The Distribution Index Calculator for Estonian

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

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia
[2] Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden
E-mail: Ene.Vainik@eki.ee, Ahti.Lohk@eki.ee, Geda.Paulsen@eki.ee

Abstract

Lexicographers working with such morphologically rich languages as Estonian face the task of detecting the lexicographic status of some word forms that look like case forms of nouns but can behave as function words to a certain degree. Hence, a measurable criterion for making a word form an autonomous headword is needed. The present paper describes the idea and development of a tool called the Distribution Index Calculator (DIC) for Estonian. It is a web-based application which finds the frequency data of word forms and lemmas from an annotated corpus and retrieves a statistic called the Distribution Index (DI). The DI indicates the relative prominence of a word form as compared to its expected normative level of salience. The application is described in detail and some illustrations of its performance are provided. The evaluation of its quality is as follows: a higher than critical level of DI can be trusted as an indicator of the relative autonomy of a word form, while a lower than critical level of DI does not preclude such autonomy. The DIC thus gives relative heuristics rather than absolute ratings or true-value decisions.

Keywords: language technology; lexicography; morphology; distribution of case forms; the Estonian language

1. Introduction

There is an endless source of candidates for new dictionary headwords in the era of e-dictionaries and automated compilation processes. This is so not only because of such obvious neologisms as *koroonaviirus* 'coronavirus' and *karjaimmuunsus* 'herd immunity', but also because of the effort to present fairly established word forms as autonomous headwords in a dictionary. The latter holds when such autonomy is justified, i.e. when the lexical items serve a function or meaning distinguishable from the base word (e.g. Blensenius & Martens, 2019).

Lexicographers working with such morphologically rich languages as Estonian face a specific task: to detect the lexicographic status of word forms that look like case forms of nouns but can behave as function words to a certain degree (e.g. *sõnul* : is it the noun *sõna* 'word' in plural adessive or the indecomposable adposition *sõnul* 'according to (someone's) claim' (Karelson, 2005; Paulsen et al., 2019)). The task is to establish the degree of emancipation of such word forms from the noun paradigm, and thus provide a justification for upgrading them to the status of independent headwords in dictionaries. A similar task in languages lacking case form morphology is, for example, establishing the lexicographic status of plural forms or derivatives. Practical decisions about whether to include a word form as a headword or not have to be made by lexicographers daily. Hence, a measurable (synchronic) criterion for word form emancipation is needed.

We can now introduce the first working prototype of the DIC[1]. Below, we refer to the

[1] teenus.eki.ee/d-index

theoretical underpinnings briefly, and describe the idea behind the statistic and its calculation. We also give the details of its realisation as an eight-line pseudocode and present some illustrations of how it works. The evaluation of the results was carried out as an experiment comparing the results of the DIC with the decisions made by lexicographers. The problems and future directions of development are also discussed.

1.1 Some notes about the theoretical background

The ubiquitous process of grammaticalisation offers a theoretical explanation for the phenomenon of developing new function words out of case forms of nouns (Grünthal, 2003; Habicht et al., 2011). A process called lexicalisation could be considered at play as well, as far as we talk about the emergence of new lexical units: the stand-alone headwords in a dictionary (for more references and discussion see Paulsen et al., 2021).

In Estonian, there are both already fossilised lexemes (e.g. *kõrval* 'beside' in (1b)) and (continually new) forms on their way to the status of lexical items (e.g. *äärel* 'on the edge' in (2b)) (see e.g. Karelson, 2005; Paulsen et al., 2019), which require the attention of a lexicographer:

(1) a. *Koera* ***kõrva-l*** *istub kärbes.*
dog.GEN ear-ADE sit-3SG fly
'A fly is sitting on the dog's ear.'

b. *Laps istub koera* ***kõrval***
child sit-3SG dog.GEN aside
The child is sitting next to the dog.'

(2) a. *Mees kõnnib katuse* ***ääre-l.***
man walk-3SG roof.GEN edge-ADE
'The man is walking on the edge of the roof.'

b. *Valitsus on kokkukukkumise* ***äärel.***
government is-3SG collapse.GEN edge
'The government is on the brink of collapse.'

It has been established that there are two types of processes that take place in the grammaticalisation of a lexical item: 1) semantic change from a referential meaning to a grammatical meaning (Hopper & Traugott, 2003: 1 (also called bleaching, see Heine, 2005: 578-579)), and 2) increase in the usage of a word form (see e.g. Feltgen et al., 2017). The two processes appear simultaneously. We can only think of the frequency of usage being a prerequisite for the semantic change. The essence of the process is that the lexical item is used more frequently and in different contexts than it was used before when it carried only lexical meaning. The acquired new aspects of meaning (or new functions) further reinforce the more frequent usage.

The DIC described here can provide information only about the increase in relative frequency. The implications of semantic change must be tackled in a separate module of a future lexicographic tool.

2. The Distribution Index and its calculation

Information about the relative frequency of word forms can be helpful when it comes to deciding whether a particular word form should be given the status of a headword in its own right. We have proposed an index of a statistical distribution of word forms (DI) as a heuristic for lexicographers (Vainik et al., 2021; Paulsen et al., 2021).

The idea behind the proposed DI lies in the assumption that proper forms of 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. Based on the knowledge of normal distribution, it is possible to predict the frequencies of word forms on the basis of their lemma frequencies. The 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 if 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 there is an abnormal distribution.

2.1 Normal distribution of the case forms

The normal distribution of Estonian case forms was established in a previous study (Vainik et al., 2021). In that work, the distribution data of case forms from two annotated corpora — the balanced corpus of Estonian and the morphologically tagged corpus — were compared in order to control for the constancy of the proportions. The distribution of all of the case forms (i.e. 29 combinations of number and case) demonstrated very steady proportions in both of the corpora (r = 0.999; StDev 0.000). The mean values of the two corpora were established as the norms (see Table 1).

