CoDeRooMor : A new dataset for non-inflectional morphology studies of Swedish

Volodina, Elena

Northern European Association for Language Technology (NEALT)
2021

Volodina , E , Mohammed , Y A & Lindström Tiedemann , T 2021 , CoDeRooMor : A new dataset for non-inflectional morphology studies of Swedish . in Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoLaLiDa) . NEALT Proceedings Series , no. 45 , Northern European Association for Language Technology (NEALT) , Linköping , pp. 178-189 , Nordic Conference on Computational Linguistics , Reykjavík , Iceland , 31/05/2021 . < https://www.aclweb.org/anthology/2021.nodalida-main.18.pdf >

http://hdl.handle.net/10138/339476

CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish

Elena Volodina, Yousuf Ali Mohammed
University of Gothenburg / Sweden
elena.volodina@gu.se
yousuf.ali.mohammed@gu.se

Therese Lindström Tiedemann
University of Helsinki / Finland
therese.lindstromtiedemann@helsinki.fi

Abstract

The paper introduces a new resource, *CoDeRooMor*,[1] for studying the morphology of modern Swedish word formation. The approximately 16.000 lexical items in the resource have been manually segmented into word-formation morphemes, and labeled with their categories, such as prefixes, suffixes, roots, etc. Word-formation mechanisms, such as derivation and compounding have been associated with each item on the list. The article describes the selection of items for manual annotation and the principles of annotation, reports on the reliability of the manual annotation, outlines the annotation tool Legato, and presents the dataset and some first statistics. Given the "gold" nature of the resource, it is possible to use it for empirical studies as well as to develop linguistically-aware algorithms for morpheme segmentation and labeling (cf. statistical sub-word approach). The resource is freely available through Språkbanken-Text.[2]

1 Introduction

Linguistic complexity is a fascinating phenomenon that influences language perception, language learning and language production (cf. Housen et al., 2019; Bentz et al., 2016; Newmeyer and Preston, 2014). It has been studied at different levels and with different intentions, for example from a typological perspective (e.g. Gutierrez-Vasques and Mijangos, 2020) or from a computational perspective (e.g. Branco, 2018).

Linguistic complexity also varies between individual users of the same language, which makes it possible to use linguistic indicators to differentiate betweeen language typical of advanced language users as opposed to, for instance, children or beginner learners (De Clercq and Housen, 2017; Brezina and Pallotti, 2019; Pilán and Volodina, 2018).

From a second language (L2) perspective there is a need to be able to follow how the morphological complexity develops in the learner language (e.g. Pienemann and Kessler, 2012; Bonilla, 2020) for instance by following how the learner acquires more inflectional forms in the language but also by seeing how their vocabulary growth can be related to the acquisition of rules of word formation. The latter is a rather underdeveloped research area and it is that which has been our focus in developing the *CoDeRooMor* resource – we want to be able to follow how word families (cf. Bauer and Nation, 1993) grow and how awareness of word-formation mechanisms develops in language learners.

Morphology, as one of the dimensions of linguistic complexity, covers *word formation* in terms of compounding and derivational morphology as well as *inflectional morphology*, such as grammatical affixes that words take to reflect number, definiteness, gender, etc. Most publications on morphological complexity deal with studies of the inflectional dimension of morphology (e.g. Brezina and Pallotti, 2019; Forsberg and Bartning, 2010), with a few rare exceptions (e.g. Bolshakova and Sapin, 2020), which is not surprising. While automatic text annotation pipelines are able to process inflectional morphology (cf. morpho-syntactic descriptors available for corpora in the Korp search interface (Borin et al., 2012, 2016)), there is a lack of corpora containing analysis of the morphemes consituting the word lemmas. This is due to the absense of gold standard resources that can be used for training automatic tools (e.g. Kettunen, 2014). This is hypothetically also the reason why we rarely find lexical resources organized

[1] *CoDeRooMor* - **Co**mpounding, **De**rivation, **Roo**t **Mor**phology (and more)

[2] https://spraakbanken.gu.se/en/resources#refdata

by word family principles (cf. Bauer and Nation, 1993), even though there is a clear interest in that kind of resources in connection to vocabulary testing (e.g. Sasao and Webb, 2017) and psycholinguistic and cognitive research (e.g. Amirjalili and Jabbari, 2018).[3]

In the currently pursued project, *Development of lexical and grammatical competences in immigrant Swedish*,[4] funded by Riksbankens Jubileumsfond, we are looking for ways to characterize the language typical of second language (L2) learners of Swedish from different perspectives based on the analysis of two learner-specific corpora (see Section 3). Based on those corpora, we have generated a sense-based wordlist, Sen*Lex, manually segmented each item on the list into morphemes and labeled those for their morpheme categories (Section 4). The intention is to use this resource for empirical studies as well as for the development of automatic morphological segmentation and consequent morpheme classification for Swedish. We expect this type of annotation to facilitate deeper studies into lexical and morphological complexity, language acquisition patterns, associative learning mechanisms and the like. The resource can also be of interest in pedagogical studies and applications.

