CEFR-graded morpheme family for L2 Swedish

Elena Volodina, Språkbanken Text, Sweden (presenter)

Based on work by: Elena Volodina, Yousuf Ali Mohammed, Therese Lindström Tiedemann and annotators

What is a morpheme family?

- Words that share morphological components form so-called **morphological families**, which are sets of words that are *related morphologically* (sharing a morpheme) and *semantically* (sharing a meaning)

  Nikolaev et al. (2019) [cognitive studies]

- The **derivation family** is defined as a network of *derivationally related lexemes.*

  Körtvélyessy et al. (2020) [typological studies]

- **Derivational families**, i.e. clusters of lemmas in derivational relationships.

  Zeller et al. (2013) [computational lexicography]

- Inclusion of a related form of a word within a **word family** depends on criteria involving *frequency, regularity, productivity and predictability*. These criteria are applied to English *affixes* so that the *inflectional affixes* and the most *useful derivational affixes* are arranged into a **graded** set of seven levels.

Bauer and Nation (1993) [L2 acquisition studies]

Bauer & Nation (1993)

		Word families	
2	develop develops developed developing	wood wood's woods wooded	bright brighter brightest
3	developable undevelopable developer(s) undeveloped	woody woodiest woodier woodiness	brightly brightish brightness
4	development(s) developmental developmentally		
5	developmentwise semideveloped antidevelopment	wooden	brighten
6	redevelop predevelopment	anti-wood	

Swedish L2 morpheme family

Morpheme ⬍	CEFR level ⬍	Morpheme Category ⬍	Lemgram ⬍	Sense ⬍	Word Class ⬍	Saldo Word Class ⬍	Word formation ⬍	Receptive ⬍ ⧉	Productive ⬍ ⧉
a	C1	Derivational suffix	skapa..vb.1	skapa..1	Participle (PC)	Verb (vb)	Derivation	0.13 (2)	9.33 (49)
ing	C1	Derivational suffix	meningsfull..av.1	meningsfull..1	Adjective (JJ)	Adjective (av)	Compound	0.06 (1)	8.95 (47)
is	A1	Derivational suffix	kompis..nn.1	kompis..1	Noun (NN)	Noun (nn)	Derivation	6.57 (17)	92.77 (38)
sk	B2	Derivational suffix	karolinsk..av.1	karolinsk..1	Adjective (JJ)	Adjective (av)	Derivation	0.00 (0)	5.39 (24)
in	B2	Derivational suffix	karolinsk..av.1	karolinsk..1	Adjective (JJ)	Adjective (av)	Derivation	0.00 (0)	5.39 (24)
ar	A1	Derivational suffix	lärare..nn.1	lärare..1	Noun (NN)	Noun (nn)	Derivation	13.13 (34)	51.27 (21)
skap	B1	Derivational suffix	äktenskap..nn.1	äktenskap..1	Noun (NN)	Noun (nn)	Derivation	0.45 (6)	8.33 (19)
en	B1	Derivational suffix	äktenskap..nn.1	äktenskap..1	Noun (NN)	Noun (nn)	Derivation	0.45 (6)	8.33 (19)

6

SweL2P, suffix -skap

Total rows : 53

Morpheme ⇅	CEFR level ⇅	Morpheme Category ⇅	Lemgram ⇅	Sense ⇅	Word Class ⇅	Saldo Word Class ⇅	Word formation ⇅	Receptive ⇅	Productive ⇅
skap	B1	Derivational suffix	äktenskap..nn.1	äktenskap..1	Noun (NN)	Noun (nn)	Derivation	0.45 (6)	8.33 (19)
skap	B2	Derivational suffix	kunskap..nn.1	kunskap..1	Noun (NN)	Noun (nn)	Derivation	1.72 (26)	0.90 (4)
skap	B1	Derivational suffix	partnerskap..nn.1	partnerskap..1	Noun (NN)	Noun (nn)	Derivation	0.15 (2)	1.31 (3)
skap	B1	Derivational suffix	vänskap..nn.1	vänskap..1	Noun (NN)	Noun (nn)	Derivation	0.60 (8)	0.88 (2)
skap	C1	Derivational suffix	föräldraskap..nn.1	föräldraskap..1	Noun (NN)	Noun (nn)	Derivation	0.57 (9)	0.19 (1)
skap	B2	Derivational suffix	medborgarskap..nn.1	medborgarskap..1	Noun (NN)	Noun (nn)	Derivation	0.26 (4)	0.22 (1)
skap	B2	Derivational suffix	kunskapsmässig..av.1	kunskapsmässig..1	Adjective (JJ)	Adjective (av)	Derivation	0.00 (0)	0.22 (1)
skap	B2	Derivational suffix	språkkunskap..nn.1	språkkunskap..1	Noun (NN)	Noun (nn)	Compound	0.20 (3)	0.22 (1)

