GÖTEBORGS UNIVERSITET

SPRÅKBANKENTEXT

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF HELSINKI

Swedish morphological profile

Elena Volodina, Språkbanken Text, Sweden (presenter)

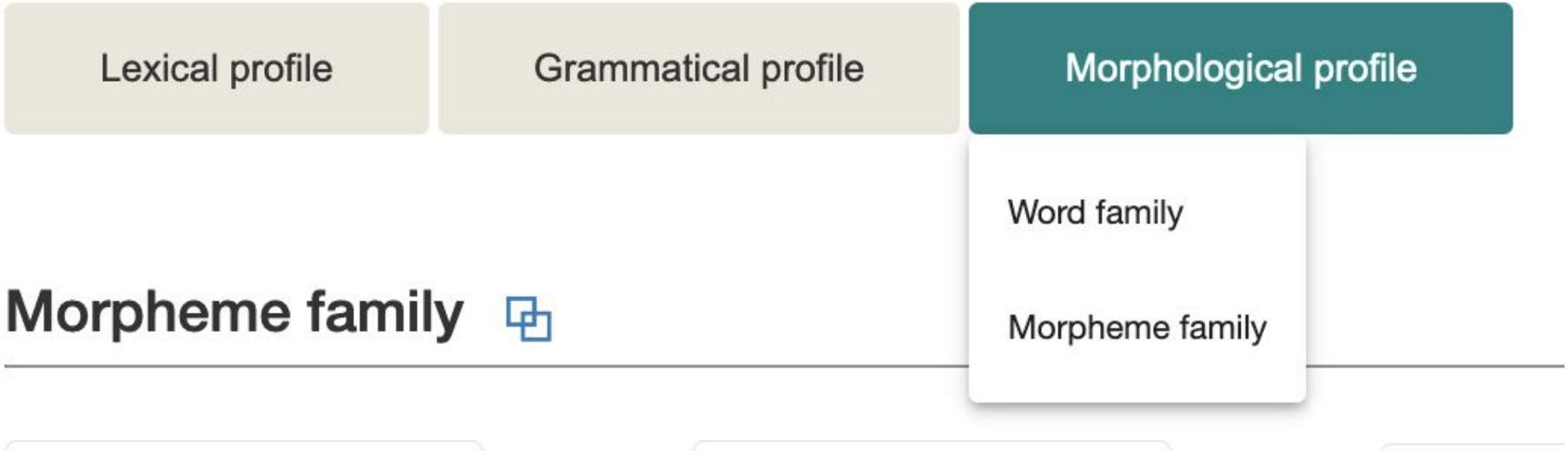
Based on work by: Elena Volodina, Yousuf Ali Mohammed, Therese Lindström Tiedemann and annotators

Demo

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

Login: demo

Swedish L2 Morphological profile



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

Morpheme family

Search morpheme: ling

Search word

Morpheme Category
- [ ] Binding morpheme
- [ ] Derivational prefix
- [x] Derivational suffix
- [ ] Inflectional ending
- [ ] Root morpheme
- [ ] Unclassified / unknown

Word Class | Saldo Word Class | Word formation

A1 | A2 | B1 | B2 | C1 | C2

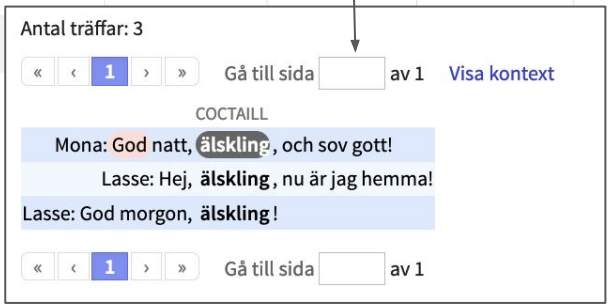
Tables | Graphs | Statistics | Download | Clear all