2 Related work

Morphemic segmentation is an important NLP task which is applied to machine translation, cognate identification, linguistic typological studies, and the like (Sennrich et al., 2015; Miestamo et al., 2008). The task of morpheme segmentation consists of the identification of morpheme boundaries within a word, and classifying them by their category. Most work has been focused on inflectional morphology and on classification of the endings by their syntactic and grammatical functions, such as gender, number, tense indicators (e.g. Cotterell et al., 2019).

Identification of word formation morphemes (roots, suffixes, prefixes) and their subsequent classification is a more complicated task, and until recently most approaches have been targeting only morpheme boundary identification using unsupervised or semi-supervised approaches, for example a language independent approach taken in Morfessor (Creutz and Lagus, 2007; Smit et al., 2014) or sub-word identification techniques (e.g. Gutierrez-Vasques and Mijangos, 2020).

Only recently have datasets with labeled data started to appear, and depending on their size, neural networks are used for experimentation with more complicated tasks including both morpheme segmentation and labeling of word formation morphemes (e.g. Bolshakova and Sapin, 2020; Sorokin and Kravtsova, 2018).

Morphology has not been one of the major strands of research on Swedish, neither as an L1 (native speaker language) nor as an L2 (second language learners). There also has not been a lot of interest in the development of tools and resources in relation to Swedish morphology except for Saldo morphology (Borin et al., 2013) which is used in annotation of Swedish texts and which primarily includes inflectional paradigms. Due to its language independence, Morfessor (Smit et al., 2014) offers a possibility to annotate words morphologically in any language and works relatively well on concatenative languages, including Swedish. The output consists of several suggestions for word segmentation into morpheme constituents.

In recent years interest has increased in finding ways to study different forms of complexity in connection to second language acquisition and learner corpora (Housen et al., 2019). However as Housen et al. say, morphological complexity has not been at the centre of attention. When studies have looked at morphological complexity they have also tended to focus primarily on inflectional morphology.

The resource we present in this paper is aimed at non-inflectional morphology of Swedish and can be used in a variety of NLP and linguistic tasks, including within the second language acquisition domain, and is filling a gap by offering a richly annotated dataset for morphological studies.

3 Item selection

To limit the annotation work to only the most relevant items, which in our context means items of relevance for second language learners of Swedish, we have used two source corpora:

- COCTAILL (Volodina et al., 2014), a corpus of coursebook texts that learners of Swedish

[3] e.g.https://www.ltu.se/research/subjects/teknisk-psykologi/nyheter/Nytt-projekt-om-barns-lasformaga-1.203355

[4] https://spraakbanken.gu.se/en/projects/l2profiles

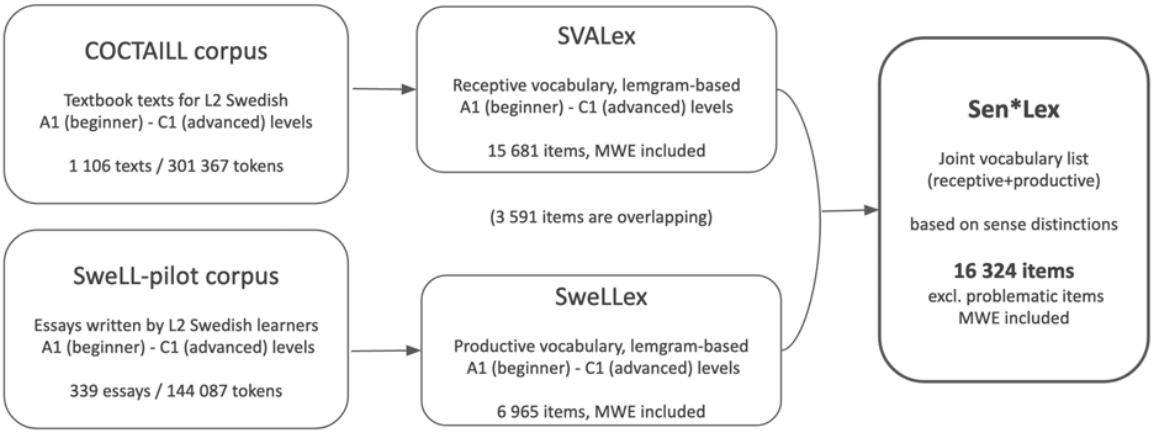

Figure 1: Selection of items for morphological annotation

as a second language (L2 Swedish) read as part of their proficiency courses, and

- SweLL-pilot (Volodina et al., 2016a), a corpus of essays written by adult learners of Swedish as a second language

Both corpora have indications of levels of proficiency according to the Common European Framework of Reference, CEFR (Council of Europe, 2001), and contain texts and essays at five out of six defined levels: A1 (beginner), A2, B1, B2, C1 (advanced).

From the two corpora, two lists of lemgrams (i.e. baseforms of the words + their corresponding parts of speech, POS) have been generated, namely:

- SVALex (François et al., 2016), consisting of L2 Swedish receptive vocabulary, and
- SweLLex (Volodina et al., 2016b), containing L2 Swedish productive vocabulary.