Related resources (a few examples)

- CELEX (Dutch, German, English) lexicons based on L1 corpora
  - https://catalog.ldc.upenn.edu/LDC96L14
  - contain derivational and compositional structure, inflectional paradigms (among others)
  - protected by paywall
- DErivBase (German) - resource with morphological families
  - https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/derivbase/
  - based on German CELEX
- Derivational networks (40 EU languages)
  - https://www.degruyter.com/document/doi/10.1515/9783110686630/html
  - 3 word classes á 30 core root lexemes across languages
- Word formation networks (Polish, Czech, French, Spanish)
  - Lango et al. (2021); DeriNet (lexical network for Czech)
- Word families (English)
  - Bauer & Nation (1993) https://academic.oup.com/ijl/article-abstract/6/4/253/941347
  - Coxhead (1998) An Academic Word List (570 word families)
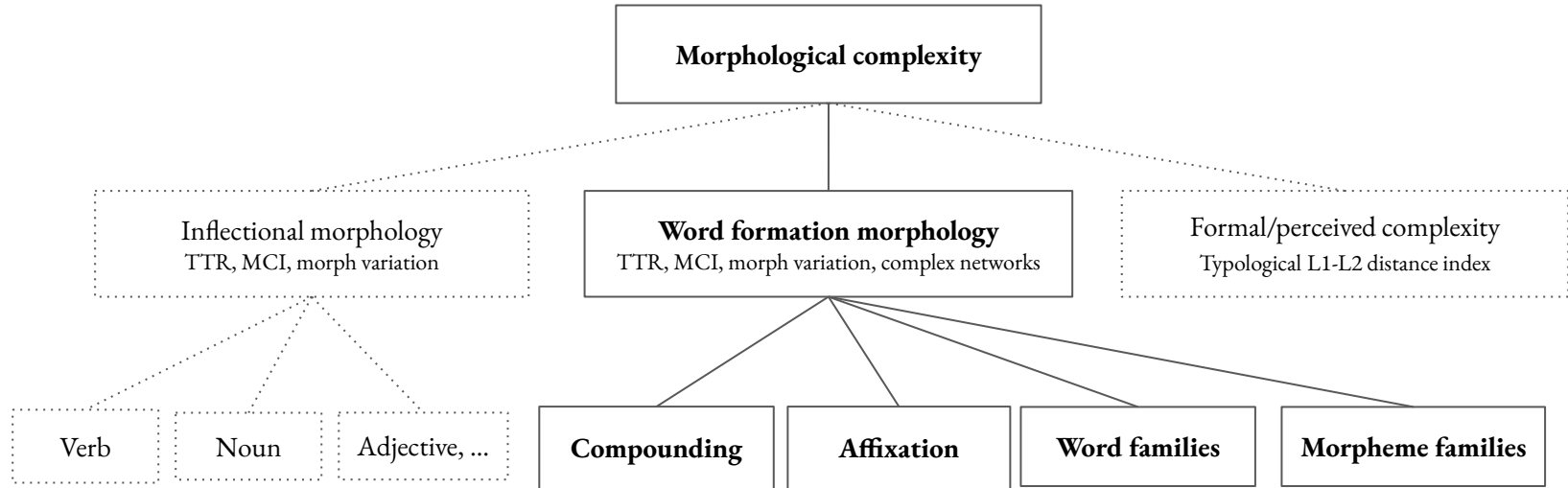
Why did we do it?

