ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук
Образовательная программа
«Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему «Морфологический анализатор языка руули с использованием lexd и twol»

Тема на английском “Morphological Parser for Ruuli Language in lexd and twol”

Студент 3 курса
группы № 202
Бурлаков Филипп Юрьевич

Научный руководитель
Сериков Олег Алексеевич,
приглашённый преподаватель

Москва, 2023 г.

Content

1. Introduction

This paper is dedicated to the process of making a morphological parser for the Ruuli language.

Ruuli[1] (also known as Ruli, Ruruuli-Lunyala, Luruuri-Lunyara, Ruruli-Lunyara, Ruruuli-Runyala and Luduuli) is a threatened Bantu language spoken in Uganda by 160,000 people (2002 census) [Lewis et al 2015]. Being a threatened language, it is also expectedly a low-resource language, which makes it inefficient to apply machine learning and deep learning methods in order to build a morphological parser, which can help automatize glossing and allow to make a corpus of texts with morphological parsing for theoretic research.

Having said that, the best option is to use tools based on transducers, which is an instance of the rule-based approach. Here, we apply *lexd* and *twol* formalisms of Helsinki Finite-State Transducer toolkit (HFST) [Lindén et al 2011], more or less corresponding to morphological deep and surface structures respectively. The first one represents morphemes as they are stored in our lexicon, while the other takes into account rules that change their appearance. For instance, English plural morpheme is *-s*, and it is its deep representation, however, there are certain rules, which change its form to *-es*, like in contexts, where it goes after letter *x*: *fox* + *-s* > *fox-s* (deep representation) > *fox-es* (surface representation).

Therefore, during our work we regulated morphological combination rules at *lexd* step, and phonological and morphophonological rules at *twol* step.

We intended to make an accessible and user-friendly tool for scientists that are currently working with texts in Ruuli or will take part in the research in the future, so after building the parser, we made a simple web interface that is easy to use.

All the information about the language was taken from the grammar sketch [Namyalo et al 2021] and via collaboration with one of its authors, Alena Witzlack-Makarevich.

[1] JE103; Glottocode: ruul1235, ISO 639-3: ruc

Data, code, and materials we used are available on our Github page[2].

2. Development of the morphological parser

2. 1. Overview

As we have already mentioned, the morphological parser was developed using transducer-based tools, which implies manual description of the grammar. The work was divided into stages, one for each part of speech. Obviously, we mainly focused on the parts of speech with inflection categories, others were added as simple elements during preprocessing of the dictionary.

For each part of speech, we described it combinatory potential (i.e., all the grammatically possible forms) and relevant morphophonological and phonological rules that affect graphic representation of a word (as some of the processes can cause change in pronunciation, but not orthography).

After building the rules, we processed the dictionary for it to fit the patterns. At the final stage of working on the parser we made an attempt to estimate its accuracy using the glossed corpus of ELAN records and examples from the grammar sketch.

The pipeline can be described as:

1. Writing the rules as they appear in the grammar
    a. Orthography and phonology (*twol* rules)
    b. Morphology (*lexd* rules)
        i. Nouns
        ii. Adjectives
        iii. Pronouns
            - personal
            - demonstrative
            - possessive
        iv. Verbs
2. Dictionary preprocessing

[2] https://github.com/PhilBurub/Morphoanalyzer-Ruuli/

3. Estimation

   a. ELAN files preprocessing

   b. Examples from the grammar preprocessing

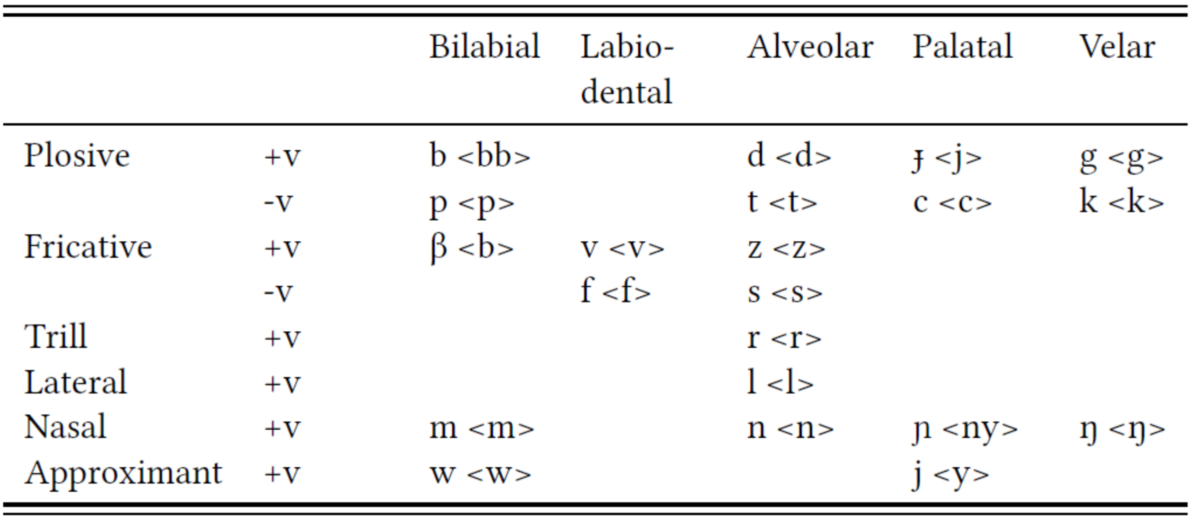
   c. Accuracy

2. 2. Rules

2. 2. 1. Orthography and phonology

Consonant phonetic inventory with graphic correspondence to each phoneme is presented in the table taken from the grammar:

Table 1. Consonants and corresponding graphemes [Namyalo et al 2021].



