This makes it possible to establish an indicative level of DI of the stage near the indeclinable words. The threshold for ambiforms approaching the characteristics of uninflected words can thus be assigned a provisional value of 0.63.

5. Conclusions

This study aimed to examine the effect of the distributional character of case forms of nouns that have already been or may be decategorised into other parts of speech. We tested the D-index developed a part of this study to detect the deviating frequency of case forms in different settings. PoS-tagging discrepancies between the morphological analyser and the combined dictionary enabled us to study the effect of "inured" and absent decategorisation on the D-index score. The results suggest that for the outcome to be most authentic, the noun-based ambiforms should be analysed without the decategorisation influence, i.e. the D-index analysis should be applied in the pre-PoS-disambiguation stage.

The threshold levels of DI posited in the previous study seemed to function relatively well as indicators of the underrepresentation, normal and moderate and critical overrepresentation of forms. The threshold value of 0.13, the marker of heightened frequency, appears to hold. The analyses of different groups of ambiforms suggest that the upper part of the critical overrepresentation (0.63), as a quite broad stage, could be preserved for the stage of "approaching the characteristics of uninflected words". A closer study of the ambiforms in this upper area is recommended for future research.

In our opinion, the D-index contributes statistical corpus post-processing information in certain stages of the lexicographic workflow: the specification of a lexeme's status as a headword and its PoS affiliation. For easy and fast access to a form's D-index, we have developed the Distribution Index Calculator for Estonian. It is a web-based application that retrieves the frequency data of word forms and lemmas from an annotated corpus and retrieves DI statistics on a lexicographer's workbench (see Vainik et al., 2021).

Since the results of the D-index (and the PoS-tagger) analysis depend on the outcome of morphological dissection, the future development of the natural language processing tasks is also relevant for our purposes. In this article, we have tested one morphological disambiguator available for the Estonian language; the other possibilities are currently the Universal Dependencies PoS Tagger[14] and the TreeTagger[15]. The development of a pre-trained language model, such as Bert, has shown promising results in PoS and morphological tagging of Estonian (see Kittask et al., 2020), which has the potential to also improve the results of the D-index calculus.

In the process of examining the D-index in use, we have determined that "dry" statistical analysis has the potential to give us new knowledge about language. The qualitative study of the groups selected for the analysis in this study and possibly the adjustment of the threshold values of the D-index form an interesting prospect for future research. There are also broader questions arising from this study, for instance: Could the D-index help improve corpus tagging systems? Can it be used in other languages? As an answer to the first question, we suggest that the D-index could help to choose the PoS that is more likely correct in disambiguation processes. The D-index itself is quite readily applicable to other morphologically rich languages, given that the norms of the forms are established.

6. Acknowledgements

We are thankful to the three anonymous referees for their valuable comments on this article. This work was supported by Estonian Research Council grant PSG227.

7. References

Blensenius, K. & von Martens, M. (2019). Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek &

[14] https://cloud.gate.ac.uk/shopfront/displayItem/tagger-pos-et-maxent1

[15] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

C. Tiberius (eds.). *Proceedings of eLex 2019 conference. 1–3 October 2019. Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 660–675.

Brinton, L. J. & Traugott E. C. (2005). *Lexicalization and language change.* Cambridge: CUP. DOI: 10.1017/CBO9780511615962.

CombiDic = *The EKI Combined Dictionary.* (2020). Hein, I., Kallas, J., Kiisla, O., Koppel, K., Langemets, M., Leemets T., Melts, M., Mäearu, S., Paet, T., Päll, P., Raadik, M., Tiits, M., Tsepelina, K., Tuulik, M., Uibo, U., Valdre, T., Viks, Ü. & Voll, P. Institute of the Estonian Language. Accessed at: Sõnaveeb 2020. https://sonaveeb.ee. (5 March 2021)

The Estonian Collocations Dictionary = *Eesti keele naabersõnad.* (2019). Kallas, J., Koppel, K., Paulsen G. & Tuulik, M., Institute of the Estonian Language. Accessed at: http://www.sonaveeb.ee. (14 February 2020)

Ekilex. Accessed at: https://ekilex.eki.ee/ (20 March 2021)

The Explanatory Dictionary of Estonian = *Eesti keele seletav sõnaraamat* I–VI. (2009). M. Langemets, M. Tiits, T. Valdre, L. Veskis, Ü. Viks, P. Voll (eds.). Institute of the Estonian Language. Tallinn: Eesti Keele Sihtasutus. Accessed at: http://www.eki.ee/dict/ekss/. (5 April 2021)

Grünthal, R. (2003). *Finnic Adpositions and Cases in Change.* Suomalais-Ugrilaisen Seuran toimituksia 244. Helsinki: Finno-Ugrian Society.

Habicht, K., Penjam, P. & Prillop, K. (2011). Sõnaliik kui rakenduslik ja lingvistiline probleem: sõnaliikide märgendamine vana kirjakeele korpuses. *Estonian Papers in Applied Linguistics* 7, pp. 19–41.

Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics*, 39(6), pp. 1041–1070.

Heine, B. & Kuteva, T. (2007). *The genesis of grammar. A reconstruction.* Oxford: Oxford University Press.