Complexity as a proxy of learner language development
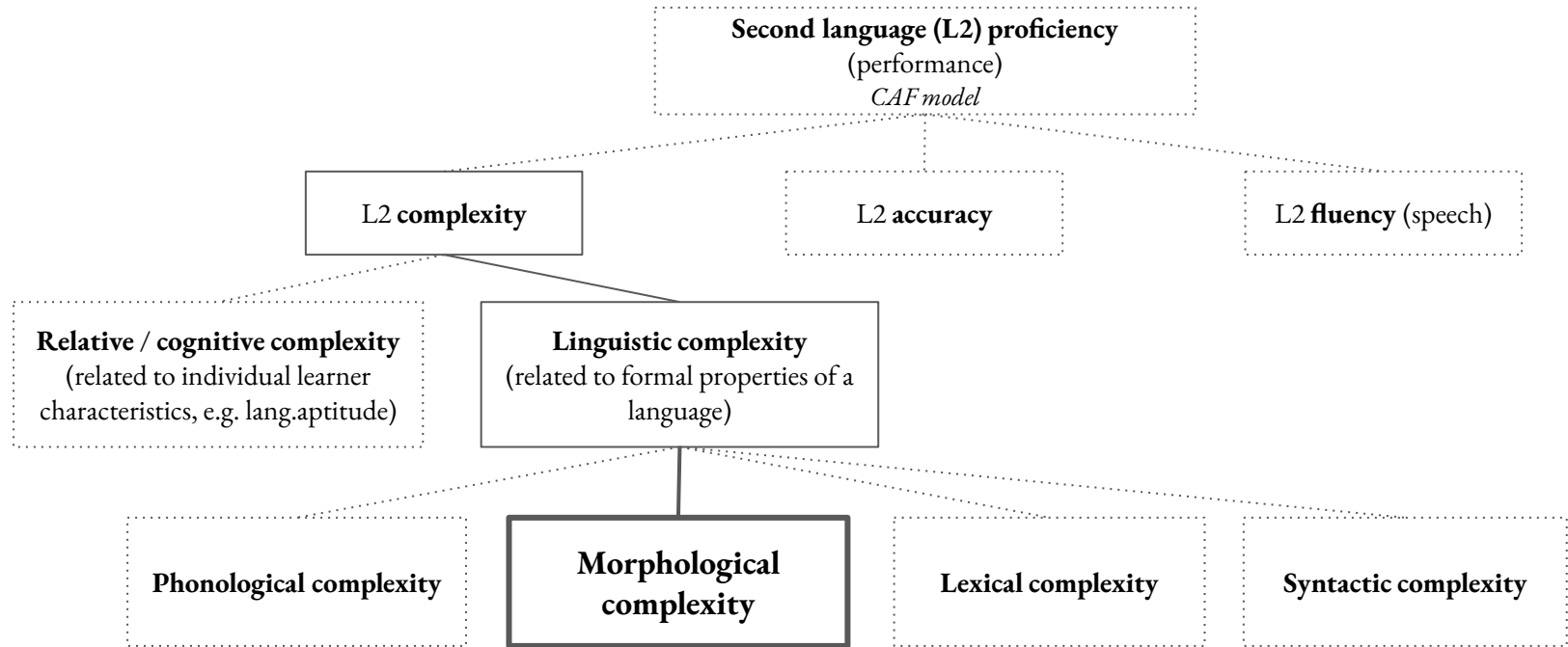
Our focus

Complexity measures (infl. morphology)

- Syntactic variety (Forster & Skenann, 1996; Ellis & Yan, 2005)
- Inflectional diversity index (Malvern et al., 2004)
- Normalized mean size of paradigm (Xanthos & Gillis, 2010)
- MCI - Morphological Complexity Index (Brezina & Palotti, 2016):
  - http://corpora.lancs.ac.uk/vocab/process_text_morph.php

Complexity measures (deriv. morphology)

- ...yet to be developed

Within L2 context

- Relation of morphological awareness and language proficiency
- Awareness of derivational morphemes
- Priming effect of roots/affixes on learning of new vocabulary (cf word families)
- Vocabulary testing
- Pedagogical applications / ICALL (e.g. exercises)
- ...

Outside L2 context

- Typological studies of languages
- Computational complexity of languages
- Psycholinguistic studies
- Cognitive studies
- Machine translation
- Cognate identification
- ...

L2 profiles for Swedish

L2 profiles for Swedish

L2P - Documentation

L2P - MWE experiment

L2P - experiment with word difficulty

Full name: Development of lexical and grammatical competences in immigrant Swedish, RJ, 2018-2020

Background

Sweden has a growing number of immigrants, the need for courses and coursebooks in Swedish as a second language (L2) is increasing, as is the demand for standardized tests and qualifications. This project intends to study the development of lexical and grammatical competences in L2 learners of Swedish.