		Bilabial	Labio-dental	Alveolar	Palatal	Velar
Plosive	+v	b <bb>		d <d>	ɟ <j>	g <g>
	-v	p <p>		t <t>	c <c>	k <k>
Fricative	+v	β <b>	v <v>	z <z>		
	-v		f <f>	s <s>		
Trill	+v			r <r>		
Lateral	+v			l <l>		
Nasal	+v	m <m>		n <n>	ɲ <ny>	ŋ <ŋ>
Approximant	+v	w <w>			j <y>	

And the same for vowels based on the grammar information:

Table 2. Vowels and corresponding graphemes.

	Front		Central		Back	
	-l	+l	-l	+l	-l	+l
Close	i <i>	ii <i:>			u <u>	uu <u:>
Mid	e <e>	ee <e:>			o <o>	oo <o:>
Open			a <a>	aa <a:>		

Most of the speakers treat *r* and *l* as interchangeable allomorphs of the same phoneme, so we did not make any distinction between them. On the deep level we generalized both to *l*, and on the surface level we allow both of them to take place:

(1) `l:l l:r`

Hiatus resolution takes a notable place in Ruuli phonology. There are at least three processes that occur in such contexts: glide formation from vowels /i/, /o/ and /u/, progressive assimilation of vowel /a/ (viewed in the grammar as a combination of elision and compensatory lengthening) and fusion, which only takes place between a verb root and an applicative suffix. Also, according to the examples, there is a (morpho)phonological process, which has not been studied yet, that deletes /a/ vowel or changes it to a glide /j/. Other common processes are nasal assimilation and postnasal fortition of liquids and approximants.

There are also rules that do not change orthography of morphemes, such as palatalization. They were not considered in the framework of this project as our morphological parser analyses only text data.

2. 2. 2. Morphology

Nouns

It is well known that Bantu noun system is described with the concept of noun classes, which are considered to be a word classifying gender-like notion. There is such a process as a class shift, which means that a word changes its noun class, usually with a regular meaning shift, and 'replaces' its initial suffix with another one, which is commonly somehow more semantically marked [Lutz 2021]. Class shifts can express plurality, augmentative and diminutive senses.

For nominal patterns we separate class prefixes from roots and take into account combinatorial possibilities of each root. Singular-plural class shift is described in the dictionary, while diminutive and augmentative ones are deduced from the rules described in grammar, such as augmentatives of singular count nouns would belong to class 20 and attach its prefix *gu*-. More detailed overview of the way we worked with the dictionary to find possible class matches will be given in the Dictionary preprocessing section.

The other feature of nominal morphology in Ruuli is augment prefix, which precedes the class ones. It is said to have three allomorphs (*a*-, *o*- and *e*-) depending on the noun class. Moreover, some vowel-initial nouns do not attach the augment prefix.

Our goal was to restrict the possible combinations according to the rules. In order to do so, on the stage of dictionary preprocessing we marked three features of a noun root:

1. Singular-plural classes of a noun.
2. If a noun is uncountable (or mass).
3. If a noun is able to attach the augment prefix.

As an example of a noun root entry:

(2) `<bank_account>:akawunti[9a,10a,naug]`

Here, we have an English gloss inside the angle brackets, Ruuli root after the colon and its combinatorial possibilities inside the square brackets. It means that the root *akawunti*, "a bank account", attaches prefixes of 9a and 10a noun classes and does not attach the augment prefix. Besides that, it does not have a *mass* marking, therefore, it can (theoretically) attach the prefix of class 20 as its augmentative derivation (which is a common way for countable nouns to form augmentative meanings). A slightly different example would be the following:

(3) `<mud>:nyangata[8,mass]`

The root *nyangata* does not have a singular-plural noun class shift, so here we have the *mass* marking and only one noun class marking (which is common for uncountable nouns). However, mass nouns (again, theoretically) can attach other classes' prefixes for augmentative-diminutive variants, such as classes 12, 13 and 14. Also there is no *naug* marking which means it can combine with the augment prefix.

Adjectives

Ruuli adjectives are relatively simple to describe, as they agree with their head nouns in class and augment, which follow the same morphophonological rules as in the noun system. Therefore, we did not restrict its compatibility with noun prefixes in any way.

Pronouns

Personal pronouns are simply a closed class that can only attach the additive focus marker.

Demonstrative pronouns agree with their head in class. Also, medial demonstratives always bear the augment prefix that follows its regular rules.

Possessive pronouns consist of a possessee prefix (which marks agreement with a head in class), an associative marker (stem) and a possessor suffix.

Verbs

Below we present a full morphological pattern of Ruuli verbs. This is a slightly elaborated version of what is shown in the grammar with additional information from the article [Molochieva et al 2021].

Table 3. Morphological pattern of verbs.