Case	**DIC**	**Leipzig Glossing**	**Singular**	**Plural**
nominative	n	NOM	0.262	0.068
genitive	g	GEN	0.217	0.053
partitive	p	PART	0.102	0.037
additive	adt	ADT	0.011	
illative	ill	ILL	0.005	0.002
inessive	in	INE	0.042	0.007
elative	el	ELA	0.028	0.009
allative	all	ALL	0.028	0.008
adessive	ad	ADE	0.044	0.010
ablative	abl	ABL	0.004	0.001
translative	tr	TRA	0.027	0.002
terminative	ter	TER	0.002	0.000
essive	es	ESS	0.004	0.001
abessive	ab	ABE	0.001	0.000
comitative	kom	COM	0.021	0.006

Table 1. Normal distribution of declinable words in Estonian

The norms were deduced relying on data on all types of declinable word classes: nouns, adjectives, numerals and pronouns. As such, the norms serve as generalised benchmarks for comparison.

2.2 Formula for calculating the DI

In order to calculate the DI for an ambiform (i.e. a word form ambiguous in respect to its lexicographic status, also referred to as a wicked word form later in this paper), we need to guess which case form of which particular lemma it might be, i.e. the word form has to undergo tentative morphological analysis. For example, the word form *sõnul* would be interpreted tentatively to be the plural adessive case form of the lemma *sõna* ‘word’.

To calculate the DI we need: 1) the observed frequency of the word form in a corpus (Z), 2) the norm of that particular case form (number + case) taken from a table of such norms (e.g. Table 1), and 3) the frequency of the lemma in a corpus (X). The DI is calculated according to the following formula:

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

2.3 The scale of DI values

The value of the DI can (theoretically) vary from nearly -1 to 1. Values near zero indicate normal distribution, and negative values indicate that the word form is under-represented as compared to its expected frequency. Values above zero indicate that the word form is used more often than expected by the norm. On a few occasions, a value can exceed 0.9, which indicates that the frequency of the lemma and the frequency of case forms are very close, i.e. the word occurs mostly in a certain case form. For example, *tikutulega* [match light-COM] '(search) diligently' occurs 2,547 times and the lemma *tikutuli* ‘match light’ 2,587 times in ENC2019. Lemmas of such case forms lack the normal paradigm, and their distribution is far from normal.

In an empirical study that compared the DIs of proper case forms to ambiforms, we were able to establish a tentative threshold value of DI (0.130). Values equal to or greater than the threshold are considered to show abnormal distribution (Vainik et al., 2021). Values higher than zero but lower than the threshold show moderate deviation from the normal distribution. The tentative scale of values and labels is presented in Table 2.

Values of DI	Label
< -0.05	*normist väiksem* ‘under-represented’
-0.05 ... < 0.05	*normaalne* ‘normal distribution’
0.05 ... < 0.130	*normist suurem* ‘moderate over-representation’
0.130	*kriitiline* ‘critical over-representation’

Table 2. Values and labels used in DIC

3. Description of the development of the calculator

3.1 The designed DIC functionalities

The DIC is a web-based application accessible to everyone. It takes an *ambiform* as input from the user and retrieves corpus data (frequencies of the word form and the suspected lemma), as well as the suspected morphological form. The tool calculates the distribution index of the input form and compares it to the ranked scale of word form emancipation. The DIC provides the outcome with a verbal label of the detected tendency of the distribution. The labels reflect the values determined in Table 2 (see the previous section): normist väiksem ('under-represented'), normaalne ('normal'), normist suurem ('moderate'), and kriitiline ('critical').

3.2 Prerequisites for building the DIC application

There are some inevitable prerequisites for creating the DIC application: 1) knowledge of the valid normal distribution of case forms (number + case, abstract), 2) the established scale of DI values, 3) the availability of an expeditious module for morphological analysis, and 4) the availability of a morphologically annotated corpus for retrieving the frequency data of forms and lemmas.

3.3 The main components of the application

The DIC application is written in the Python programming language and it uses the micro web framework Flask. Due to the specifics of the application, it is necessary to use two software components: one that performs a morphological analysis of the entered ambiform and another that requests statistical information about the frequency of the ambiform and its potential base forms from a representative corpus of texts.

The morphological analysis has to provide information about lemmas, parts of speech and the forms corresponding to the ambiform. In the current prototype, we use EstNLTK (version 1.6.7), which is a natural language toolkit for Estonian written in Python. It provides resources for basic NLP tasks: tokenisation, morphological analysis, lemmatisation, named entity recognition etc. (Orasmaa et al., 2016: 2460). Alternative tools for morphological analysis, such as R-package UDPipe[2], are not available yet[3]. The EstNLTK toolkit also seems natural because its tagging system coincides with that used by Sketch Engine: the platform that lexicographers are most familiar with. From a practical viewpoint, it is preferable to avoid discrepancies in tagging, e.g. it would be helpful to find similar long-tags when it comes to looking into the concordances of the particular ambiform in SketchEngine.

The second component of the DIC makes automated HTTP requests to the Estonian

[2] See more https://www.rdocumentation.org/packages/udpipe/versions/0.8.5, https://www.r-bloggers.com/2018/02/a-comparison-between-spacy-and-udpipe-for-natural-language-processing-for-r-users/, https://universaldependencies.org/

[3] Kairit Sirts, personal communication.

National Corpus 2019 (ENC2019), which is available on the SketchEngine platform. The requests are performed by using the Sketch Engine API[4]. ENC2019 is currently the newest and largest automatically annotated corpus of the Estonian language (approx. 1.5 billion words). The corpus is annotated with the EstNLK toolkit (version 1.6.7). The precision of the annotation is not yet known. Some problems with the compilation and annotation processes of Estonian corpora are discussed by Koppel (2020).

3.4 The DIC algorithm

The DIC algorithm performs a sequence of activities when calculating the D-index. The sequence is provided by an eight-line pseudocode, as follows (and explained below):

```
1: word ⟵ user entered ambiform
2: norm_freq ⟵ read_from_file
3: lemmas, postags, forms ⟵ estnltk_morf_anal(word)
4: for i ⟵ [1, …|lemmas|]:
5:       X, Z ⟵ query_from_SkE(word, lemmas[j], postags[j])
6:       Y ⟵ norm_freq[forms[j]]
7:       D_index ⟵(Z - X * Y) / X
8:       DI_label ⟵ find_di_label(D_index)
```

Rows 1 and 2: A user enters the input data —a `word`— and the `norm_freq` is read from a file. The `norm_freq` is the normal (expected) distribution of word forms, and it is previously specified based on the balanced corpus of Estonian and the morphologically disambiguated corpus (see section 2.1 above).