The approach used in the generation of the above lists has been reused by us to generate a new list based on senses (i.e. a list where each entry corresponds to a unique combination of baseform+POS+sense) once the pipeline for Swedish could assign word senses (Nieto Piña, 2019) based on Saldo senses (Borin et al., 2013). This work resulted in *Sen*Lex* (a sense-based variant of SweLLex and SVALex in one), a publication on which is currently under preparation. The non-problematic items of this latter list have been used for the morphological annotation.[5]

Figure 1 shows the basic information about the two source corpora and the three vocabulary lists. Sen*Lex includes both single-word items and multi-word expressions (MWEs), and contains word senses coming from both learner essays and course books. A certain amount of items overlap, i.e. occur in both corpora; whereas some items are homographs within the same part of speech (cf. `vara, verb` – Eng. 'be' and Eng. 'last'), but have several distinct senses. These latter items may have identical morphological analysis despite having several entries in the list, but it is also possible that they have different morphological annotation as is the case with the verb `vara`, where the root is `var-` in both lemmas but the final `-a` is seen as derivational in one sense and inflectional in the other, since `vara` (Eng. 'be'), has the imperative form `var!` and the verb `vara`, (Eng. 'last') has the stem and imperative form `vara!`, not that you are ever likely to use it in the imperative.

The *CoDeRooMor* morphological dataset that we are presenting is, thus, not all-covering for modern Swedish. However, given the nature of second language learning, the most central items should be represented in the list, therefore making it relatively comprehensive. [6]

[5]By problematic items we mean the items that have automatically been assigned multiple lemgrams or failed to be assigned a lemgram. These items are left for future work.

[6]We would also like to note that the set includes c. 500 triplets consisting of lemgrams which are verbs and part-of-speech-tag participle, e.g. `cykla` 'to ride a bike' + PC (participle). Since we annotate lemgrams these have then been annotated as the lemma of the verb, rather than one of the participles. We did look into annotating them as participles, but in fact each of these items can include occurrences in the data which are a combination of present participles or past participles, or even supine forms (a form etymologically re-

Word formation	Definition	Example
Abbreviation	words consisting of the initial components of a word or several words, including chemical abbreviaions and some blends	AB (aktiebolag) (cf. Eng. 'ltd' = 'limited'), Au (Sw. guld, Eng. 'gold')
Compound	words formed by adding together two stems	skol+bok ('school book')
Derivation	words formed by adding a prefix or a suffix to a stem	sorg**lig** ('sad')
Lexicalized form	words that cannot be reduced to baseforms, e.g. MWEs	Aftonblad**et** (name of a tabloid), järnspik**ar** (a swearword)
Root lexeme	words consisting of a root only or a root and an inflectional suffix	bok ('book'), adjö ('goodbye'), and**e** ('spirit')
Unknown	reserved for difficult or uncertain cases including most first names	alzheimers (name of a disease), kalender ('calendar')

Table 1: Taxonomy of word formation mechanisms with definitions and examples

4 Annotation principles

The aim of the morphology annotation work consisted in

- segmenting each lexical item (lemgram+POS+sense) into morphemes
- assigning a word formation description to the item according to a taxonomy (Table 1)
- categorizing each morpheme according to a taxonomy of morpheme categories (Table 2).

The items were analysed at the lemgram level and hence the work did not include annotation of inflectional forms/morphemes, with a few exceptions (see below).

For example, `oändlighet, noun` ('infinity') was

1. segmented into four morphemes
   `o-änd-lig-het`
2. each morpheme received a label:

```
o: prefix
änd: root
lig: derivational suffix
het: derivational suffix
```

3. the word formation of the item was labeled as `derivation`

The taxonomy of morphemes is presented in Table 2. Most of the categories are self-explanatory, but some need to be explained.

• The category of *real root* should not be taken as representing an actual morpheme, but is used to catch cases of alternative spellings of the same root, and hence a form of allomorphy. This was done so that we could collect all words with the same root, including alternative root spellings, into a *word family* to create a word family resource for L2 Swedish (cf. Bauer and Nation, 1993).

Morph. category	Explanation	Example
p	derivational prefix	**för**djupa
r	root (orthographic)	**kaot**isk
rr	real root	**kaos** (kaotisk)
s	derivational suffix	kaot**isk**
f	infix*	kedj**e**brev
i	inflectional suffix	i_höst**as**
?	unknown	ir**on**i

Table 2: Taxonomy of morpheme categories and examples. * Swe. *fogemorfem*

• The category of *inflectional suffix* was added to cover some suffixes that change in other inflectional forms in the paradigm, e.g. as the final morpheme -a in `skol`**`a`** ('school', noun) since the plural is `skol`**`or`** and the compounding stem is also simply `skol`, e.g. `skol-gård` ('school yard'); and also the final morpheme -a in `läs`**`a`** since it is not part of the imperative, which in Swedish is usually seen as the verb stem `läs!` and nor is it part of the tense inflection `läs`**`er,`** `läs`**`te,`** `läs`**`t`**. Furthermore, we needed to catch cases of lexicalized forms that are not reducible to the (otherwise existing) baseforms, e.g. `järnspik`**`ar`** (a swear word literally meaning 'iron nails'). Yet another reason for this category was the presense of multi-word expressions, e.g. `i_det_stor`**`a`**`_hel`**`a`** ('in general'), where some of the constituent parts are always used in an inflected form whereas other parts might be possible to inflect.