Position	**Sense/Inventory**	**Restrictions**
TA1	narrative prefix	
NEG1	standard negation	*with *NEG2*
SBJ	subject indexing	
NEG2	prohibitive	*with *NEG1*
TA2	tense prefixes, progressive, persistive	*with subjunctive suffix *progressive and persistive with other aspectual affixes
OBJ	object indexing	
root		
APPL	applicative suffix	portmanteau form with *TA3*
CAUS	causative suffix	portmanteau form with *TA3*
RECP	reciprocal suffix	portmanteau form with *TA3*
TA3	perfective suffix	*with Final *with other aspectual affixes
PASS	passive suffix	portmanteau form with *TA3*
Final	subjunctive/ final vowel	*with *TA3* only subjunctive with the negation of near future
Post-final	habitual suffix	*with other aspectual affixes

The verb is the most complex part of speech in the morphological system. Besides the wide range of affixes and their compatibility patterns, we had to consider unique morphophonological rules. This leads to verbs being the least thoroughly described and causes many flaws, that we will discuss later in the Estimation part.

The hardest aspect is the morphophonology of the perfective suffix. There are several possible ways roots combine with the suffix: it may be added regularly (which is not very frequent) or the perfective morpheme may be fused with the root in a simple way, imbricated into it or interwoven with it. Those irregular cases involve consonant deletion, addition of fossilized affixes and enclitics, and combination of a root and the suffix in a discontinuous way, sometimes with simultaneous fusion.

The pattern is normally hard to predict based on the root. Fortunately, in the dictionary, for each verb the authors present its perfective form. Within our rules, we used it as a shortcut: we use provided irregular perfective forms as bases, so the parser does not have to build them itself. This may cause problems with so-called 'portmanteau forms' (which are fused forms of several suffixes) and we are going to focus on that during our future development.

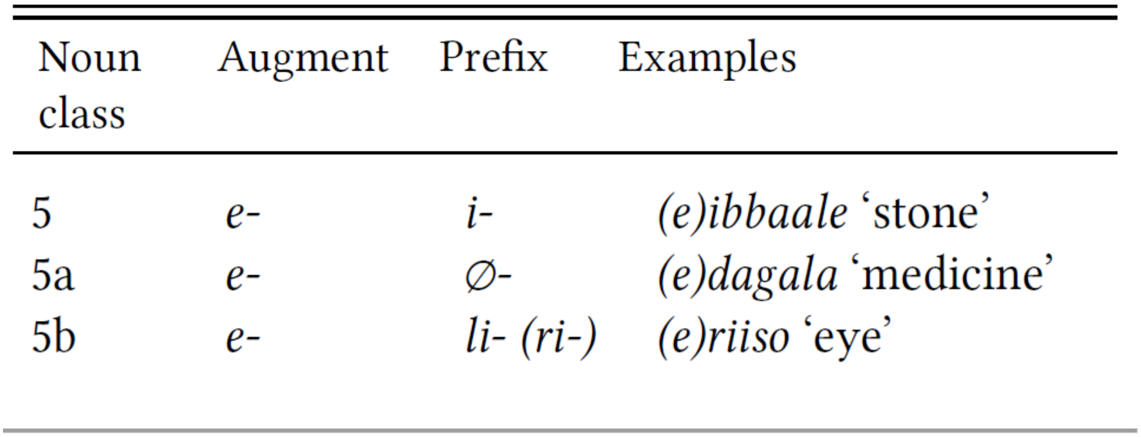
2. 3. Dictionary preprocessing

The most complicated parts of dictionary preprocessing consisted in verb and noun systems, which we will describe in more details. Otherwise, all we did was reassigning part of speech tags in order to make the inventory more consistent.

2. 3. 1. Nouns

As we have already mentioned, we stored bare noun roots in the transducer's vocabulary and assigned its compatible classes. To do so, we needed to cut off the class prefixes and detect the class. The main problem here was that in the dictionary, there are not marked, as we may call it, "subclasses". What we mean by this are those pseudoclasses that syntactically perform the same agreement as their main class, however, morphologically either do not bear any class prefixes ('*a*' classes) or bear a distinct one (the '*b*' class). In the grammar they go under notation of the number of their main class and a letter. For instance, here is a part of the class table from the grammar that shows class 5 and its two subclasses:

Table 4. Noun classes 5, 5a and 5b, and their noun class prefixes [Namyalo et al 2021].



Noun class	Augment	Prefix	Examples
5	*e-*	*i-*	*(e)ibbaale* ‘stone’
5a	*e-*	*∅-*	*(e)dagala* ‘medicine’
5b	*e-*	*li- (ri-)*	*(e)riiso* ‘eye’

Therefore, to detect the right noun class including such subclasses we needed an algorithm presented in the Dictionary Preprocessing notebook, which matches the noun to its class and separates the root.

The plural pairing class presented in the dictionary was simply added to the list of compatible forms alongside with its possible augmentation and diminutive noun classes based on countability. We decided to mark nouns without plural pairings as uncountable in order to speed up the process and avoid additional processing.

2. 3. 2. Verbs

Preprocessing of verbs turned out to be less complex but also less efficient, as we will see in the Estimation part.

Firstly, we detected regular words, i.e., those verbs, perfective forms of which did not involve any non-linear morphophonological processes described in the previous part. Also, we stored the vowel of the perfective suffix, which is said to harmonize with the last vowel of the root. By doing so, we avoided complicated rules and stored the vowel shown in the dictionary instead as an inflectional class.