Row 3: All of the possible `lemmas`, `postags`, and `forms` are found for the entered word using the Estonian morphological analyser estnltk_morf_anal (EstNLTK is the Python library for Estonian language processing and analysis; see section 3.3 above).

Row 4: Repeat the sentences in rows 6 and 9 as many times as there are elements in the lemmas list (`|lemmas|`).

Row 5: The `query_from_SkE` method queries SketchEngine based on the `word`, from which we separate the information about the frequency of occurrence of the word (`Z`) and the frequency of occurrence of the `lemmas[j]` at the `postags[j]` (`X`).

Row 6: The program finds the norm proportion `Y` for the `forms[j]` in the dictionary `norm_freq`.

Row 7: Based on `X`, `Y` and `Z, the D_index` is calculated.

Row 8: Using the predefined scale, the `find_di_label` method is used to find the rating

[4] See more at https://www.sketchengine.eu/documentation/api-documentation/

label (`DI_label`) corresponding to the `D_index`.

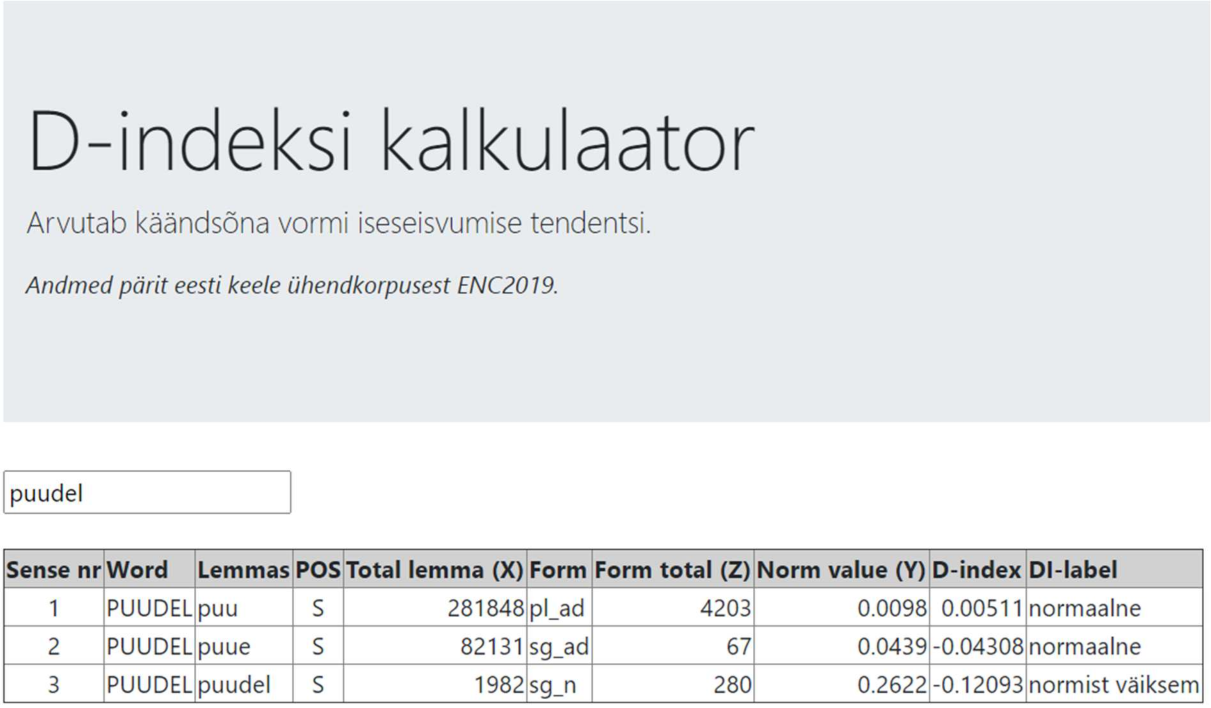
The number of D-indices of a single word (nominal) depends on how many initial lemma-postag forms morphological analysis and SkE query yield.

4. The DIC at work

The DIC works on the web. It can be opened in a separate window of a web browser while working in Ekilex or checking corpus data via SketchEngine. It is supported by the most common browsers (it has been tested on Microsoft Edge, Mozilla Firefox, Chrome, Vivaldi, and Brave).

Figure 1 presents the user interface of the DIC. The title translates as "A calculator of D-index" and the subtitle as "It calculates an autonomy tendency for case forms of declinable words" and "The data is retrieved from the corpus ENC2019". There is a search box below the title and further below are situated tabular fields for the results of a query. There will be as many rows presented as there are different interpretations provided by the morphological analysis.

The form entered, *puudel,* has three homographic readings as different case forms (plural adessive, singular nominative and singular adessive, respectively) of three different lemmas: *puu* 'tree', *puudel* 'poodle', and *puue* 'disability. The distribution rates and labels of these interpretations are presented in the last two columns. It can be concluded that the frequency of the form *puudel* is normal or below, no matter for which case and lemma it stands.



D-indeksi kalkulaator

Arvutab käändsõna vormi iseseisvumise tendentsi.

Andmed pärit eesti keele ühendkorpusest ENC2019.

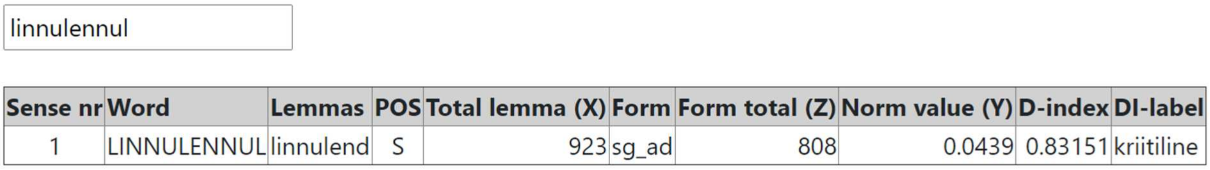
puudel

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	PUUDEL	puu	S	281848	pl_ad	4203	0.0098	0.00511	normaalne
2	PUUDEL	puue	S	82131	sg_ad	67	0.0439	-0.04308	normaalne
3	PUUDEL	puudel	S	1982	sg_n	280	0.2622	-0.12093	normist väiksem

Figure 1. The user interface of DIC.