• During the annotation process an additional category - question mark <?> - was introduced for dubious cases that needed further discussion, e.g.

lated to the past participle but which only occurs in the past tense with the auxiliary verb `ha` 'to have'. We will return to these items in future work.

a in a-kassa ('unemployment benefit fund') or the on in ironi, ironisk ('ironic, ironical'). In most cases later comparisons helped resolve these issues and enabled the classificiation of the morpheme into one of the main morpheme categories.

The taxonomy of word formation mechanisms follows from Table 1, and is based on SAG (Teleman et al., 1999) and Haspelmath (2002). Where a word was a derivation based on a compound (e.g. all-var-lig, 'serious') or a compound which consisted partly of a derivation (e.g. å-bäk-e, 'monstrosity'), only word formation mechanism that gave us the final word was annotated, i.e. all-var-lig was annotated as a derivation and å-bäk-e as a compound. Detailed description of our annotation principles is available in our guidelines.[7]

To prepare a reliable resource for analysis of Swedish morphology, two authoritative resources have been used for major guidance in our annotation work: *the Swedish Academy Grammar*, SAG (Teleman et al., 1999) and two contemporary lexicons from the Swedish Academy: *the Contemporary Dictionary of the Swedish Academy*, Svensk ordbok, SO and *The Swedish Academy Glossary*, Svenska akademiens ordlista, SAOL (Sköldberg et al., 2019), both available through https://svenska.se/. To get access to the information in the lexicons, the Swedish Academy further allowed us to match our list of items against the SO/SAOL database, download the aspects of interest and integrate them into our annotation tool where annotators could consult them or copy to work further based on that.[8]

Each item in the SO/SAOL database contains division markers within the word, indicating where two morphemes meet (see Figure 2). Dots and vertical lines are used as notations, where the vertical line has a higher priority and is seen as a major word boundary. However, no information is provided about exactly what each morpheme stands for, e.g. whether it is a derivational suffix,

Figure 2: SO-SAOL analysis

an inflection or a root. They also do not provide marking of the compounding / derivational infix (Swe. *fogemorfem*), since their notation has the primary goal to indicate to the user where a word can be hyphenated, and infixes are always then attached to the stem.

5 Annotation workflow and visualization

A team of three highly qualified annotators performed the annotation under the supervision of a project researcher. During the first month the three annotators went through a training period where they worked in parallel with the project researcher and annotated 100 new items per week plus reannotated items from previous weeks when need be. Based on the parallel items, comparisons were run on both the morphological analysis and the word formation assignment of each item. The guidelines were refined to take care of any remaining unclarities or disagreements.

The 400 items that were annotated by the 4 members of the morphology group during the training period have been used for calculating Inter-Annotator Agreement (IAA) which we report using Krippendorff's Alpha in Table 3. As can be seen from the Table, the agreement was consistently high during all training steps, with segmentation being the most agreed upon annotation type (0.93) and labeling the one with most disagreements (0.86). However, the agreement is considered to be acceptable with values over 0.75, and very high with values over 0.9, which makes us believe that the annotation of *CoDeRooMor* is very reliable and of high quality.

After annotating 400 items in parallel, the rest of the items were divided between the 3 annotators with weekly meetings to monitor progress and discuss problematic cases. Before each meeting the project researcher got a morpheme-based

[7] https://docs.google.com/document/d/1G5PEfeDEKg4dAZaupj6FmUUWBGiegiqagzXgTA3cDSY/

[8] We initially discussed an opportunity to use automatic pre-processing for detection of morpheme boundaries, e.g. using SWETWOL tool (Karlsson, 1992) or Morfessor (Creutz and Lagus, 2007), but instead opted for expert morpheme boundary indication performed by trained lexicographers and available through the SAOL/SO, as described in this subsection.

Annotation type	1-100	-200	-300	-400
Segmentation	0.87	0.87	0.89	0.93
Labeling	0.86	0.86	0.89	0.86
Segmentation+Labeling	0.85	0.85	0.87	0.88
Word formation	0.89	0.89	0.94	0.91

Table 3: IAA measure using Krippendorff's Alpha, reported for each 100-word portion.

comparison where disagreements and partial disagreements were identified, and these could then be checked by the project researcher and discussed as needed at the meeting. The team came up with a solution and amended the guidelines to ensure systematic annotation in the future. After each meeting the annotators were expected to correct any items that had been picked up in the comparison to adhere to the agreed or revised principles. The guidelines have been a living document all through the process. An article with a more detailed description of the linguistic principles of segmentation and labeling is under preparation (Lindström Tiedemann et al., In Prep.). Once the annotation was completed the project researcher once again checked disagreements, partial disagreements and also searched through the data consistently for certain strings to find possible inconsistencies. Based on this some further corrections were done according to the guidelines, e.g. if a suffix had been annotated as an inflectional suffix but should be a derivational suffix according to the guidelines this was corrected.