Jakubíček, M. (2021). Morphology is an open problem of NLP. Talk given at the Workshop on Parts of Speech. Tallinn: Institute of the Estonian Language. Available at: https://portaal.eki.ee/component/content/article/101-projektid/3414-workshop-on-the-role-of-parts-of-speech-in-language-technology.html.

Kaalep, H-J. & Vaino, T. 2001. Complete Morphological Analysis in the Linguist's Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pp. 9—16, Tartu. Available at: http://www.cl.ut.ee/yllitised/smugri_toolbox_2001.pdf.

Kasik, R. (2015). *Sõnamoodustus* [Word formation]. Tartu: Tartu University Press.

Karelson, R. (2005). Taas probleemidest sõnaliigi määramisel [Once again on the problems of assigning the PoS]. *Estonian Papers in Applied Linguistics* 1, 53—70.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Information Technology*, 105, pp. 116–127.

Kittask, C., Milintsevich, K. & Sirts, K. (2020). Evaluating Multilingual Bert for Estonian. In A. Utka, J. Vaičenonienė, J. Kovalevskaitė & D. Kalinauskaitė (eds.). *Human Language Technologies – The Baltic Perspective.* IOS Press, pp.

19–26. (Frontiers in Artificial Intelligence and Applications). DOI: 10.3233/FAIA200597.

Koppel, K. (2020). *Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele* [Corpus-Based Automatic Detection of Example Sentences for Dictionaries for Estonian Learners]. PhD thesis. Tartu: Tartu University Press.

Koppel, K., Tavast, A., Langemets, M. & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Proceedings of the eLex 2019 conference. 1–3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 434–452.

Langemets, M., Kallas, J., Norak, K. & Hein, I. (2020). New Estonian Words and Senses: Detection and Description. *Journal of the Dictionary Society of North America* 41 (1), pp. 69–82.

Laur, S., Orasmaa, S., Särg, D. & Tammo, P. (2020). EstNLTK 1.6: Remastered Estonian NLP Pipeline. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 7152–7160.

Orasmaa, S., Petmanson, T., Tkatšenko, A., Laur, S. & Kaalep, H-J. (2016). EstNLTK – NLP Toolkit for Estonian. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & P. Stelios (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, Slovenia: ELRA, pp. 2460–2466. http://www.lrec-conf.org/proceedings/lrec2016/pdf/332_Paper.pdf

Paulsen, G., Vainik, E., Tuulik, M. & Lohk, A. (2019). The lexicographer's voice: word classes in the digital era. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) *Proceedings of eLex 2019 conference. 1–3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 319–337.

Paulsen, G.; Vainik, E.; Tuulik, M. (2020). Sõnaliik leksikograafi töölaual: sõnaliikide roll tänapäeva leksikograafias [On word classes in contemporary lexicography: The lexicographers' view]. *Estonian papers in applied linguistics*, 16, pp. 177–202. DOI: 10.5128/ERYa16.11.

Sahkai, H. (2008). Konstruktsioonipõhine keelemudel ja sõnaraamatumudel [A construction-based model of language and dictionary]. *Estonian Papers in Applied Linguistics*, 4, pp. 177–186.

Tavast A., Koppel K., Langemets M. & Kallas J. (2020). Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, Vol. 1., Greece: Democritus University of Thrace, pp. 215–223.

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 & Simon Krek (eds.) *Proceedings of the XVIII EURALEX International*

Congress: EURALEX: Lexicography in Global Contexts. Ljubljana, Slovenia.

Tkachenko, A. & Sirts, K. (2018). Neural Morphological Tagging for Estonian. In Muischnek, K. & Müürisepp K. (eds.). *Human Language Technologies – The Baltic Perspective.* IOS Press. (Frontiers in Artificial Intelligence and Applications), pp. 166—174. DOI: 10.3233/978-1-61499-912-6-166.

Vainik, E., Paulsen, G. & Lohk, A. (2020). A typology of lexical ambiforms in Estonian. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, Vol. 1. Alexandroupolis, Greece: Democritus University of Thrace, pp. 119—130.

Vainik, E.; Paulsen, G. & Lohk, A. (2021). Käändevormist sõnaks: mida näitab sagedus? [From inflected form to a word: the role of frequency]. Accepted by *Estonian Papers in Applied Linguistics*, 17.

Vainik, E.; Lohk, A. & Paulsen, G. (2021). The Distribution Index Calculator for Estonian. *Proceedings of eLex 2021 conference.* 5—7 July 2021, Brno, Czechia. Brno: Lexical Computing CZ, s.r.o.

Veskis, K.; Liba, E. (2010). Automatic Tagger Evaluation. Syntax assignment report. NGSLT (Nordic graduate school on language technology) NLP course 2008. Available at: http://teataja.ee/veskis-liba-syntax-assignment-modified.pdf

Viitso, T-R. (2003). Structure of the Estonian language: Phonology, morphology, and word formation. In M. Erelt (ed.) *Estonian language.* Tallinn: Estonian Academy Publishers, pp. 9—92.

Viks, Ü. (1992). *Väike vormisõnastik. I: Sissejuhatus & grammatika; II: Sõnastik & lisad* [A Concise Morphological Dictionary of Estonian. I: Introduction & Grammar; II Dictionary and Appendices]. Tallinn.