For all the irregular verbs we extracted the prefix vowel and stored both, non-perfective, and perfective forms. Moreover, we tried to take into account one-syllable verbs, however, their morphophonology requires further development.

2. 4. Estimation

We used two sources to estimate our morphological parser and to find those mistakes we would need to consider, the ELAN annotated corpus and examples from the grammar, and extracted morphologically complex wordforms from each one. The ELAN corpus was fully preprocessed automatically, and the grammar examples required manual extraction, which is why one of them is big, but has a lot of errors, while the other one is small, but highly reliable. Testing on both gives us much information about the problematic patterns.

2. 4. 1. ELAN files preprocessing

This could be an effective corpus for the estimation; however, it does not have word alignment, which means that words are not matched to their glossing. And sometimes words that are written separately on the surface level happen to be glossed as a single one and vice versa. This caused many mistakes in the golden standard and made it much less reliable, as sometimes wordforms match their gloss set incorrectly.

After preprocessing, the golden standard based on the ELAN files included 4896 wordforms.

2. 4. 2. Examples from the grammar

We manually chose words from different examples from distinct sections in the grammar and preprocessed them into the same format as the ELAN corpus.

After preprocessing, the golden standard based on the examples included 113 wordforms.

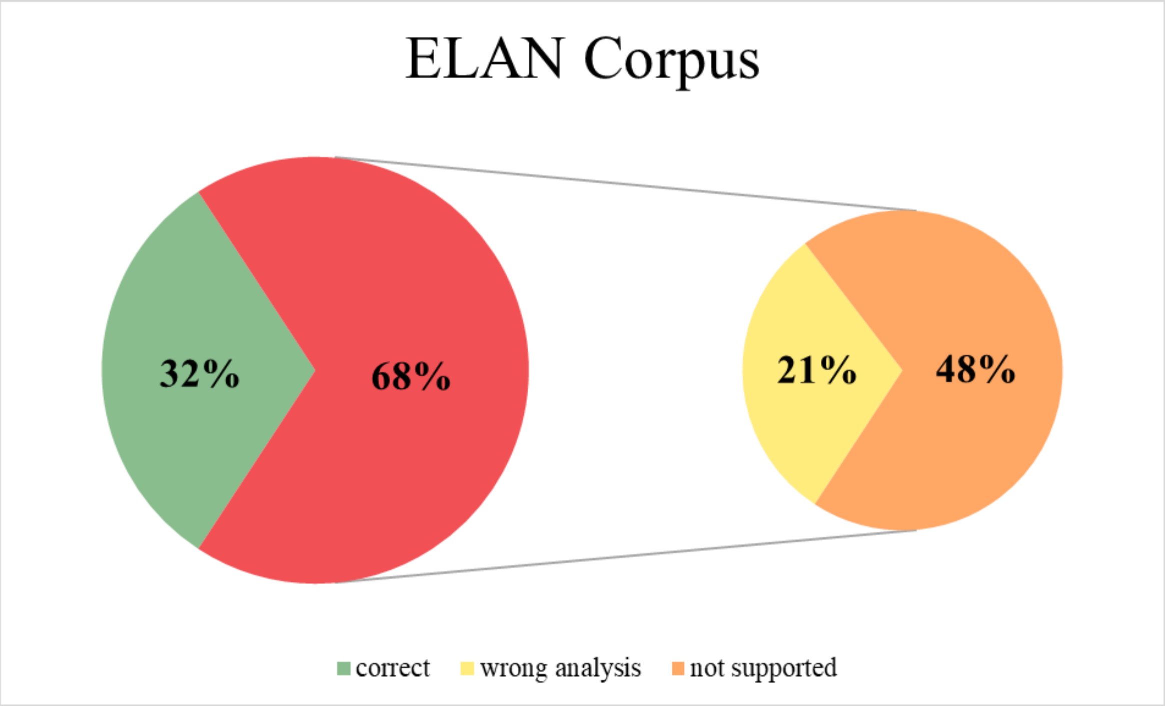
2. 4. 3. Estimation

The process of estimation was slightly different for our sources. In the case of the ELAN corpus we required all the glosses listed in the standard to be included in at least one of the analyses of the algorithm to count it as a correct parsing, whilst for our manually collected examples we only considered a result to be correct only in case all the glosses besides roots matched. We did not use the ELAN collection for a strict evaluation, because there are cases of morphemes being aligned to an incorrect word, so this mean was supposed to lower the impact of such mistakes.

During the evaluation we separately counted the cases when the algorithm could not suggest any analysis of a specific word form and those when the parser did not have the correct analysis among its suggestions. Moreover, we counted the glosses each of the failed words to have some representation of part of speech distribution.

ELAN Corpus

Figure 1. Percentage of errors and correct answers during estimation using ELAN Corpus.