To ensure consistency of the annotation work, a tool for lexicographic annotation *Legato* (Alfter et al., 2019) was implemented within the framework of the project, see Appendix A. The tool requires annotators to log in to save their annotations. The functionality of the tool allows the annotator to see

- the current item as a lemgram, the lemgram part of speech, the part of speech tag and its first level of occurence in the source corpora
- sense descriptor from the Saldo lexicon
- examples from the corpora
- two fields where previous annotation for the annotator appears when available
- two fields with annotations from the Swedish Academy lexicons (SO and SAOL)
- a text area for entering "Current values" for the analysis

In addition, the tool offers possibilities to open guidelines, check a list of previously "skipped items" or click on supportive links (among others, COCTAILL corpus hits for the current item and SAOL/SO hits). To navigate between the items, it is possible to "jump" to another item at a certain numeric index, search for some specific items or filter items.

Furthermore, the tool also allows each annotator to download their own annotated words with time stamps for inspection of the results. The project researchers can, in addition, download the annotations from all annotators, to generate several types of comparisons and statistics, and download a full set of annotated words.

6 *CoDeRooMor* dataset description

The *CoDeRooMor* dataset (version 1.0) contains 16 230 analyzed lemgrams[9] representing 4 429 unique roots, 259 unique derivational suffixes, 155 unique prefixes and 12 unique binding morphemes (infixes), see Table 4. Table 4 shows statistics over all morphemes in the dataset with some examples, number of times these morphemes appear in the lexemes in the Sen*Lex list, number of times they are used in the running tokens in the COCTAILL corpus (coursebooks) and in the SweLL-pilot corpus[10] (essays).

The five most frequent root morphemes in the Sen*Lex items on the *CoDeRooMor* are:

- `ut` (313 words in the "family", each containing that root), e.g. `utbildning`, ('education')
- `i` (272 words), e.g. `i` (preposition), ('in')
- `för` (228 words), e.g. `överföra`, ('transfer')
- `upp` (225 words), e.g. `kolla upp`, `uppdrag`, ('check up', 'assignment')
- `till` (189 words), e.g. `tillbaka`, ('back')

If we instead look at the five most frequent root morphemes in the corpora, the most common in Coctaill are `ha` (13 933 words), `var` (13 597

[9] We started with 16 324 triplets (lemgram + POS + sense), but we had to invalidate some lemgrams which were incorrectly lemmatized and not found in the data when doublechecking. We found these items since they were unexpected in learner data and were therefore doublechecked in the corpora by the project researcher supervising the annotation.

[10] The calculations were performed on a new version of SweLL-pilot, from 2020, which contains an extended collection of essays compared to Volodina et al. (2016a), namely 490 essays and 156 988 tokens (as compared to 339 essays and 144 087 tokens in the 1st version)

Morpheme category	Unique count	Sen*Lex	COCTAILL	SweLL-pilot	Examples
root	4429	23 987	471 056	142 381	**mat**bord, **kär**leksaffär, sago**lik**
suffix	259	10 062	91 646	28 638	markn**ad**, kost**sam**, milit**är**
prefix	155	2 183	19 828	5 489	**kon**sonant, ny**re**noverad
infix	12	1 089	3 441	1 641	känn**e**dom, kvinn**o**rörelse
inflection	32	3 067	88 641	28 810	sak**er**_och_ting, Medelhav**et**, läs**a**

Table 4: Statistics per morpheme type in the three resources

words), `och` (13 154 words), `gå` (13 046 words) and `kunn` (12 528 words). In the SweLL-pilot they are `kunn` (6 962 words), `att` (5 007 words), `och` (4 737 words), `var` (4 726 words) and `jag` (4 690 words).

Examples of other frequent root morphemes are

• `liv` with 73 family members, e.g. `affärsliv, livmoder, leva_livet`, ('business life', 'uterus', 'live_the_life'); and

• `sam` with 53 family members, e.g. `samtal, samhällelig, sambo`, ('conversation', 'societal', 'partner').

On inspection we can see that these family groupings need to be refined to be separated further into proper "families", so that words containing unrelated homographic roots do not accidentally end up in the same family. To give one example, the `sam`-family at the moment contains both `samisk` ('Sami', adj.) and `samhälle` ('society'), which should be separated into two different families since a morpheme is the smallest meaningful unit in language and therefore each root should have only one meaning and homographs should be separated.