- [ ] Add column for related roots
- [x] Show only first occurrence

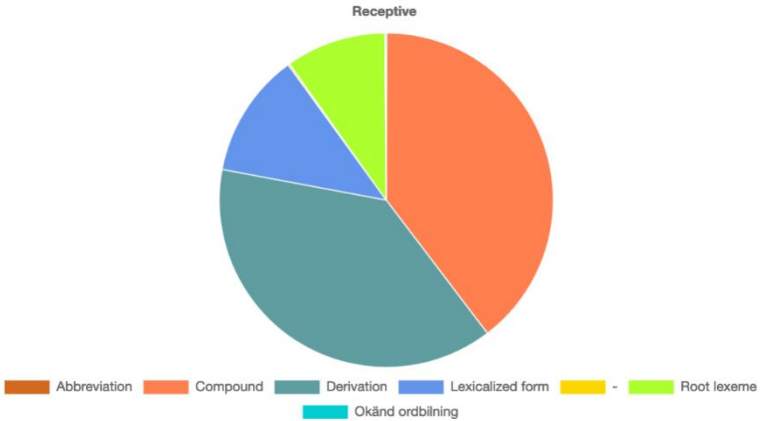
Tables - desc

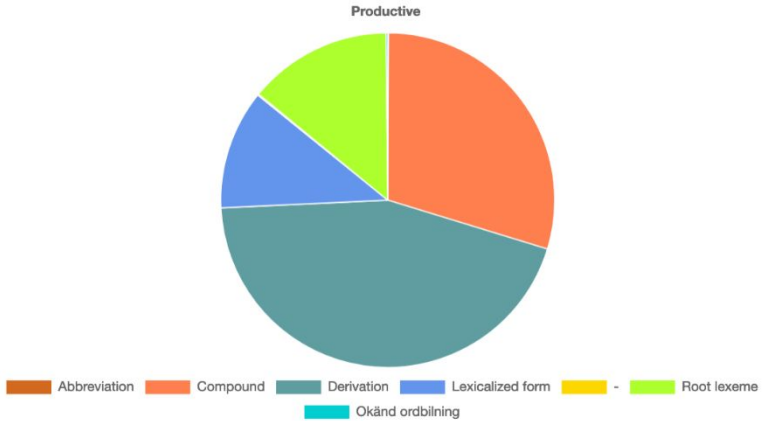
Total rows : 12

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)			



Word formation mechanisms

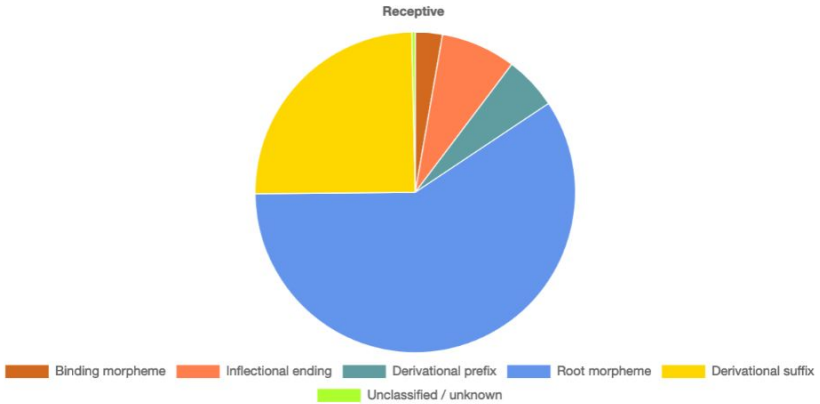


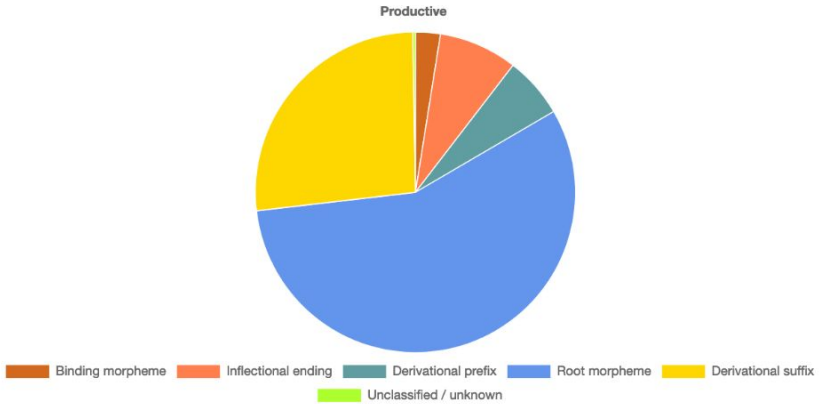


Word formation	Receptive
Abbreviation	17
Compound	14344
Derivation	13871
Lexicalized form	4367