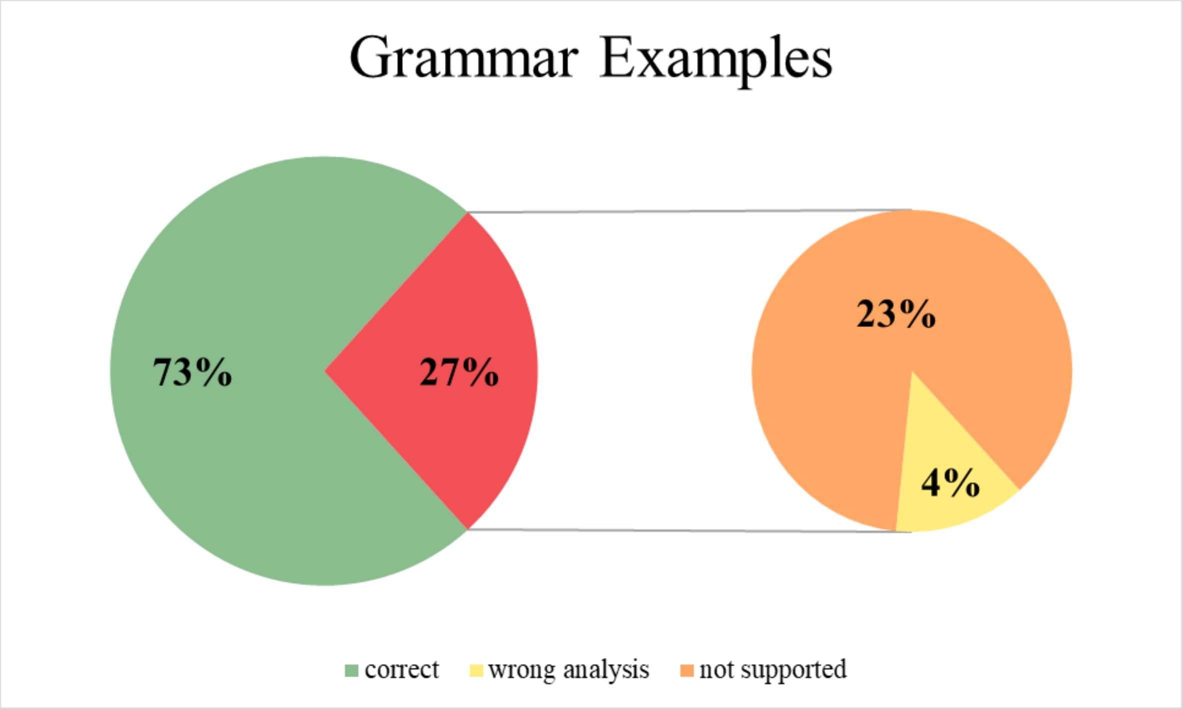


To get the right impression, we should keep in mind that the quality of this testing material is rather low, therefore this percentage cannot be considered a reliable one. Our goal was to detect the most problematic places using a big representative sample. According to the bar plots shown in **Appendix 1** and **Appendix 2**, verbs happen to be the most difficult part of speech to process, as half of the glosses in the first top-20 and 40% of the second one are verb-specific (such as FV, SBJ, PFV etc.). It was expected that the perfective form of verbs would be among the least successful to process, so unsurprisingly the gloss PFV is found in both charts. Some elements could be explained mainly with the frequency, as FV, which is by far the most common morpheme among verbs. In the framework of our future work, we will pay much attention to errors related to TAM forms (PROG, PST, NAR and such).

There happen to be no noun-specific glosses, as noun classes are also used in the verb system and adjectives to express agreement and augment prefix can also be found in its modifiers.

Grammar Examples

Figure 2. Percentage of errors and correct answers during estimation using Examples from the grammar.



Here, we get more representative data owing to the manual check of the material, therefore, this percentage can be considered a valid estimation.

Analysis of the most problematic morphemes of unsupported words presented in **Appendix 3** does not add much to what has been mentioned before. Moreover, it is quite likely that there was some sort of bias in the sample since it is rather small. For instance, we see locative noun classes (LOC) pop up in the higher positions of the plot despite them being not that common (cf. their position on the ELAN Corpus plot). Nevertheless, this estimation covered all the relevant topics and is easy to apply for fast improvements that can be led to the accurate list of cases of incorrect analysis.

There were only 9 glosses of the words, the correct parsing of which was not suggested by the algorithm. This shows that these are specific cases and makes it easy to improve in the future.