General description

We intend to perform the study through two corpora: coursebook texts and learner essays, both marked up for proficiency levels according to the Common European Framework of References (CEFR). The corpora will be processed by computational methods, after which the results will be analysed by linguists, lexicographers, grammarians, teachers and language assessors - both linguistically, and based on theory of teaching, to find ways of identifying minimal or central (need-to-know) vocabulary and grammar scopes, as well as

Project duration
2018-01-01 – 2021-12-31

Project members

Elena Volodina
(Projektledare)
elena.volodina@svenska.gu.se

Therese Lindström
Tiedemann (Researcher)
University of Helsinki

Yousuf Ali Mohammed
(Systems developer)
yousuf.ali.mohammed@svenska.gu.se

David Alfter (Systems developer (part-time) & PhD student (part-time))
david.alfter@svenska.gu.se

https://spraakbanken.gu.se/en/projects/l2profiles

What is Swedish L2 profile?

Swedish L2 profile

(will be) freely available at: https://spraakbanken.gu.se/larkalabb/svlp

Lexical profile	Grammatical profile	Morphological profile

Sen*Lex ⧉

Search word	Word Class: Noun (NN) ▾	Saldo Word Class ▾	Noun declension ▾	Gender ▾

A1	A2	B1	B2	C1	C2

Tables	Graphs	Statistics	Download	Clear all

☐ Show only first occurrence   *Tables - description* ⧉   *Filters - description* ⧉

Total rows : 2221

CEFR level ⇅	Word ⇅	Lemgram ⇅	Sense ⇅	Word Class ⇅	Saldo Word Class ⇅	Receptive ⇅ ⧉	Productive ⇅ ⧉
A2	skola	skola..nn.1	skola..1	Noun (NN)	Noun (nn)	11.92 (82)	93.53 (214)
A2	kompis	kompis..nn.1	kompis..1	Noun (NN)	Noun (nn)	2.76 (19)	62.94 (144)
A2	dag	dag..nn.1	dag..1	Noun (NN)	Noun (nn)	23.54 (162)	52.45 (120)
A2	år	år..nn.1	år..1	Noun (NN)	Noun (nn)	37.35 (257)	40.21 (92)

19

Noun declension

Receptive

- 1. -or (blomma, blommor)
- 2. -ar (älv, älvar)
- 3. -er (dans, danser; (bok, böcker))
- 4. -r (sko, skor)
- 5. -n (bo, bon; nöje, nöjen)
- 6. - (bär, bär)
- Other (lexikon, lexika)

Productive

4. -r (sko, skor): 6

- 1. -or (blomma, blommor)
- 2. -ar (älv, älvar)
- 3. -er (dans, danser; (bok, böcker))
- 4. -r (sko, skor)
- 5. -n (bo, bon; nöje, nöjen)
- 6. - (bär, bär)
- Other (lexikon, lexika)
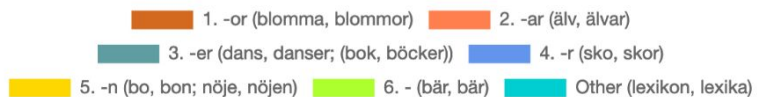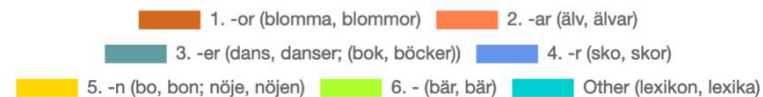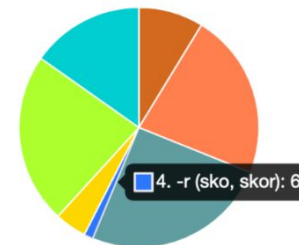
Noun declension	Receptive
1. -or (blomma, blommor)	211
2. -ar (älv, älvar)	475
3. -er (dans, danser; (bok, böcker))	572

Noun declension	Productive
1. -or (blomma, blommor)	43
2. -ar (älv, älvar)	110
3. -er (dans, danser; (bok, böcker))	124

Gender

Receptive

Legend: Common (en älv, en bil) · Common or neuter · Neuter (ett ägg, ett år) · No gender · Always plural - no gender

Productive

Common or neuter: 4
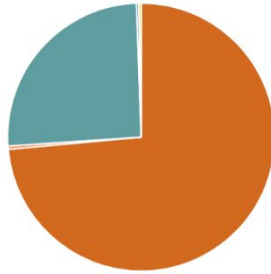
Legend: Common (en älv, en bil) · Common or neuter · Neuter (ett ägg, ett år) · No gender · Always plural - no gender