4.1 Illustrations

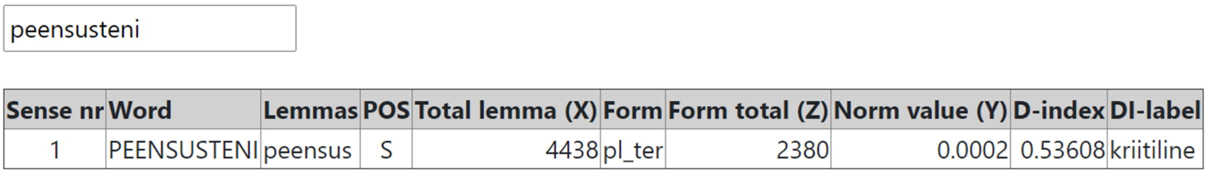
In the following, we present some examples of how the DIC works. Here is a short list of word forms: *kombel, lahus, nõusolekul, linnulennul, peensusteni, alguses, habemega, lehes* and *sõlmes*. The results of the analysis are presented in Figure 2 (a—k). We have omitted the title sections to save space. The illustrations are grouped in descending order according to their DI values (and labels).



linnulennul

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	LINNULENNUL	linnulend	S	923	sg_ad	808	0.0439	0.83151	kriitiline

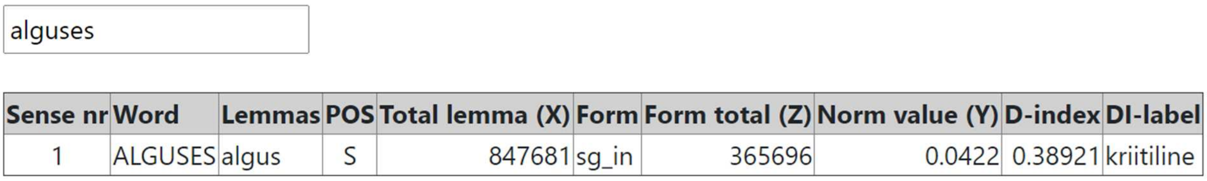
a) *linnnulennul* [bird.fly-ADE] 'very fast'



peensusteni

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	PEENSUSTENI	peensus	S	4438	pl_ter	2380	0.0002	0.53608	kriitiline

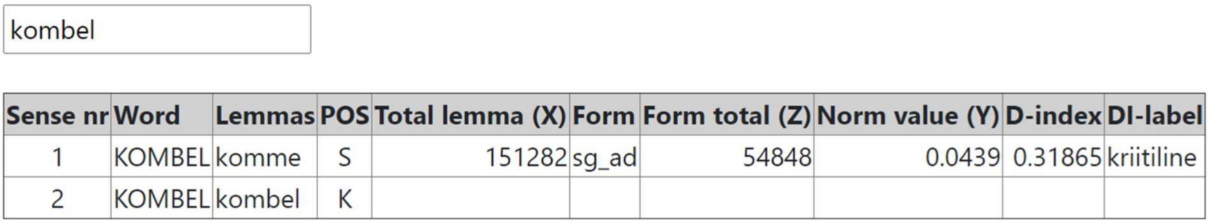
b) *peensusteni* [detail-PL.TER] 'scrupulously'



alguses

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	ALGUSES	algus	S	847681	sg_in	365696	0.0422	0.38921	kriitiline

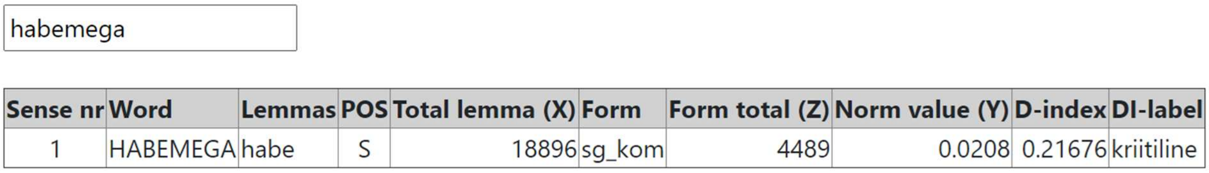
c) *alguses* [beginning-INE] 'at the beginning'



kombel

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	KOMBEL	komme	S	151282	sg_ad	54848	0.0439	0.31865	kriitiline
2	KOMBEL	kombel	K						

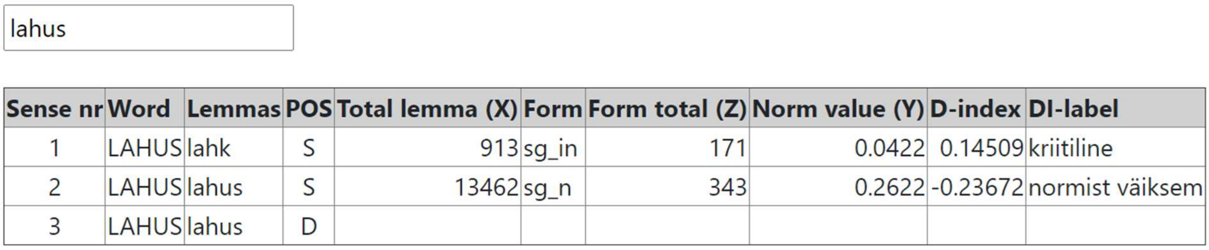
d) *kombel* [manner-ADE] 'in a way'



habemega

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	HABEMEGA	habe	S	18896	sg_kom	4489	0.0208	0.21676	kriitiline