3. Web-interface

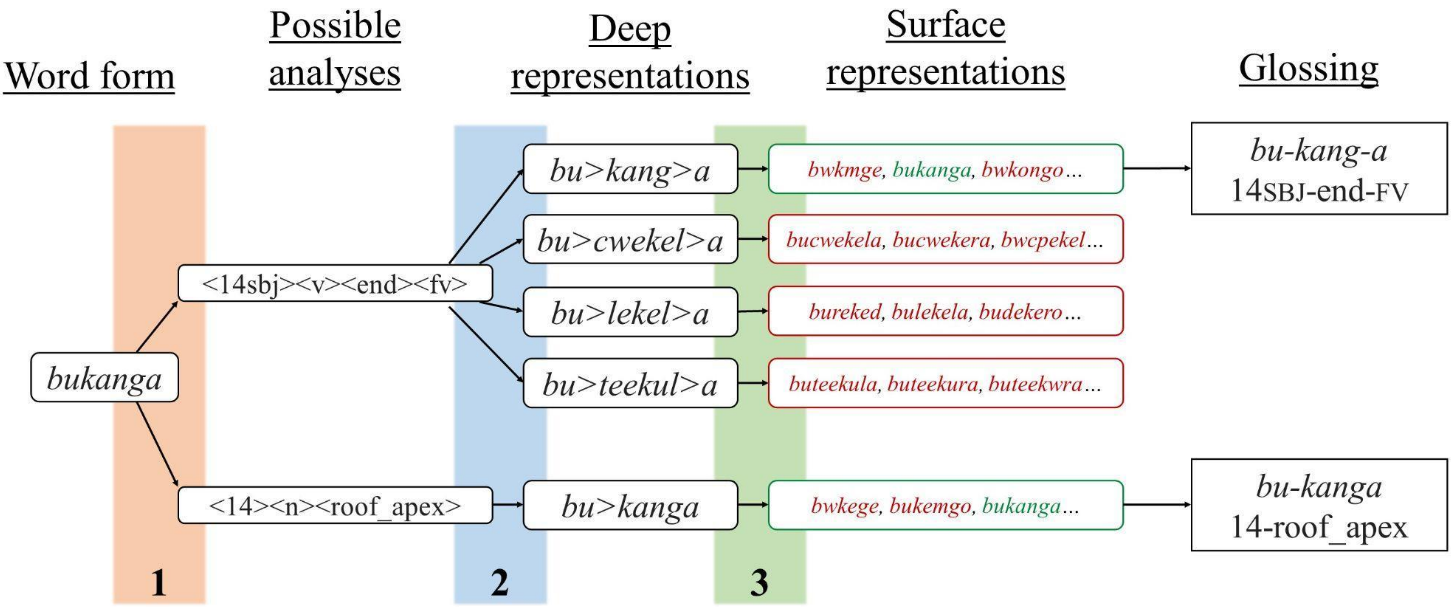
At this point we agreed upon a minimalistic design of the interface. The search page consists only of a search box and the title, and the result page lists all the possible options of glossing.

Seemingly, HFST formalisms are not designed for an end-to-end glossing mechanism, so the way we had to develop it may seem inefficient. It includes three queries to the different parts of the transducer:

1. Basic morphological parsing query. An input is a **wordform** and an output is a list of possible alternatives each being a list of **glosses**.
2. Glosses to deep representation of morphemes. An input is a list of **glosses,** and an output is a list of possible **deep representations** of these glosses.
3. Deep representation to surface representations. An input is a **deep representation,** and an output is a list of possible **surface representations**, or **wordforms**.

On the scheme below these steps are represented with the arrows with coloured background boxes.

Scheme 1. The processing of the wordform *bukanga*.



After the third query (the green box) we get a list of possible surface representations. If this list contains an initial word form, we show the corresponding analysis on the page, otherwise we get rid of it. The biggest problem is that on this stage we get those surface representations overgenerated which makes this processing slow.

As soon as we upload the web interface on the server, it will be easily accessible with automatic means.

4. Conclusion

During this project we succeeded to build a working prototype of a morphological parser for the Ruuli language.

At this point, its accuracy can be estimated at 73% and further improvements are yet to be made, as we already have all the information about the flaws collected during our tests.

Finally, it can already be used by the scientists to gloss word forms via our web-interface, that we will publish on our domain in the nearest future or that can be already set up locally.

Sources

Lewis et al 2015 — Lewis M. P., Simons G. F., Fennig Ch. D. *Ethnologue: Languages of the World, Eighteenth edition*. Dallas, Texas: SIL International. 2015

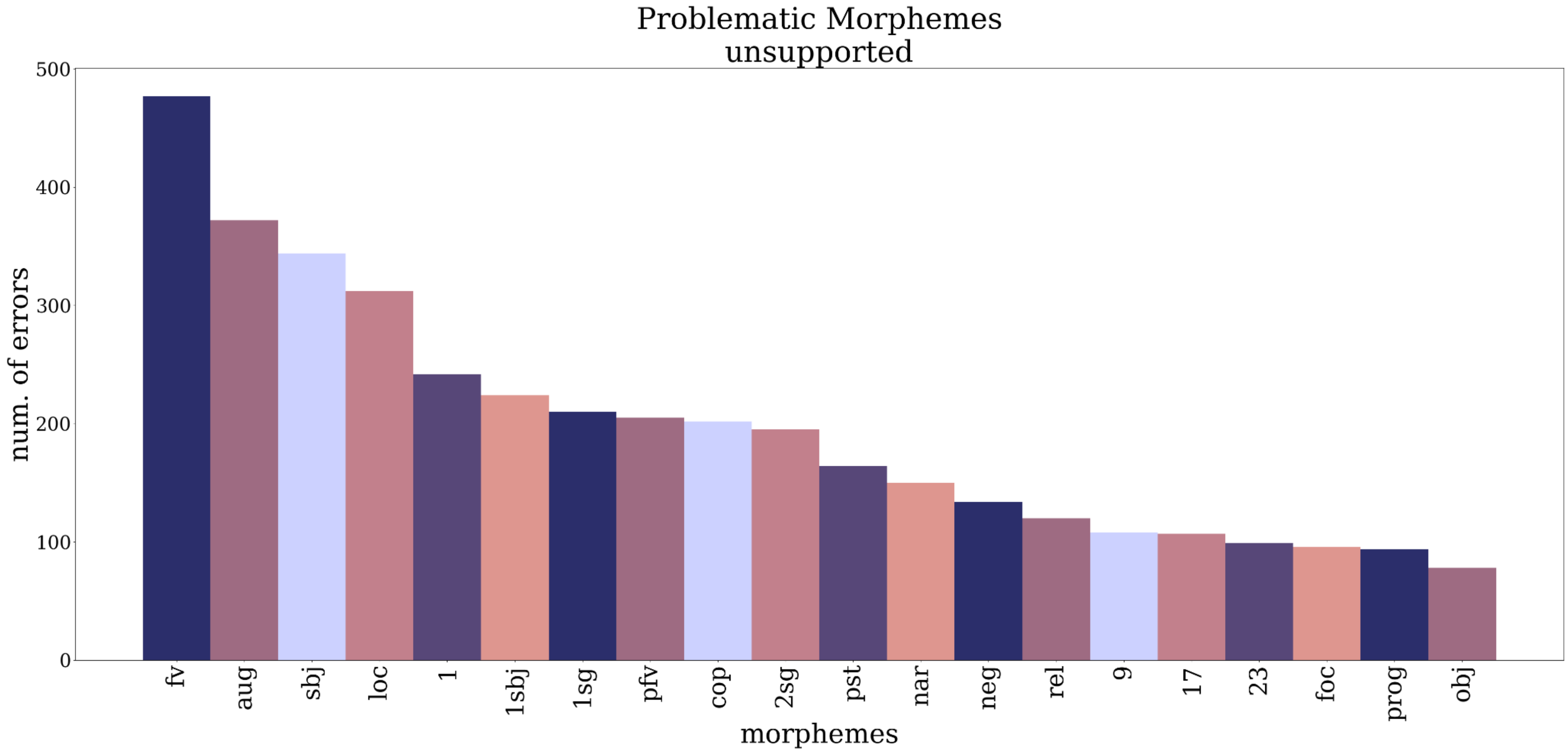
Lindén et al 2011 — Lindén K., Axelson E., Hardwick S., Pirinen T. A., Silfverberg M. HFST-framework for compiling and applying morphologies // *International Workshop on Systems and Frameworks for Computational Morphology*. Berlin: Springer, 2011. pp. 67–85