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

Morpheme categories





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

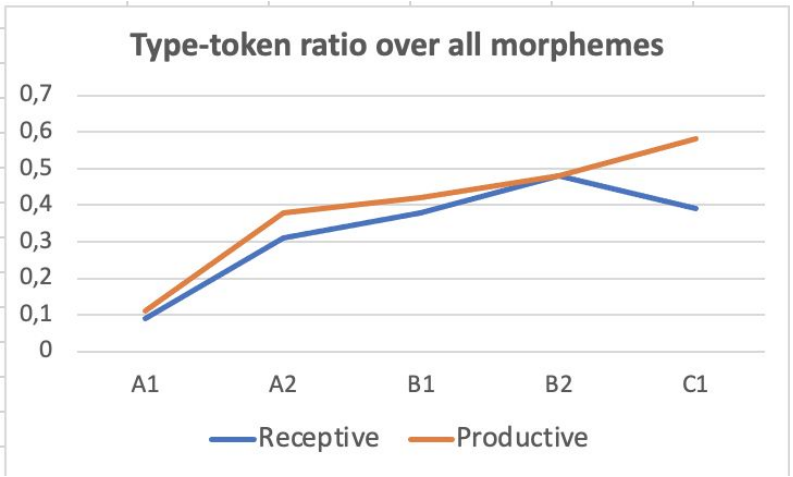
Morpheme Category	Productive
Binding morpheme	138
Inflectional ending	436
Derivational prefix	338
Root morpheme	3108

Word

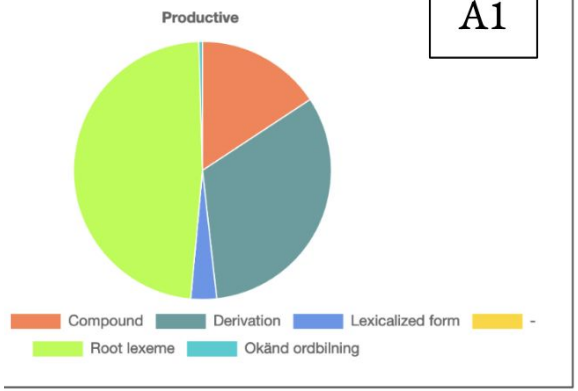
Statistics: type-token ratio

CEFR level

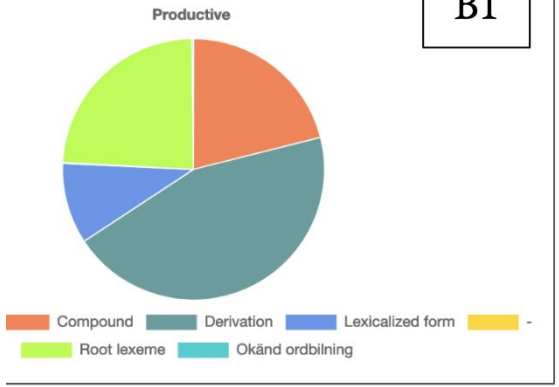
CEFR level	Receptive (Type)	Productive (Type)	Receptive (Token)	Productive (Token)	Receptive (TTR)	Productive (TTR)
A1	2166	464	25015	4083	0.09	0.11
A2	5134	714	16367	1866	0.31	0.38
B1	10327	1184	27224	2829	0.38	0.42
B2	10916	1310	22569	2748	0.48	0.48