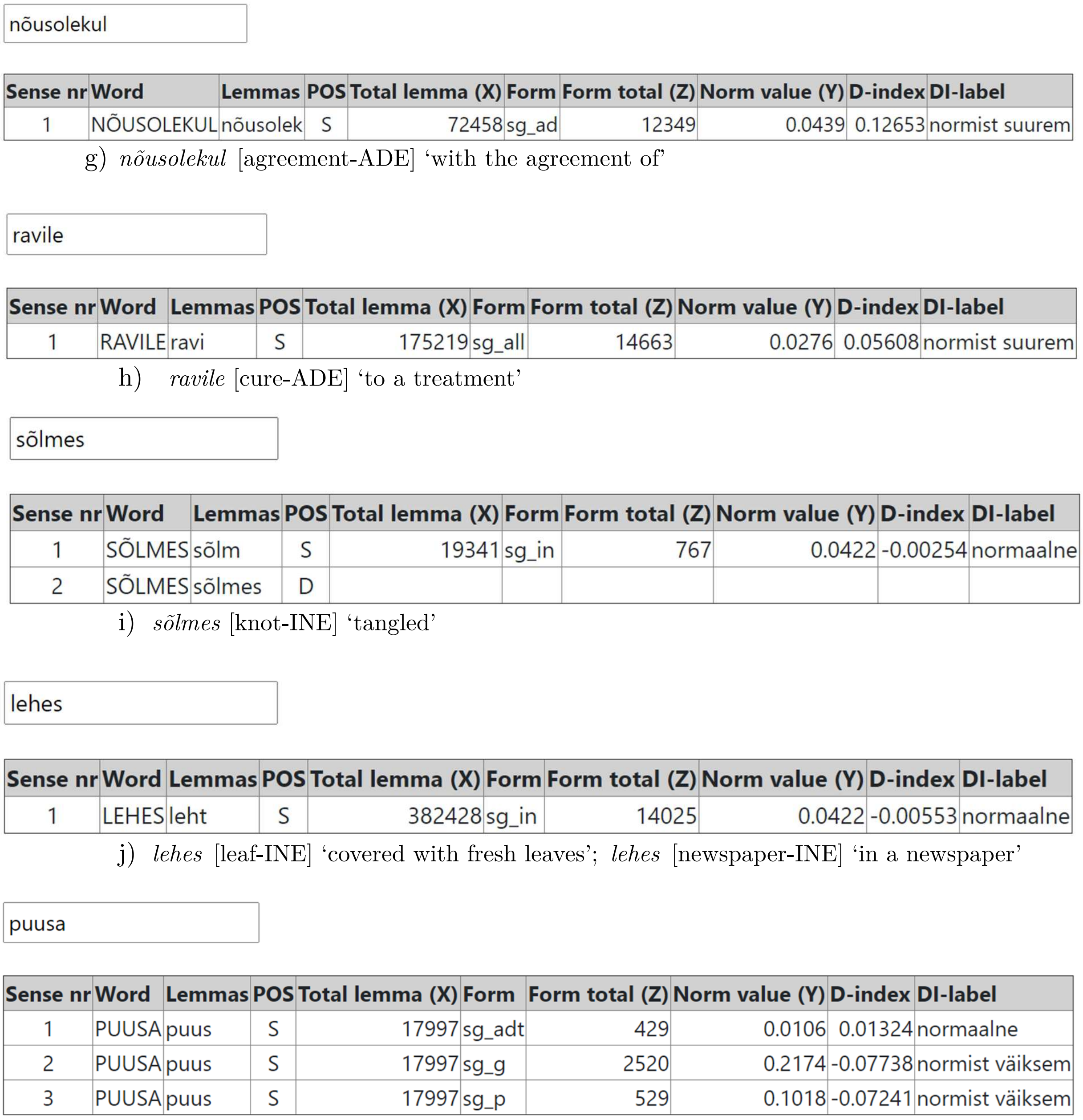
e) *habemega* [beard-COM] 'outdated'



lahus

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	LAHUS	lahk	S	913	sg_in	171	0.0422	0.14509	kriitiline
2	LAHUS	lahus	S	13462	sg_n	343	0.2622	-0.23672	normist väiksem
3	LAHUS	lahus	D						

f) *lahus* [division-INE] 'separated (from)'



nõusolekul

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	NÕUSOLEKUL	nõusolek	S	72458	sg_ad	12349	0.0439	0.12653	normist suurem

g) *nõusolekul* [agreement-ADE] ‘with the agreement of’

ravile

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	RAVILE	ravi	S	175219	sg_all	14663	0.0276	0.05608	normist suurem

h) *ravile* [cure-ADE] ‘to a treatment’

sõlmes

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	SÕLMES	sõlm	S	19341	sg_in	767	0.0422	-0.00254	normaalne
2	SÕLMES	sõlmes	D						

i) *sõlmes* [knot-INE] ‘tangled’

lehes

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	LEHES	leht	S	382428	sg_in	14025	0.0422	-0.00553	normaalne

j) *lehes* [leaf-INE] ‘covered with fresh leaves’; *lehes* [newspaper-INE] ‘in a newspaper’

puusa

Sense nr	Word	Lemmas	POS	Total lemma (X)	Form	Form total (Z)	Norm value (Y)	D-index	DI-label
1	PUUSA	puus	S	17997	sg_adt	429	0.0106	0.01324	normaalne
2	PUUSA	puus	S	17997	sg_g	2520	0.2174	-0.07738	normist väiksem
3	PUUSA	puus	S	17997	sg_p	529	0.1018	-0.07241	normist väiksem

k) *puusa* [hip-ADT] ‘(to) akimbo’

Figure 2. Illustrations of the DIC at work

It appears that the critical values of the DI vary considerably (from 0.8 down to the threshold value of 0.130). High D-indices can characterise forms with high, moderate and low lemma frequencies in absolute terms (compare c, b and a in Figure 2, for example). This is also the case with a normal distribution (compare i and j in Figure 2, for example). The comparability of the distributions, independent of the frequencies of forms or lemmas in absolute terms, is considered to be the advantage of the DI as a statistic (see Paulsen et al, 2019; Vainik et al., 2021). The examples d, f, i and k in Figure 2 illustrate the case when a form has more than one interpretation according to the corpus tagging. In some cases, there are homographic readings of the ambiform (e.g. f in Figure 2, where the form *lahus* can be interpreted as belonging to two alternative lemmas: *lahk* ‘division’ and *lahus* ‘dilution’). Another kind of multiplicity of interpretations originates in the decategorisation of certain case forms of nouns (e.g. d, f, i and k in Figure 2; see also Paulsen et al., 2019) and interpreting those as indeclinable words (adverbs — D, adpositions — K). Decategorisation may or may not diminish the DI value as a case form (as in i and d in Figure 2, respectively). The effects of decategorisation and the accuracy of corpus tagging are discussed in more detail in another paper (Paulsen et al., 2019). The case in j where the word *leht* is polysemous is the most complicated. The calculator is unable to distinguish the meanings and sums up all of the occurrences of both the form and its lemma. Thus, the potential over-representation of a form in one particular meaning, e.g. 'covered with fresh leaves', will go unnoticed on a purely statistical basis.