Gender	Receptive
Common (en älv, en bil)	1537
Common or neuter	8
Neuter (ett ägg, ett år)	533

Gender	Productive
Common (en älv, en bil)	346
Common or neuter	4
Neuter (ett ägg, ett år)	140

Total rows : 649

CEFR level ⇅	Word ⇅	Lemgram ⇅	Sense ⇅	Word Class ⇅	Saldo Word Class ⇅	Receptive ⇅ 🔲	Productive ⇅ 🔲
A1	skola	skola..nn.1	skola..1	Noun (NN)	Noun (nn)	11.20 (29)	249.02 (102)

Antal träffar: 102

« ‹ **1** 2 3 4 5 › »   Gå till sida [ ] av 5   **Visa kontext**

SPIN (stödjer ej utökad kontext)

min **skolan** ligger på N-gata på N-gata.
I **skolan** är 300 elever.
NN arbetar i **skolan** .
Ja bra i **skolan** pratar svenska.
Min **skolan** heter NN.
Min **skola** ligger på N-gata på N-gata.
Det finns fyra våningar i **skolan** .
i **skolan** eleverna finns 16.
Jag slutar **skolan** kl. 2:00.
Jag kommer **skolan** sprbangn och but.
Jag tycker om bild och Lesnär på music och Simu **skola** .

22

What is L2 Swedish Morphological Profile?

Swedish L2 Morphological profile

Lexical profile	Grammatical profile	Morphological profile

Word family

Morpheme family

Morpheme family

(will be) freely available at: https://spraakbanken.gu.se/larkalabb/svlp

Morpheme family

Search morpheme

ling

Search word

Search word

Morpheme Category

- ☐ Binding morpheme
- ☐ Derivational prefix
- ☑ Derivational suffix
- ☐ Inflectional ending
- ☐ Root morpheme
- ☐ Unclassified / unknown

Word Class

Word Class ▾

Saldo Word Class

Saldo Word Class ▾

Word formation

Word formation ▾

A1	A2	B1	B2	C1	C2

Tables	Graphs	Statistics	Download	Clear all

☐ Add column for related roots   ☑ Show only first occurrence   *Tables - desc...*

Total rows : 12

Morpheme ⇅	CEFR level ⇅	Morpheme Category ⇅	Lemgram ⇅	Sense ⇅	Word Class ⇅	Saldo Word Class ⇅	Word formation ⇅	Receptive ⇅	Productive ⇅
ling	A1	Derivational suffix	älskling..nn.1	älskling..1	Noun (NN)	Noun (nn)	Derivation	1.16 (3)	0.00 (0)
ling	A2	Derivational suffix	Moderata_samlingspartiet..pmm.1	Moderata_samlingspartiet..1	Adjective (JJ)	Multiword proper noun (pmm)	Lexicalized form	0.29 (2)	0.00 (0)
ling	A2	Derivational suffix	böckling..nn.1	böckling..1	Noun (NN)	Noun (nn)	Derivation	0.29 (2)	0.00 (0)
ling	B1	Derivational suffix	änkling..nn.1			Noun (nn)			

Tables - des...

älskling

Definition :
- älskling (svenska.se)

Morphology :
- r: älsk s: ling

SAOL Analysis :
- älsk·ling

SO Analysis :
- älsk·ling·en

Lemgram ⇅
älskling..nn.1
lingon..nn.1
lingonsylt..nn.1
Moderata_sam

Antal träffar: 3

« ‹ **1** › »   Gå till sida [   ] av 1   Visa kontext

COCTAILL

Mona: God natt, **älskling**, och sov gott!

Lasse: Hej, **älskling**, nu är jag hemma!

Lasse: God morgon, **älskling**!