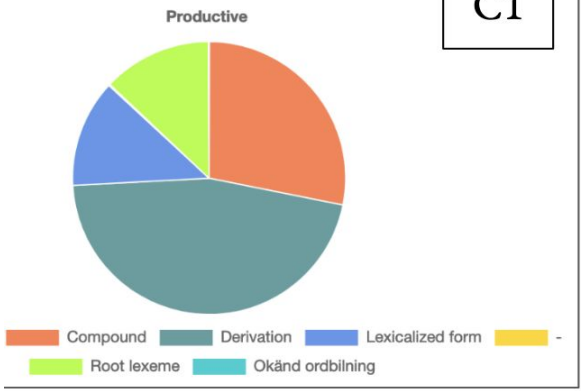
A1


Productive
Compound
Derivation
Lexicalized form
-
Root lexeme
Okänd ordbilning

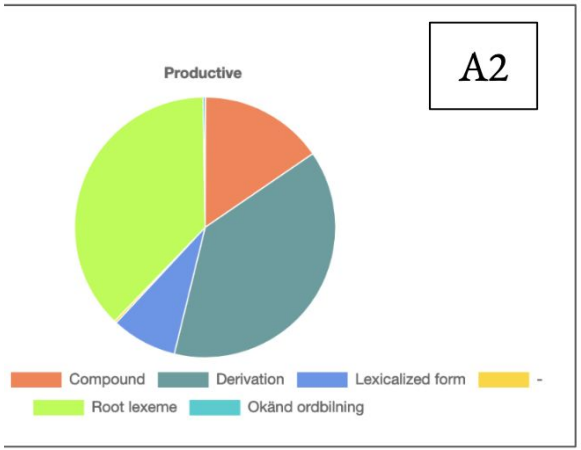
B1


Productive
Compound
Derivation
Lexicalized form
-
Root lexeme
Okänd ordbilning

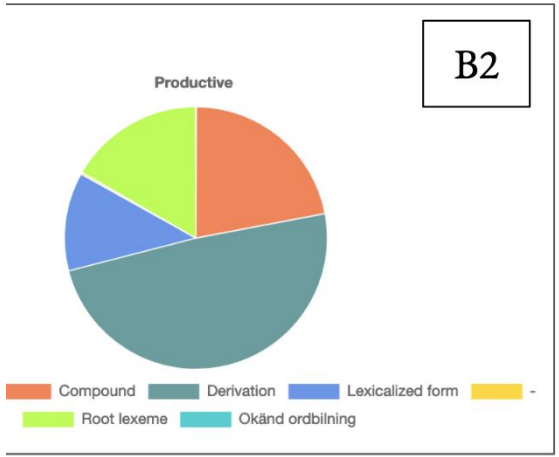
C1

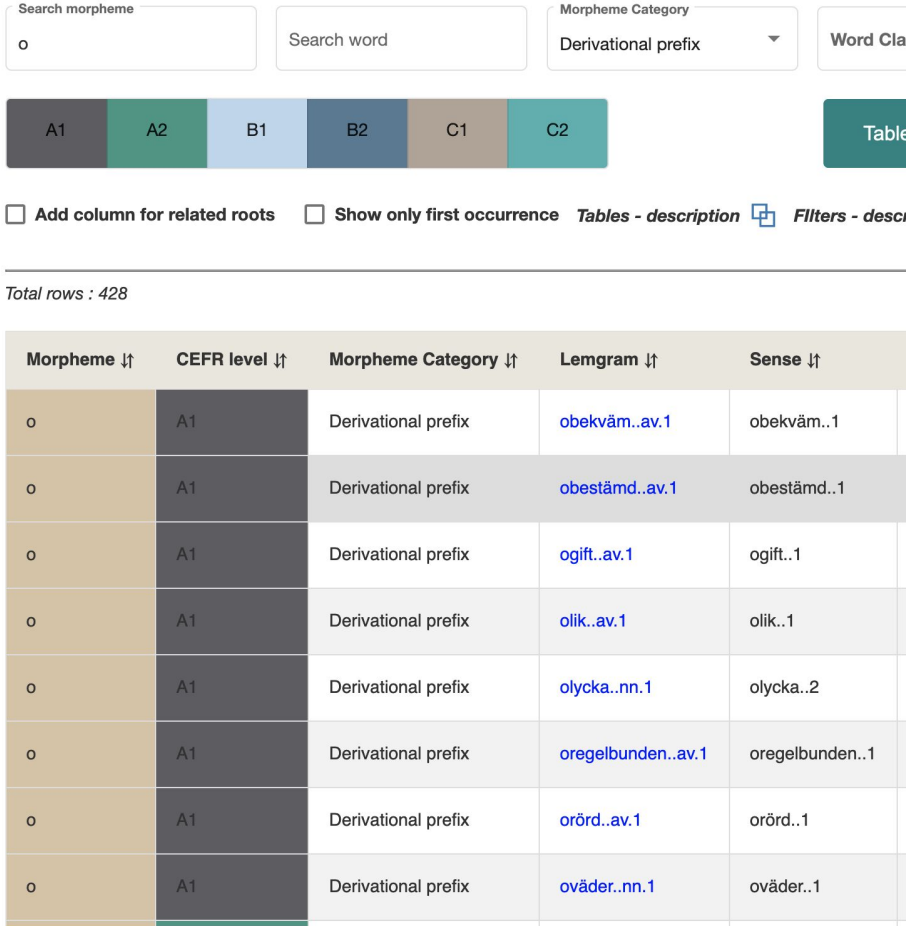

Productive
Compound
Derivation
Lexicalized form
-
Root lexeme
Okänd ordbilning

A2

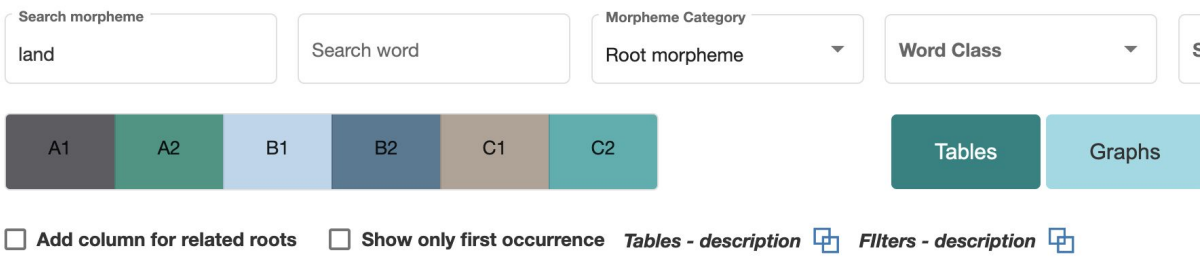

Productive
Compound
Derivation
Lexicalized form
-
Root lexeme
Okänd ordbilning

B2


Productive
Compound
Derivation
Lexicalized form
-
Root lexeme
Okänd ordbilning

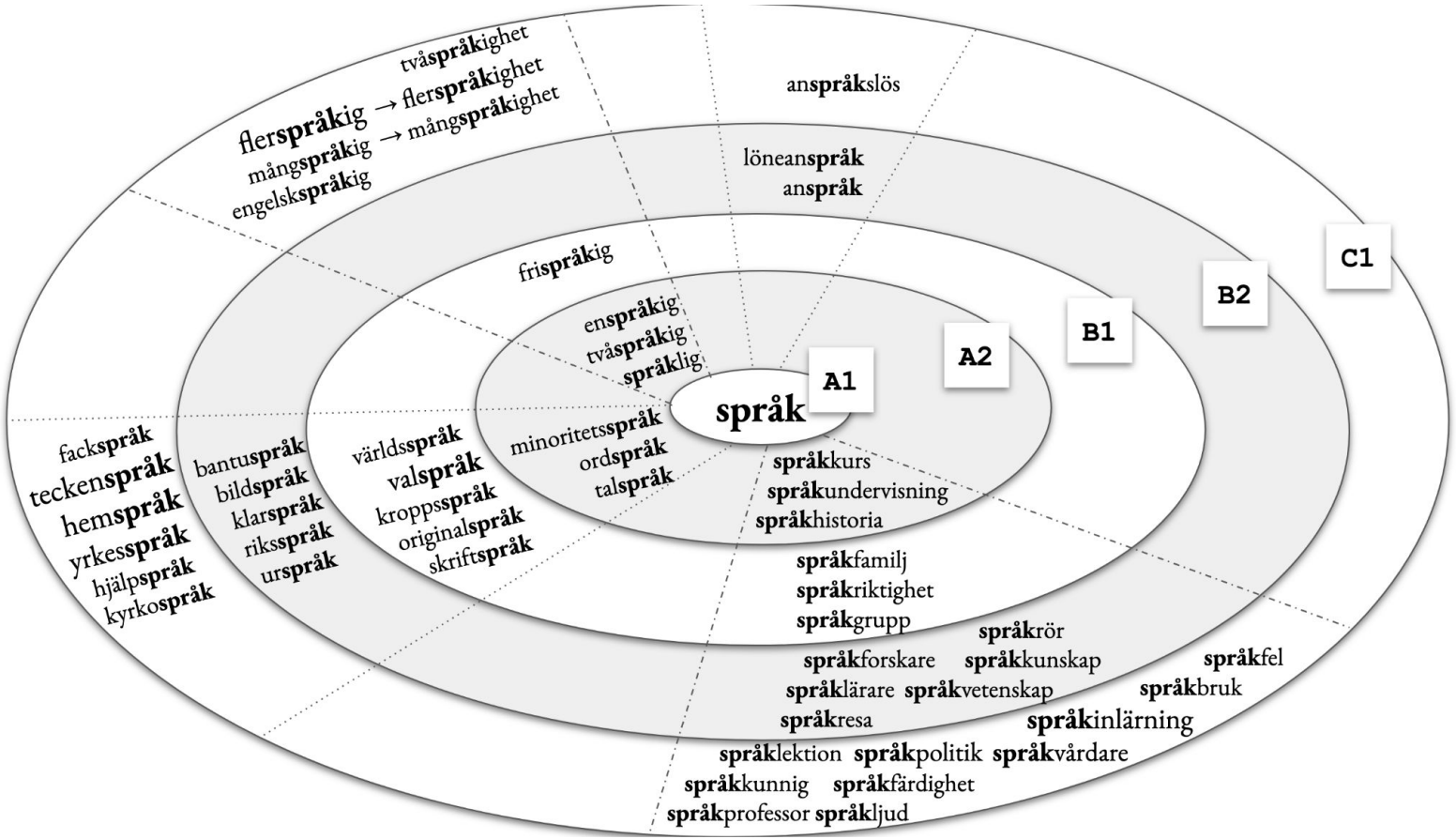

Search morpheme
o
Search word
Morpheme Category
Derivational prefix
Word Cla
A1
A2
B1
B2
C1
C2
Table
Add column for related roots
Show only first occurrence
Tables - description
Filters - desc
Total rows : 428
Morpheme
CEFR level
Morpheme Category
Lemgram
Sense
o A1 Derivational prefix obekväm..av.1 obekväm..1
o A1 Derivational prefix obestämd..av.1 obestämd..1
o A1 Derivational prefix ogift..av.1 ogift..1
o A1 Derivational prefix olik..av.1 olik..1
o A1 Derivational prefix olycka..nn.1 olycka..2
o A1 Derivational prefix oregelbunden..av.1 oregelbunden..1
o A1 Derivational prefix orörd..av.1 orörd..1
o A1 Derivational prefix oväder..nn.1 oväder..1