4.2 Evaluation of the DIC and its results

4. 2.1 Quantitative parameters

A single query by DIC took 1-1.5 seconds on average during the test period of the prototype. We noticed delays, occasionally, at times when Sketch Engine was slow anyway (for unknown reasons). The speed of the DIC is related to the smoothness of queries by Sketch Engine because the DIC retrieves its frequency data via the Sketch Engine API (see Section 3.3).

4.2.2 Quality of the results

The quality of the DIC can be estimated by comparing its output with some kind of approved standard. It is reasonable to assume that the decisions made by lexicographers so far can be used as a standard in this respect. As the problem to be solved by the assistance of the DIC is whether to include a particular word form in a dictionary as a stand-alone headword or not, we can use the DI level of the case-form-like approved headwords as a standard.

In the following, we describe the experiment of calculating DIs for a set of not yet established word forms and comparing their DI levels with similar case forms of nouns that have been approved as headwords in the CombiDic. We chose headwords from the CombiDic that are analysed as case forms only, in corpus texts by Vabamorf, and whose DIs thus purely represent their distribution as nouns and are not distorted by occasional decategorisation (see Paulsen et al., 2019 for discussion).

Table 3 presents the 30 ambiforms with their DIs based on the data of ENC2019[5]. The rows are arranged so that shared forms (number + case) are presented together. The groups are accompanied by data on their number in our database and average DI levels, as well as by examples of some headwords from the CombiDic, with the maximum and minimum value of the DI in each subcategory. The DI values exceeding the tentative threshold (0.130) set by the previous research (Vainik et al., 2021) are boldfaced in Table 3.

Eleven ambiforms out of 30 appeared to demonstrate critical over-representation (0.130), eight demonstrated moderate over-representation (0.05 ... 0.129) and eleven ambiforms

[5] An excerpt from our database of such ambiforms; see Paulsen et al. (2019).

demonstrated normal distribution (-0.04 ... 0.04).

A comparison with the DI values of the approved case-form-like headwords shows that their average well exceeds the threshold, which indicates that the approved forms generally tend to be distributed abnormally. There is remarkable variation, however, in each subcategory: the items with maximum DI values tend to be rather high (close to 0.95 occasionally) while the minimum DI values demonstrate perfectly normal distribution.

This observation — that word forms with only moderate or normal salience in a corpus (as measured by their DI) are approved as autonomous headwords in the CombiDic – can be explained in many ways. Firstly, the statistical distribution has not been the (main) concern in deciding lexicon membership. The CombiDic is an aggregated super dictionary by nature, and has inherited its content from many dictionaries compiled independently (Koppel et al., 2019, and Tavast et al., 2020). Secondly, the semantics of the word forms has naturally been the main concern in lexicography. The headwords with minimum DI levels in Table 3 are very special in terms of composition and meaning, mostly reflecting a kind of rural or robust undercurrent in the Estonian lexicon, which originates in the lifestyle of peasants. The word forms have been considered worth including in the dictionary because dictionaries are expected to assist in understanding literary and historical texts, too, and cannot be pure reflections of the newest corpora. Thirdly, the variance in the DI levels of dictionary headwords is great because not all language changes are traceable in the corpus data. The consistency of the corpus affects the statistical results obtained from it. Some case forms of nouns in our database of ambiforms just represent colloquial changes of usage that are not yet directly detectable using a corpus of written language. For example, in Table 3 the forms with normal DI levels, *VIGADETA* [mistake-PL.ABE] 'errorless' and *PÕHJUSENA* [põhjus-ESS] 'as caused', are in no way different from their approved analogues with normal DI levels: *takistusteta* [obstacle-PL.ABE] 'without obstacles' and *tulemusena* [result-ESS] 'as a result', respectively.

Table 3. Distribution indices of 30 ambiforms not present in the CombiDic compared to similar case forms present in the CombiDIc.

Ambiforms not included in the CombiDic			**Ambiforms approved as headwords in the CombiDic**			
			Average of the group			**Extremes of the group**
Ambiforms	**DI**	**Label**	**Form**	**N**	**Ave**	**Ambiforms with Max and Min values**
KOOSKÕLAS [harmony-INE] ‘in accord’	**0.756**	**critical**	sg in	67	**0.275**	*otseloodis* [stright.plummet-INE] 'vertically straight'
LÄHEDUSES [contiguity-INE] ‘nearby’	**0.547**	**critical**				
STIILIS [style-INE] 'à la mode'	**0.357**	**critical**				
KODUS [home-INE] 'at home'	**0.314**	**critical**				
LAPSEPÕLVES [childhood-INE] 'in childhood'	**0.279**	**critical**				
HÄDAS [trouble-INE] 'in trouble'	**0.187**	**critical**				
PAANIKAS [panic-INE] 'in a panic'	0.114	moderate				
RONGKÄIGUS [procession-INE] 'in procession'	0.112	moderate				
VARJUS [shadow-INE] 'in the lee of'	0.056	moderate				
MURES [worry-INE] 'worried'	0.054	moderate				*köies* [rope-INE] 'belayed'
RAAMES [frame-PL.IN] 'in the context of (smth)’	0.091	moderate	pl in	7	**0.326**	*üldjoontes* [general.line-PL.INE] ‘in general terms’
LEEKIDES [flame-PL.INE] 'in flame'	0.084	moderate				
PIIRES [border-PL.INE] 'within'	0.036	normal				
KORDADES [time-PL.INE] '(many) times'	0.008	normal				*litsides* [whore-PL.INE] 'sleep around'
VAHELDUSEKS [variance-TRA] 'for a change'	**0.447**	**critical**	sg tr	6	**0.282**	*tarbeks* [need-TRA] 'for'
VÕRDLUSEKS [comparison-TRA] 'for comparison'	**0.224**	**critical**				
PROOVIKS [try-TRA] 'on approval'	0.021	normal				
TANTSUKS [tants-TRA] 'for a dance'/'into a dance'	0.007	normal				*saateks* [accompany-TRA] 'for background'