Lutz 2021 — Lutz M. Noun Classes and Plurality in Bantu Languages // Hofherr P. C., Doetjes J. (eds.). *The Oxford Handbook of Grammatical Number*. Oxford: Oxford University Press. 2021

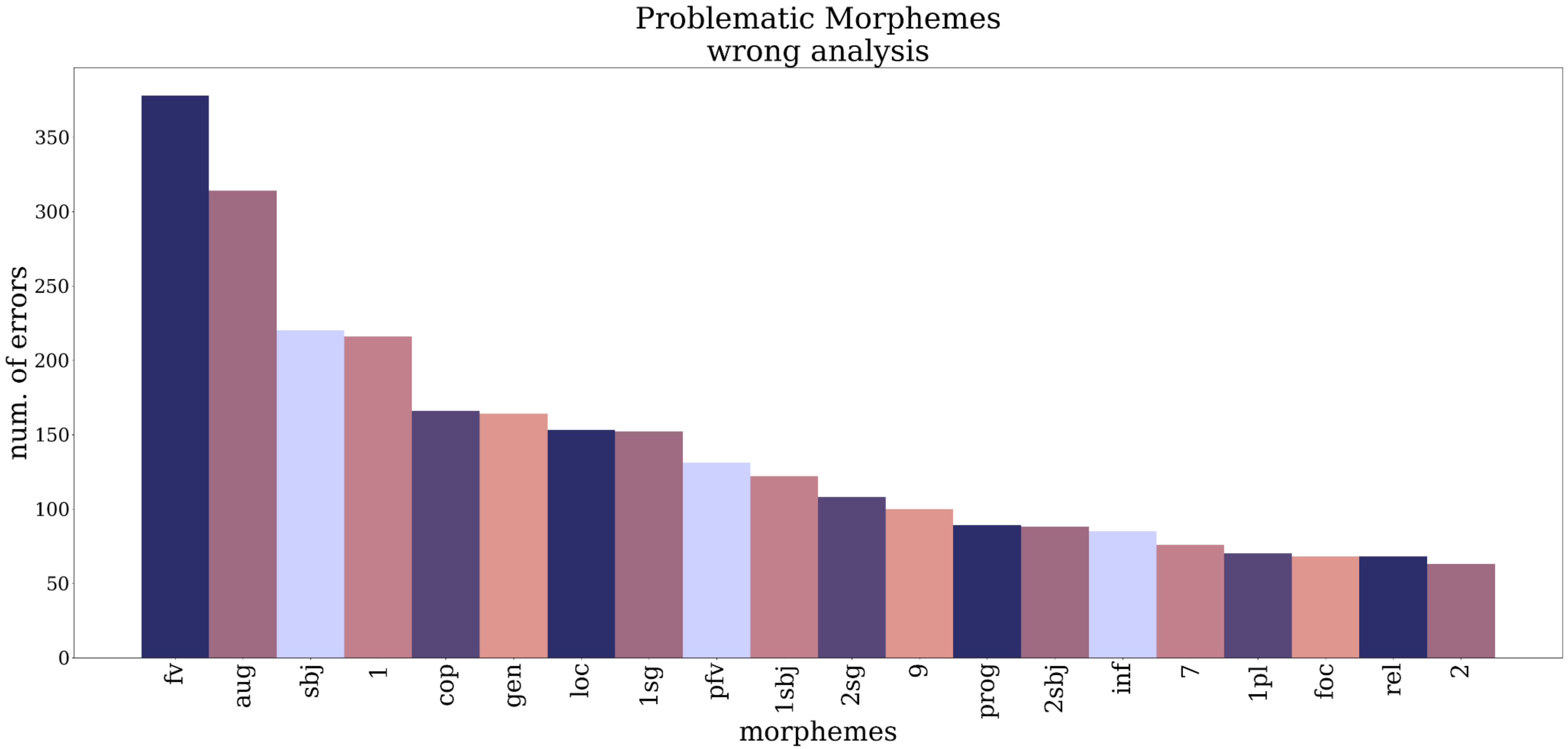
Molochieva et al 2021 — Molochieva Z., Namyalo S., Witzlack-Makarevich A. Phasal Polarity in Ruuli (Bantu, JE.103) // Kramer R. (ed.). *The Expression of Phasal Polarity in African Languages*. Berlin, Boston: De Gruyter Mouton, 2021. pp. 73-92