Taking a look at the most frequent derivational morphemes (prefixes and suffixes), we can see that in the annotated wordlist

• the most common prefixes in the wordlist are `för-` (380 words), `be-` (299 words), `o-` (256 words), `re-` (112 words), `pro-` (85 words) as in `förälder, besök, odjur, reagera, problem` ('parent', 'visit', 'beast', 'react', 'conference', 'problem')

• and the most common derivational suffixes in the wordlist are `-a` (1 894 words), `-er` (640 words), `-ning` (443 words), `-ig` (433 words), `-ar` (378 words) as in `idrotta, aktivera, utbildning, duktig, ägare` ('do sports', 'activate', 'education', 'smart', 'owner').

There are several prefixes that only occur once in the dataset (wordlist and corpora), e.g. `abs-, fysio-, ko-` as in `abstrakt, fysiologisk, koefficient` ('abstract', 'physiological', 'coefficient'). In cases such as `koefficient` it would be good to consider comparison to allomorphs such as `kon`, but this currently needs to be done manually. Some of the least common suffixes are `-ej, -enn` as in `pastej, persienn` ('paté', 'Venetian blind') which only occur once in the dataset (wordlist and corpora). In the dataset it is also possible to access the frequency in relation to the number of occurrences in the L2 corpora we work with.

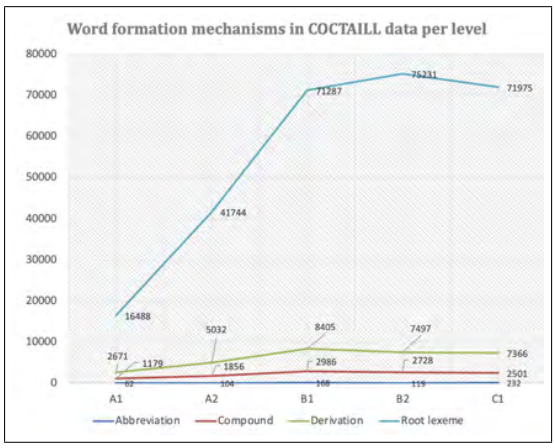

Figure 3: Statistics (raw count) over word formation mechanisms in the course book data

From the initial exploration of the word formation mechanisms in the two source corpora, we can see that *root lexemes* clearly dominate (Figures 3 and 4), followed by *derivation* and *compounding*. *Abbreviation* is hardly represented, nor are *lexicalized forms* that we haven't even included into the graphs. The hypothetical reason for the overrepresentation of *root lexemes* can be the fact that most frequent words in the language, namely prepositions, particles and conjunctions, are root lexemes and therefore add to the running statis-

Lemgram	Sense	POS	Analysis	Segment.	Pattern	RealRoot	WordForm	CEFR
adekvat..av.1	adekvat..1	JJ	p:ad r:ekv s:at	ad-ekv-at	p:r:s		derivation	C1
adla..vb.1	adla..1	PC	r:adl s:a	adl-a	r:s	rr:adel	derivation	B2
adel..nn.1	adel..1	NN	r:adel	adel	r		root_lexeme	B1
adelsman..nn.1	adelsman..1	NN	r:adel f:s r:man	adel-s-man	r:f:r		compound	B1
adjektiv..nn.1	adjektiv..1	NN	p:ad r:jekt s:iv	ad-jekt-iv	p:r:s		derivation	A2
adjö..in.1	adjö..1	IN	r:adjö	adjö	r		root_lexeme	A2

Table 5: *CoDeRooMor* dataset by lemgram, an excerpt

Morpheme	Identifier	Category	Frequency	Examples
a	s	suffix	1 605	leverera, lugna_sig, meritera, narkotika, pumpa, rasa
er	s	suffix	577	abdikera, intrigera, politiker, kritiker, motivera, tekniker
tid	r	root	128	arbetstid, nutid, skoltid, livstid, dåtid, deltid
ny	r	root	46	nyinköpt, nykokt, nykomling, nyligen, nymodighet, Nynäshamn
o	p	prefix	240	olaglig, olämplig, olik, olika, olikhet, oljud
re	p	prefix	105	reaktionstid, rebell, rebellisk, recensent, recensera, recension
s	f	infix*	803	fredstid, landsfader, riksbank, tvångsgift, landsdel, riksdag
o	f	infix*	76	vilopaus, sagobok, sannolik, sociolog, vilorum, typografi

Table 6: *CoDeRooMor* dataset by morpheme with examples, an excerpt. *Swe. *fogemorfem*

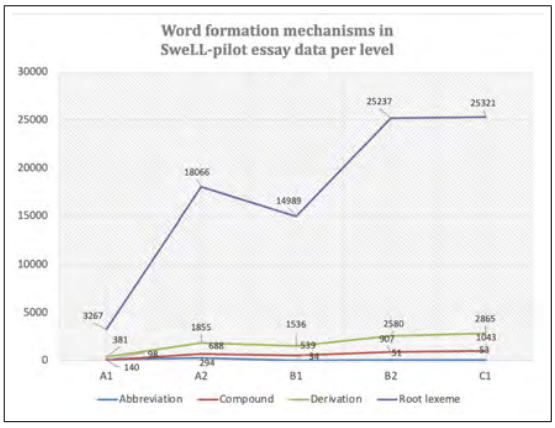

Figure 4: Statistics (raw count) over word formation mechanisms in the learner essay data