Table 3 *(continued)*

Ambiforms not included in the CombiDic			**Ambiforms approved as headwords in the CombiDic**			
			Average of the group			**Extremes of the group**
Ambiforms	**DI**	**Label**	**Form**	**N**	**Ave**	**Ambiforms with Max and Min values**
JÕUGA [force-COM] 'by force'	0.059	moderate	sg kom	16	**0.365**	*kamaluga* [hand-COM] 'handful'
HINGEGA [soul-COM] 'passionately'	0.031	normal				
ÜLLATUSEGA [surprise-COM] 'with surprise'	0.006	normal				*kapaga* [cup-COM] 'in quantities'
PENSIONILE [pension-ALL] 'pension off'	**0.151**	**critical**	sg all	7	**0.172**	*tagaplaanile* [back.ground-ALL] 'to the background'
MINEKULE [leaving-ADE] 'to be leaving'	-0.008	normal				*verele* [blood-ALL] 'into bleeding'
HINNANGUL [estimate-ADE] ‘as estimated’	**0.528**	**critical**	sg ad	51	**0.346**	*esmapilgul* [first.glance-ADE] ‘at first glance’
VÕIMUL [power-ADE] 'in power'	0.028	normal				*pasal* [shit-ADE] 'diarrhea'
RÕÕMUST [joy-ELA] 'because of joy'	0.013	normal	sg el	6	**0.170**	*surmasuust* [death.mouth-ELA] 'escape death' *esirinnast* [forefront-ELA] 'from the forefront'
PÕHJUSENA [põhjus-ESS] 'as caused'	0.006	normal	sg es	4	**0.42**1	*kulutulena* [wildfire-ESS] 'extensively' *tulemusena* [result-ESS] 'as a result'
TÜKKIDEKS [piece-PL.TRA] 'into pieces'	0.074	moderate	pl tr	1	**0.267**	*ribadeks* [strip-TRA] ‘into strips’
ANDMETEL [data-PL.ADE] 'based on data'	**0.197**	**critical**	pl ade	7	**0.563**	*savijalgadel* [clay.foot-PL.ADE] 'shaky' *sulgpatjadel* [feather.pillow-PL.ADE] 'treasured'
VIGADETA [mistake-PL.ABE] 'errorless'	0.007	normal	pl ab	2	**0.206**	*viperusteta* [glitch-PL.ABE] 'without a glitch' *takistusteta* [obstacle-PL.ABE] 'without obstacles'

The results of the experiment suggest that the lexicographers could include the eleven ambiforms with critical DI values in Table 3 in a dictionary without hesitation while with the others additional — preferably semantic — consideration is needed. On the other hand, the status of word forms already included in the CombiDic can be validated — to some degree — automatically, based on their higher than threshold DI values.

The overall quality rating of the DIC can be formulated in this way: a higher than critical level of DI can be trusted as an indicator of the relative autonomy of a word form, while a lower than critical level of DI does not preclude such autonomy. The DIC thus provides relative heuristics rather than absolute ratings or true-value decisions.

5. Conclusion and discussion

There is a need for a measurable criterion when deciding the lexicographic status of some wicked case forms of nouns in Estonian that can take the meaning and function of indeclinable function words. We have proposed a distribution index (DI) as such a measure. The DI can be used as an indicator of the correspondence of a particular form's actual frequency with its predicted — in the normal distribution of case forms — elicitation degree.

We have described the steps taken to develop an application — the Distribution Index Calculator (DIC) — which can be used by lexicographers when working with wicked word forms (called ambiforms in this paper and elsewhere (e.g. Vainik et al., 2020; Paulsen et al., 2019)). The purpose of such an application is to provide the lexicographer with more elaborate statistical information than absolute frequencies and to process further annotated corpus data with the aim of developing a more specific indicator of the degree of grammaticalisation. We have described the prerequisites and the main components of the application, as well as having provided the algorithm.

As a result, the DIC is a web-based application accessible to everyone. It takes an *ambiform* as an input from the user and retrieves corpus data (frequencies of the word form and the suspected lemma), as well as the suspected morphological form. The tool calculates the distribution index of the input form and compares it to the ranked scale of word form autonomy. The DIC provides the outcome with a verbal label about the detected tendency of the distribution.

A substantial part of the paper was devoted to providing examples of the DIC at work and to comparing the results of the DIC with the decisions made by lexicographers when approving such forms for the CombiDic of Estonian. The conclusion was that the DIC provides relative heuristics rather than absolute ratings or true-value decisions. This is because a higher than critical level of DI can be trusted as an indicator of the relative autonomy of a word form, while a lower than critical level of DI does not preclude such autonomy, and additional inspection of the case forms is needed.

The idea of the DI and the calculator providing indices as a measurable statistic is based on the assumption that the case forms generally follow a constant proportion (i.e. their normal distribution) in corpus texts. It has also been stated by Koppel (2020) that "[...] patterns of

Estonian words are well established and rarely debated among lexicographers [...]". However, the existence and categorisation of wicked case forms has been quite a problem for lexicographers (Paulsen et al., 2019; Karelson, 2005). The question of upgrading lexical items that traditionally were sub-headwords in dictionaries to headwords has arisen in the context of aggregating autonomous dictionaries into the unified CombiDic (and its underlying database, Ekilex) (Koppel et al., 2019; Tavast et al., 2020).

One can argue that the DIC does a task similar to the Sketch Engine's function "frequent constructions", i.e. revealing the relative prominence of certain forms. However, as the DI is based on a comparison with the normative distribution of case forms, our tool provides an instant comparison with the norm and is thus more informative about possible deviations. We believe that the DIC can be useful for lexicographers as it provides the results of the calculation, as well as information about the existence of alternative interpretations due to homonymy. No lexicographer has tried to work with the DIC yet, as it is still in development.

Since the setup of the DIC is generic, it can also be used to test the tendencies of morphological distribution in other languages with rich morphology. The language-specific normal distribution rates (number + case) need to be available and the scales have to be established beforehand. Finnish might be a good candidate for a trial[6], as there are similar grammaticalisation processes of nominal case forms (see e.g. the analysis of the grammaticalisation of body-part nouns into adpositions in Ojutkangas, 2001). "Most Finnic adpositions display elements of productive noun inflection and frequently apply one of the local case sets" (Grünthal, 2003: 47).