« ‹ **1** › »   Gå till sida [   ] av 1

Word formation mechanisms

Receptive
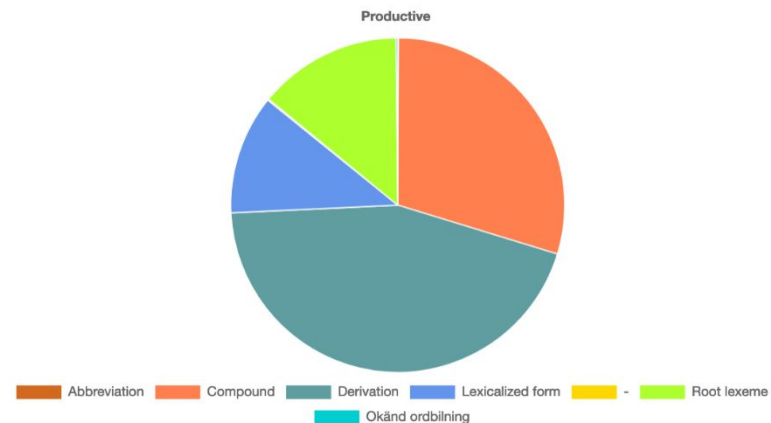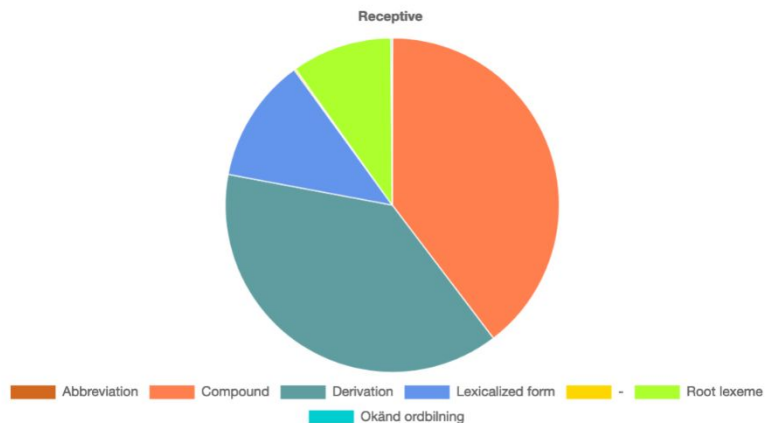
Productive

Word formation	Receptive
Abbreviation	17
Compound	14344
Derivation	13871
Lexicalized form	4367

Word formation	Productive
Abbreviation	4
Compound	1629
Derivation	2448
Lexicalized form	634

26

Morpheme categories
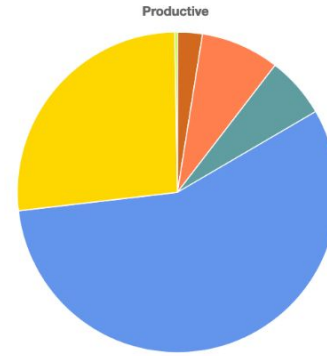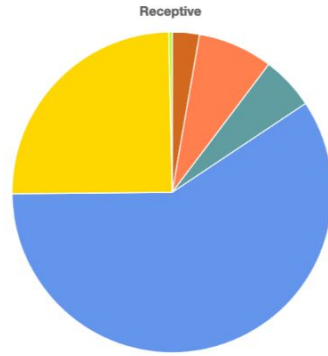
Receptive

Morpheme Category	Receptive
Binding morpheme	990
Inflectional ending	2739
Derivational prefix	1939
Root morpheme	21444

Word

Morpheme Category	Productive
Binding morpheme	138
Inflectional ending	436
Derivational prefix	338
Root morpheme	3108

Legend: Binding morpheme, Inflectional ending, Derivational prefix, Root morpheme, Derivational suffix, Unclassified / unknown

Statistics: type-token ratio

CEFR level	Receptive (Type)	Productive (Type)	Receptive (Token)	Productive (Token)	Receptive (TTR)	Productive (TTR)
A1	2166	464	25015	4083	0.09	0.11
A2	5134	714	16367	1866	0.31	0.38
B1	10327	1184	27224	2829	0.38	0.42
B2	10916	1310	22569	2748	0.48	0.48

Type-token ratio over all morphemes
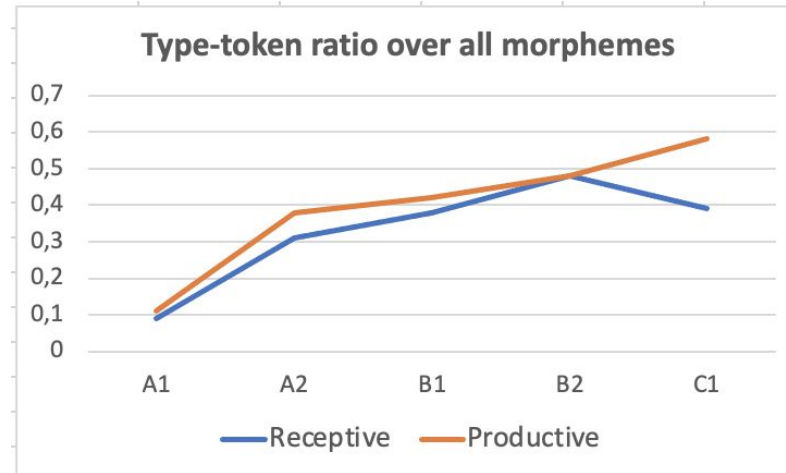
How did we do it?