Item	A1	A2	B1	B2	C1	Total
word	1369	2689	4518	4440	3211	16230
morpheme	2457	5727	11318	11732	9292	40534
morpheme/word	1.79	2.13	2.51	2.64	2.89	2.50

Figure 5: Statistics over morpheme per word at different levels of proficiency

tics. In addition some words which could also be seen as derivations are currently seen as root lexemes since the final suffix falls in other inflectional forms and hence they are counted as root lexemes, e.g. `resa` 'to travel', cf. `resa` 'journey', since the rule was that annotators should usually select a word formation which fit with the first part of the annotation (segmentation and morpheme categorization).

Using *CoDeRooMor*, it is possible to trace the morphemic complexity of the words at different stages of language development. From Figure 5 we can see that the morphemic word structure is getting more complex as proficiency develops, with the average number of morphemes per new word (based on words which first occur at that CEFR-level in our data) growing from 1.79 at the beginner level till 2.89 at the advanced level.

The dataset can be downloaded as an excel file or as a file with comma separated values (csv file format). The information can be organized in several ways:

1. with lemgrams as the main lookup items (see Table 5). The associated information per lemgram consists of:
    - lemgram
    - sense indicator (Saldo-based)
    - part of speech
    - analysis by morpheme
    - real root (if applicable)
    - word segmentation boundaries
    - word morpheme patterns
    - word formation mechamism
    - the CEFR level (level of first occurence)
    - frequency information from COCTAILL, by level and in total (if applicable)
    - frequency information from SweLL-pilot, by level and in total (if applicable)

2. with morphemes as the main lookup item (see Table 6 for an example). The associated information consists of:

    - morpheme, e.g. `abs`
    - identifier, e.g. `p`
    - category, e.g. `prefix`
    - number of unique words containing that morpheme in the Sen*Lex list
    - list of words containing this morpheme-category (building a ”morpheme” family)
    - frequency in Sen*Lex by level and in total, if applicable (several columns)
    - frequency in COCTAILL by level and in total, if applicable (several columns)
    - frequency in SweLL-pilot by level (several columns)

The Legato annotation tool can compile some statistics and tables for overviews and visualization, which currently is only available for project researchers. In the future, we plan to make these functions open to all users, together with making this dataset available not only for download, but also for browsing (cf. English Vocabulary Profile, Capel, 2010).

The *CoDeRooMor* dataset can be freely downloaded from Språkbanken-Text.[11]

7 Future work

The *CoDeRooMor* resource offers promising possibilities for several types of research. Research questions with Linguistics and Second Language Acquisition domain are described in detail in Lindström Tiedemann et al. (In Prep.) and are mentioned briefly in the introduction to this article. With regards to pedagogical and applied research prospects, we are currently exploring how the items can best be linked together and presented to the public as a word family resource for use both in research and in teaching. The plan is that since Swedish uses both derivation and compounding frequently the resource will show all words which have a common root as a family and there will be information about how this relates to CEFR levels based on the corpora that we mentioned above. The dataset can be effetively used for Intelligent Computer-Assisted Language Learning research, for example for exercise generation or text complexity analysis.

To visualize the resource and support research into the non-inflectional morphology, we are working on a user interface for Swedish similar to the English Vocabulary Profile (EVP)[12] and Pearson GSE Teacher Toolkit.[13] The interface has a working title *Swedish L2 Profile* (SweL2P) and is integrated into the Lärka platform[14] (Alfter et al., 2018), at Språkbanken Text (Gothenburg, Sweden). The GUI will provide possiblilites to search, filter, browse and download various L2 Swedish datasets (lexical, morphological, grammar, including *CoDeRooMor*) generated as an output of the project.

We are currently also experimenting with automatic morpheme segmentation based on the *CoDeRooMor* dataset which is showing promising results and we hope that this might result in a new functionality in the Sparv pipeline (Borin et al., 2016) allowing automatic segmentation and labeling of morpheme categories for Swedish.

The ultimate aim is to analyze learner language in a more nuanced way, where analysis of word formation morphemes could help us to look deeper into lexical and morphemic complexity and to understand language acquisition and processing better. Type token ratio (TTR) has been often used as a way to measure lexical diversity, i.e. how varied the vocabulary in a text is (see e.g. McKee et al., 2000). However recently TTR has also been used as a means of studying morphological complexity (Gutierrez-Vasques and Mijangos, 2020; Kettunen, 2014). Gutierrez et al. also explore the possibility of studying morphological complexity through entropy and CRF in relation to typological comparisons of languages. Our intention is to apply similar techniques for analysis of learner language.