Root morpheme	CEFR level	Lemgram	Sense	Word Class	Saldo Word Class	Word formation	Receptive	Productive
jul	A1	jul..nn.1	jul..1	Noun (NN)	Noun (nn)	Root lexeme	0.39 (1)	0.00 (0)
jul	A2	julbord..nn.1	julbord..1	Noun (NN)	Noun (nn)	Compound	0.15 (1)	0.00 (0)
jul	A2	juldagsmorgon..nn.1	juldagsmorgon..1	Noun (NN)	Noun (nn)	Compound	0.15 (1)	0.00 (0)
jul	A2	julklapp..nn.1	julklapp..1	Noun (NN)	Noun (nn)	Compound	0.29 (2)	0.00 (0)
jul	A2	jullov..nn.1	jullov..1	Noun (NN)	Noun (nn)	Compound	0.87 (6)	0.44 (1)
jul	A2	julotta..nn.1	julotta..1	Noun (NN)	Noun (nn)	Compound	0.15 (1)	0.00 (0)
jul	A2	jultomte..nn.1	jultomte..1	Noun (NN)	Noun (nn)	Compound	0.15 (1)	0.00 (0)
jul	B1	julfest..nn.1	julfest..1	Noun (NN)	Noun (nn)	Compound	0.08 (1)	0.00 (0)
jul	B2	julafton..nn.1	julafton..1	Noun (NN)	Noun (nn)	Compound	1.06 (16)	0.00 (0)
jul	B2	julbock..nn.1	julbock..1	Noun (NN)	Noun (nn)	Compound	0.13 (2)	0.00 (0)
jul	B2	juldag..nn.1	juldag..1	Noun (NN)	Noun (nn)	Compound	0.26 (4)	0.00 (0)
jul	B2	julfirande..nn.1	julfirande..1	Noun (NN)	Noun (nn)	Derivation	0.13 (2)	0.00 (0)
jul	B2	julgåva..nn.1	julgåva..1	Noun (NN)	Noun (nn)	Compound	0.07 (1)	0.00 (0)
jul	B2	julgran..nn.1	julgran..1	Noun (NN)	Noun (nn)	Compound	0.33 (5)	0.00 (0)



Total rows : 202

Morpheme	CEFR level	Morpheme Category	Lemgram	Sense	Word Class
land	A1	Root morpheme	Ångermanland..pm.1	Ångermanland..1	Proper noun (PM)
land	A1	Root morpheme	Jämtland..pm.1	Jämtland..1	Proper noun (PM)
land	A1	Root morpheme	Svealand..pm.1	Svealand..1	Proper noun (PM)
land	A1	Root morpheme	Thailand..pm.1	Thailand..1	Proper noun (PM)
land	A1	Root morpheme	Tyskland..pm.1	Tyskland..1	Proper noun (PM)
land	A1	Root morpheme	Uppland..pm.1	Uppland..1	Proper noun (PM)
land	A1	Root morpheme	Värmland..pm.1	Värmland..1	Proper noun (PM)
land	A1	Root morpheme	land..nn.1	land..1	Noun (NN)


flerspråkig
tvåspråkighet → flerspråkighet
mångspråkig → mångspråkighet
engelskspråkig
anspråkslös
löneanspråk
anspråk
enspråkig
tvåspråkig
språklig
frispråkig
A1
A2
B1
B2
C1
språk
minoritetsspråk
ordspråk
talspråk
världsspråk
valspråk
kroppsspråk
originalspråk
skriftspråk
bantuspråk
bildspråk
klarspråk
riksspråk
urspråk
fackspråk
teckenspråk
hemspråk
yrkesspråk
hjälpspråk
kyrkospråk
språkkurs
språkundervisning
språkhistoria
språkfamilj
språkriktighet
språkgrupp
språkrör
språkforskare
språkkunskap
språkfel
språklärare
språkvetenskap
språkbruk
språkresa
språkinlärning
språklektion
språkpolitik
språkvårdare
språkkunnig
språkfärdighet
språkprofessor
språkljud