Namyalo et al 2021 — Namyalo S., Witzlack-Makarevich A., Kiriggwajjo A., Atuhairwe A., Molochieva Z., Mukama R., Zellers M. *A dictionary and grammatical sketch of Ruruuli-Lunyala*. Berlin: Language Science Press. 2021

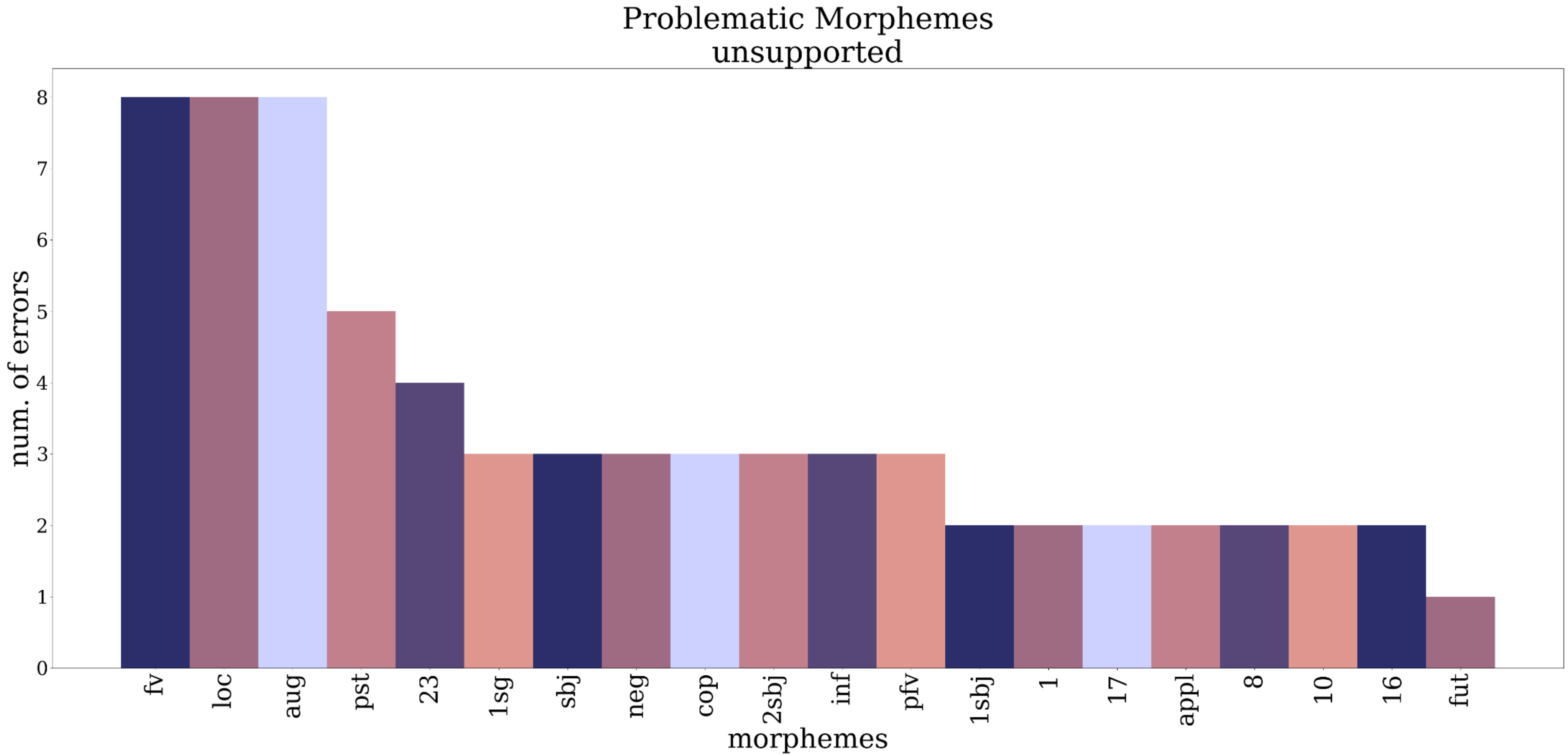
Appendix 1. 20 most popular morphemes in unsupported words from ELAN Corpus



Appendix 2. 20 most popular morphemes in incorrectly analysed words from ELAN Corpus



Appendix 3. 20 most popular morphemes in unsupported words from grammar examples


Problematic Morphemes
unsupported
num. of errors
fv
loc
aug
pst
1sg
sbj
neg
cop
2sbj
inf
pfv
1sbj
appl
fut
morphemes