Acknowledgments

This work has been supported by a grant from the Swedish Riksbankens Jubileumsfond (Development of lexical and grammatical competences in immigrant Swedish, project P17-0716:1). We acknowledge the project assistants – Beatrice Silén, Stellan Petersson and Maisa Lauriala – for their thorough annotation work; and David Alfter for

[11] https://spraakbanken.gu.se/en/resources#refdata

[12] http://www.englishprofile.org/wordlists

[13] https://www.english.com/gse/teacher-toolkit/user/lo

[14] https://spraakbanken.gu.se/larkalabb/svlp

the initial implementation of the Legato tool and the generation of the Sen*Lex list. We thank the Swedish Academy and the SAOL/SO group at the University of Gothenburg for sharing parts of their valuable datasets with us.

References

David Alfter, Lars Borin, Ildikó Pilán, Therese Lindström Tiedemann, and Elena Volodina. 2018. From language learning platform to infrastructure for research on language learning. In *CLARIN Annual Conference 2018*.

David Alfter, Therese Lindström Tiedemann, and Elena Volodina. 2019. Legato: A flexible lexicographic annotation tool. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 382–388.

Forough Amirjalili and Ali Akbar Jabbari. 2018. The impact of morphological instruction on morphological awareness and reading comprehension of efl learners. *Cogent Education*, 5(1):1523975.

Laurie Bauer and Paul Nation. 1993. Word families. *International journal of Lexicography*, 6(4):253–279.

Christian Bentz, Tatjana Soldatova, Alexander Koplenig, and Tanja Samardžić. 2016. A comparison between morphological complexity measures: typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC). Osaka, Japan, December 11-17 2016*.

Elena Bolshakova and Alexander Sapin. 2020. An experimental study of neural morpheme segmentation models for russian word forms.

Carrie Bonilla. 2020. Processability theory and corpora. *The Routledge Handbook of Second Language Acquisition and Corpora*, page 201.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language resources and evaluation*, 47(4):1191–1211.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp-the corpus infrastructure of språkbanken. In *LREC*, pages 474–478.

António Branco. 2018. Computational complexity of natural languages: A reasoned overview. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 10–19, Santa Fe, New-Mexico. Association for Computational Linguistics.

Vaclav Brezina and Gabriele Pallotti. 2019. Morphological complexity in written l2 texts. *Second language research*, 35(1):99–119.

Annette Capel. 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

Bastien De Clercq and Alex Housen. 2017. A cross-linguistic perspective on syntactic complexity in l2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, 101(2):315–334.

Fanny Forsberg and Inge Bartning. 2010. Can linguistic features discriminate between the communicative CEFR-levels?: A pilot study of written L2 French.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. In *LREC*.

Ximena Gutierrez-Vasques and Victor Mijangos. 2020. Productivity and predictability for measuring morphological complexity. *Entropy*, 22(1):48.

Martin Haspelmath. 2002. *Understanding morphology*. Routledge.

Alex Housen, Bastien De Clercq, Folkert Kuiken, and Ineke Vedder. 2019. Multiple approaches to complexity in second language research. *Second language research*, 35(1):3–21.

Fred Karlsson. 1992. SWETWOL: A comprehensive morphological analyser for Swedish. *Nordic Journal of Linguistics*, 15(1):1–45.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Therese Lindström Tiedemann, Beatrice Silén, Maisa Lauriala, Stellan Petersson, Yousuf Ali Mohammed, and Elena Volodina. In Prep. Swedish morphology and Second language acquisition.

Gerard McKee, David Malvern, and Brian Richards. 2000. Measuring vocabulary diversity using dedicated software. *Literary and linguistic computing*, 15(3):323–338.

Matti Miestamo et al. 2008. Grammatical complexity in a cross-linguistic perspective. *Language complexity: Typology, contact, change*, 23:41.

Frederick J Newmeyer and Laurel B Preston. 2014. *Measuring grammatical complexity*. Oxford University Press, USA.

Luis Nieto Piña. 2019. *Splitting rocks: Learning word sense representations from corpora and lexica*. Doctoral Thesis, University of Gothenburg.

Manfred Pienemann and Jörg-U Kessler. 2012. Processability theory. *The Routledge handbook of second language acquisition*, pages 228–247.

Ildikó Pilán and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58.

Yosuke Sasao and Stuart Webb. 2017. The word part levels test. *Language Teaching Research*, 21(1):12–30.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Emma Sköldberg, Louise Holmer, Elena Volodina, and Ildikó Pilán. 2019. State-of-the-art on monolingual lexicography for Sweden. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 7(1):13–24.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University.

Alexey Sorokin and Anastasia Kravtsova. 2018. Deep convolutional networks for supervised morpheme segmentation of russian language. In *Conference on Artificial Intelligence and Natural Language*, pages 3–10. Springer.

Ulf Teleman, Staffan Hellberg, and Erik Andersson. 1999. *Svenska akademiens grammatik*. Svenska akademien.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, 107. Linköping University Electronic Press.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016a. Swell on the rise: Swedish learner language corpus for European reference level studies. *LREC 2016*.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016b. SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå*, 130, pages 76–84. Linköping University Electronic Press.