6. Limitations of the application and suggestions for future research and development

Some limitations of this work should be noted. The first and foremost is that the results provided by the calculator depend on the accuracy of the corpus tagging. The DIC cannot go beyond the existing annotation yet. Both the corpus tagging system and morphological analysis are based on the Vabamorf (OÜ Filosoft) software, using the EstNLTK 1.6 (Python) library. This is open-source software with broad functionality created specifically to analyse the morphology of the Estonian language (Orasmaa et al., 2016: 2461). However, the wicked case forms described in this paper also cause problems for morphological analysis. This is because there is no good procedure for their disambiguation when it comes to choosing between multiple available interpretations. If a word form has been approved as an indeclinable word for the lexicon of Vabamorf, this results in a tendency for the analysis of this particular word form to be split between different interpretations with questionable accuracy. Such examples appear in illustrations d, f and i in Figure 2. Split interpretations can result in a decrease in the DI level of that form from heightened value to normal.

[6] Data regarding the distribution of case forms in Finnish is available online: https://kaino.kotus.fi/visk/sisallys.php?p=1227&fbclid=IwAR1v5oF4UqIySTckF50KwZK11VBmR8RJdHa6UNATYF9O241B1LYJ4DsbtnI and https://kaino.kotus.fi/visk/sisallys.php?p=1228

Therefore, the results of the forms with multiple interpretations cannot be fully trusted.

Another limitation is that the meanings of a polysemous word cannot be separated yet. The DIC calculates the indices of word forms as if there were only one form deductible to one particular lemma. This is shown in illustration j in Figure 2.

The DIC is in the process of ongoing development. Multiple paths forward are available in this respect: one involves improving the current prototype, e.g. by refining and fine-tuning the norms, the scale and the threshold to meet the more specific needs of lexicographers. Adding statistical information about interpretations other than case forms is one option. Another way to improve the current prototype is by extending its coverage to multiple corpora, which would enable it to follow changes in the relative salience of wicked forms in different styles, e.g. colloquial vs general usage, or by tracking diachronic changes. It is also possible to make the interface of the DIC more attractive, e.g. showing its output using visualisations.

One of the directions of future work is to try to overcome deficiencies due to the current morphological annotation of the corpus. We have thought about testing a “zero hypothesis”, i.e. ignoring the PoS definitions of morphological coding and retrieving data as “wild” word forms, summing up the numbers of the forms independently of their PoS tagging. We believe that such an approach would result in higher DI values of ambiforms with split interpretations. On the other hand, the information about their decategorisation would be lost. We are also open to trying some alternative systems of morphological tagging if available (e.g. Universal Dependencies PoS Tagger, TreeTagger and/or RFTagger).

The ultimate goal of future work is to incorporate the DIC into a more complex multi-search application, which would help lexicographers to attach POS tags to lexical units in a more systematic way. The multi-search application has to give a more comprehensive picture of a word form’s behaviour in texts. A measure of statistical distribution will be combined with measures of morphosyntactic behaviour and semantic similarity to a prototype of the suspected word class.

7. Abbreviations

Glossing: ABE – abessive case; ABL — ablative case; ADE – adessive case; ADT – additive case; ALL – allative case; COM — comitative case; ELA – elative case; ESS — essive case; GEN – genitive case; ILL – illative case; INE — inessive case; PART – partitive case; PL – plural; SG – singular; TER — terminative case; TRA – translative case

8. Acknowledgements

This work was supported by Estonian Research Council grant PSG227.

9. References

Blensenius, K. & von Martens, M. (2019). Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies. In I. Kosem et al. (eds.) *Proceedings of eLex 2019* *conference. 1–3 October 2019. Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 660–675.

CombiDic = The EKI Combined Dictionary [EKI ühendsõnastik]. (2020). I. Hein, J. Kallas, O. Kiisla, K. Koppel, M. Langemets, T. Leemets, M. Melts, S. Mäearu, T. Paet, P. Päll, M. Raadik, M. Tiits, K. Tsepelina, M. Tuulik, U. Uibo, T. Valdre, Ü. Viks, P. Voll. Institute of the Estonian Language. Accessed at: Sõnaveeb 2020. https://sonaveeb.ee. [25.2.2021].

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

Feltgen, Q., Fagard, B. & Nadal, R. (2017). Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language change. *Royal Society Open Science 4*. DOI: 10.1098/rsos.170830.

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. [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.

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

Hopper, P. J. & Traugott, E. C. (2003). *Grammaticalization.* 2nd ed. Cambridge: Cambridge University Press.

Kallas, J., Langemets, M. & Tender, T. (2019). Opening up Estonian dictionaries to European communities and language technology. In: Tender, T.; Eichinger, L. M. (Eds). *Language and Economy. Language industries in a multilingual Europe: EFNIL Conference 2019, Tallinn 2019.* Research Institute for Linguistics, Hungarian Academy of Sciences, pp. 149—154.

Karelson, R. (2005). Taas probleemidest sõnaliigi määramisel [The problems of PoS tagging revisited] *Eesti Rakenduslingvistika Ühingu aastaraamat, 1*(2004), pp. 53—70.

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 Ülikooli Kirjastus.

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.

Milintsevich, K. & Sirts, K. (2020). Lexicon-Enhanced Neural Lemmatization for Estonian. In: *Human Language Technologies – The Baltic Perspective.* IOS Press. (Frontiers in Artificial Intelligence and Applications), pp. 158—165. DOI: 10.3233/FAIA200618.

Ojutkangas, K. (2001). *Ruumiinosannimien kieliopillistuminen suomessa ja virossa. Suomalaisen Kirjallisuuden Seuran toimituksia* 845, Helsinki.

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 & S. Piperidis (eds.) *Proceedings of the*

Tenth International Conference on Language Resources and Evaluation (LREC 2016): The International Conference on Language Resources and Evaluation; Portorož, Slovenia; 2016. Portorož, Slovenia: ELRA, pp. 2460−2466. Available at: 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 the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 319−337. Available at: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_18.pdf

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.* Alexandroupolis, Greece: Democritus University of Thrace, pp. 215−223.

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., 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.