Co ● De ● Roo ● Mor

Compounding ● **De**rivation ● **Roo**t ● **Mor**phology ...and more

- … is a dataset of 16 230 Swedish words analyzed manually for inner word formation structure

Lemgram view

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

"Morpheme family" view

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

Source data → item selection

COCTAILL corpus

Textbook texts for L2 Swedish
A1 (beginner) - C1 (advanced) levels

1 106 texts / 301 367 tokens

SVALex

Receptive vocabulary, lemgram-based
A1 (beginner) - C1 (advanced) levels

15 681 items, MWE included

(3 591 items are overlapping)

SweLL-pilot corpus

Essays written by L2 Swedish learners
A1 (beginner) - C1 (advanced) levels

339 essays / 144 087 tokens

SweLLex

Productive vocabulary, lemgram-based
A1 (beginner) - C1 (advanced) levels

6 965 items, MWE included

Sen*Lex

Joint vocabulary list
(receptive+productive)

based on sense distinctions

16 324 items
excl. problematic items
MWE included

Elena Volodina, Yousuf Ali Mohammed, and Therese Lindström Tiedemann. (2021). CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish. *NoDaLiDa 2021*: Linköping Electronic University Press, Vol.178. [pdf] [[video-presentation]]

Annotation categories: word formation

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')

Guidelines: https://tinyurl.com/fbe73ran

Annotation categories: morpheme types

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*	ked**j**ebrev
i	inflectional suffix	i_höst**as**
?	unknown	ir**on**i

Annotation tool: Legato

https://spraakbanken.gu.se/larkalabb/legato

Lexicographic Annotation Tool (LEGATO)

Guidelines | Skipped items [0] | Search | Filter | External links

Quick jump to:
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Å Ä Ö

Jump to: [____] Jump
1 - 16230

Current task: **MORPHOLOGY**

Progress: 11457/16230

SALDO lemgram	Saldo POS	Part-of-Speech	CEFR level
samvete..nn.1	noun (nn)	Noun (NN)	B1

Saldo sense: **samvete..1**
Saldo primary descriptor: **känsla..2**
Saldo secondary descriptor: **rätt..2**

Examples:
Jonas : Har du aldrig fått dåligt ** samvete ** ?
Kanske hade han trots allt dåligt ** samvete ** för att han hade lämnat oss och skaffat ny familj .
Birgitta erkänner att hon borde ha riktigt dåligt ** samvete ** för det här , men hon menar att det inte drabbar någon enskild .

Previous morphology1:	Previous morphology2:	SO analysis:	SAOL analysis:
p: sam r: vet s: e		sam=vet·et	sam=vete
Copy		Copy	Copy

Current values:

r: sam
r: vet
s: e

Select word form:

○ **ROOT LEXEME**
○ **DERIVATION**
● **COMPOUND**
○ **LEXICALIZED FORM**
○ **UNKNOWN**
○ **ABBREVIATION**

36

Annotation quality

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

(Krippendorff's alpha for Inter-Annotator Agreement)

Annotators: Beatrice Silén, Stellan Petersson, Maisa Lauriala
+Therese Lindström Tiedemann (researcher)