eftermiddag
daghem
söndag
lördag
förmiddag
tisdag
onsdag
dag
dagbok
fredag
dagis
lördagsmorgon
torsdag
veckodag
vardagsrum
middag
måndag
midsommardag
dagstidning
riksdagsledamot
påskdag
pingstdag
torsdagkväll
våffeldag
allhelgonadag
mellandag
riksdag
söndagkväll
vardaglig
vardag
vardagsmat
tjugondag
juldag
födelsedag
bröllopsdag
födelsedagsbarn
daglig
helgdag
nyårsdag
riksdagsval
juldagsmorgon
vårdag
namnsdag
himmelsfärdsdag
riksdagsman
nationaldag
morgondag
arbetsdag
middagsgäst
skoldag
riksdagsparti
eftermiddagskaffe
dagligen
olycksdag
fettisdag
vardagsliv
dagboksanteckning
gårdag
vardagsmänniska
fredagskväll
födelsedagspresent
söndagspappa
fredagsmys
klämdag
sjukdag
måndagskväll
karensdag
härområdagen
söndagsmorgon
matdags
vardaglig
söndagsbilaga
askonsdag
söndagseftermiddag
höstdag
middagsmat
annandag
torsdagsmorgon
dagcentral
skärtorsdag
Svenska Dagbladet
trerättersmiddag
dagspress
dagsläge
dagblad
fettisdagsbulle
födelsedagsfest
dagordning
middagsbord
högtidsdag
semesterdag
festmiddag
långfredag
Dagens nyheter
lördagskväll
föräldradag
vinterdag
gammaldags
middagsrast
dödsdag
middagstimme
A1 (17 members) →
A2 (36 members) →
B1 (28 members) →
B2 (13 members) →
C1 (7 members)

Some definitions

Lexical units as the main entry

- word forms → running word forms (`walked`)
- lemma → base form (`walk`)
- lemgram → base form + POS (`walk, noun`)
- flemma → homonymous forms (`walk, noun & walk, verb`)
- **word family** → word grouped around the same root (`walk, walk-through`)
- **morpheme family** →

Presuppose different assumptions on vocabulary acquisition. Some discussion see in

Brown, Dale. 2018. Examining the word family through word lists. *Vocabulary Learning and Instruction* 7, no. 1: 51-65.

Stoeckel, Tim, Tomoko Ishii, and Phil Bennett. 2020. Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics* 41, no. 4 (2020): 601-606.

- 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]

Suffix -skap (*-skap* family)

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)

- 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]

word forms

lemgram (inflectional paradigms)

word formation (affixation)

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	

Coxhead (1998)

Academic Word List

The following is a list of the words from the Academic Word List that are highlighted in this dictionary. Words shown in **bold** are one of the 'parent words'.

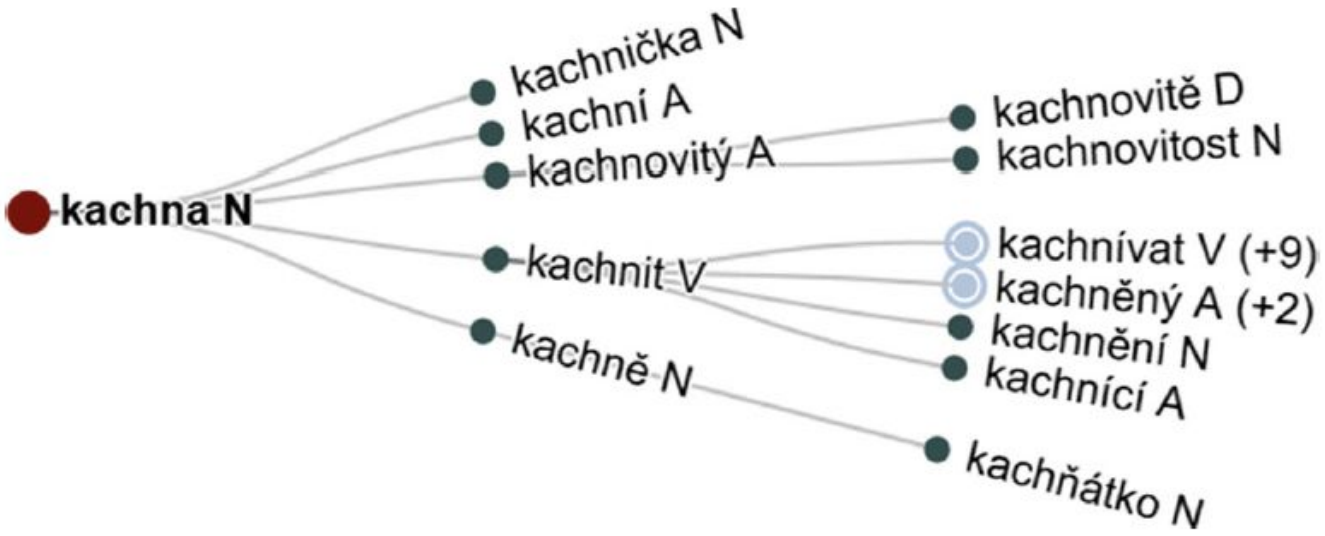
abandon[1] *v*
abandoned *adj*
abnormal *adj*
abstract[1] *adj*
abstract[2] *n*
abstract[3] *v*
abstraction *n*
academic[1] *adj*
academic[2] *n*
academy *n*
access[1] *n*