# What can we do with it?

Statistic analysis of L2 Swedish

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**

Relate statistics and levels of language development

Word formation mechanisms in COCTAILL data per level
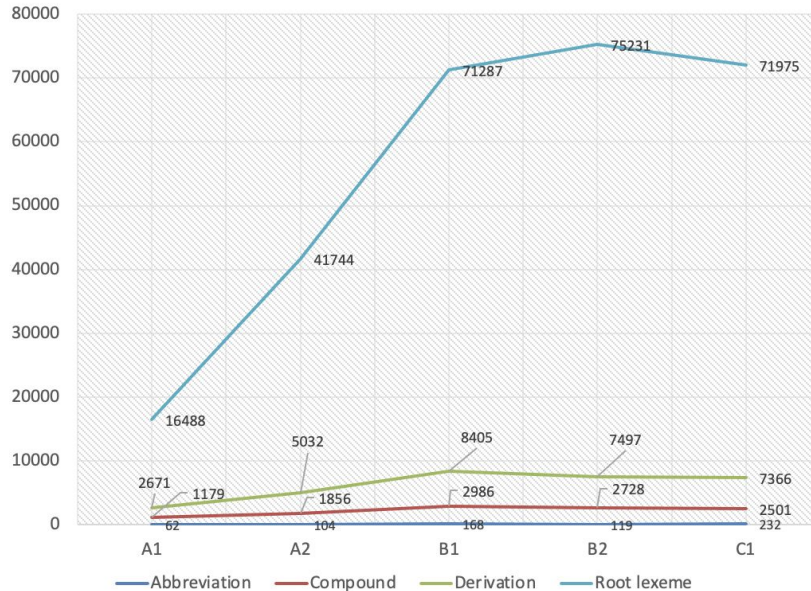
Word formation mechanisms in SweLL-pilot essay data per level
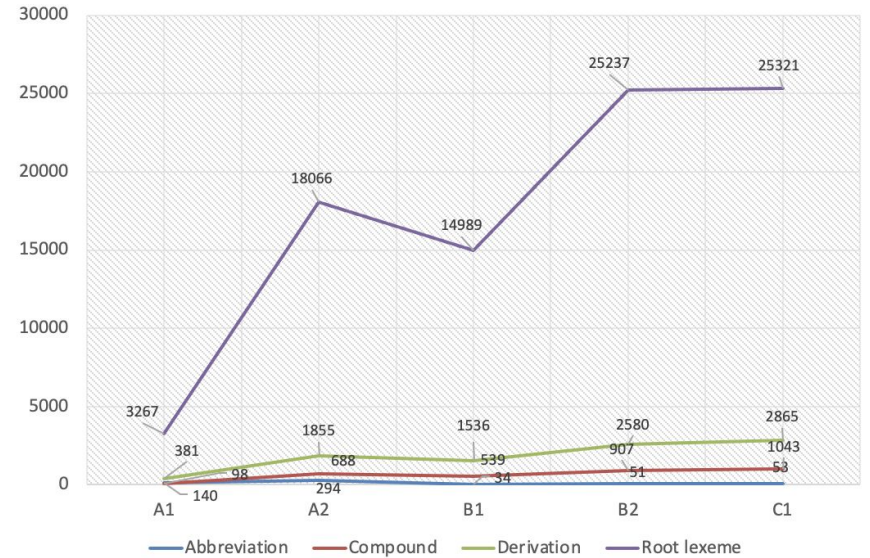
Derivational complexity

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

Morphological complexity

- MCI - morphological complexity index (cf. Brezina & Palotti, 2016)
  - http://corpora.lancs.ac.uk/vocab/analyse_morph.php
- Experiments with automatic proficiency prediction based on derivation morphology features
- ...

Automatic tools for

- ...detection of morpheme boundaries
- ...labeling morphemes for their types
- ...labeling words for their word formation type
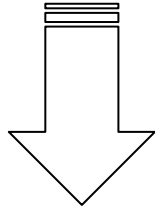
- ...with the ultimate aim/hope to add this type of analysis to automatic pipelines, primarily to Sparv

Swedish Word Family

Select root
bild

Search word

WordClass ▼

Saldo WordClass ▼

WordForm ▼

| A1 | A2 | B1 | B2 | C1 |

Table View | Graphical View | Statistical View | Download | Clear All

☑ **Only first occurance**

Total rows: 34

Root ⇅	CEFR ⇅	Lemgram ⇅	Sense ⇅	WordClass ⇅	Saldo WordClass ⇅	WordForm ⇅	Receptive ⇅	Productive ⇅
bild	A1	bild..nn.1	bild..1	Noun (NN)	noun (nn)	Root lexeme	13	5
bild	A2	universitetsutbildning..nn.1	universitetsutbildning..1	Noun (NN)	noun (nn)	Compound	1	0
bild	A2	utbilda..vb.1	utbilda..1	Verb (VB)	verb (vb)	Compound	3	0
bild	A2	utbildning..nn.1	utbildning..1	Noun (NN)	noun (nn)	Derivation	6	0
bild	A2	utbildningsnivå..nn.1	utbildningsnivå..1	Noun (NN)	noun (nn)	Compound	1	0
bild	A2	bilda..vb.1	bilda..1	Verb (VB)	verb (vb)	Derivation	2	0

44

Thank you!
Comments? Questions?

(or mailto: elena dot volodina at svenska dot gu dot se)