adjacent *adj*
adjust *v*
adjustment *n*
administration *n*
administrative *adj*
adult[1] *n*
adult[2] *adj*
advocacy *n*
advocate[1] *v*
advocate[2] n
affect *v*

analogy *n*
analyse *v*
analysis *n*
analyst *n*
analytical *adj*
analyze *v*
annual[1] *adj*
anticipate *v*
anticipation *n*
apparent *adj*
append *v*

assist[1] *v*
assistance *n*
assume *v*
assuming *conj*
assumption *n*
assurance *n*
assure *v*
attach *v*
attachment *n*
attain *v*
attainment *n*

Word formation networks: Polish, Czech, Spanish, French



Lango, M., Žabokrtský, Z., & Ševčíková, M. (2021). Semi-automatic construction of word-formation networks. *Language Resources and Evaluation*, *55*(1), 3-32. https://link.springer.com/article/10.1007/s10579-019-09484-2

Our definition

Word family is a group of words sharing

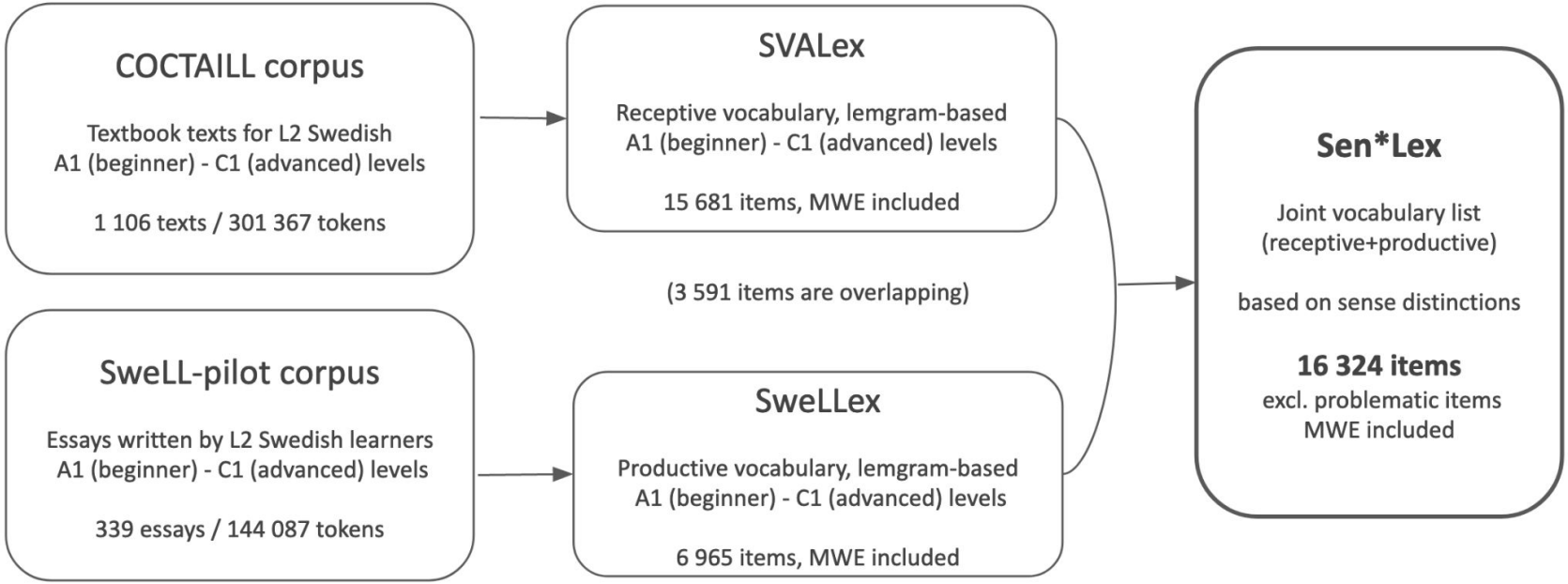
(1) the same root/base, including potential variations of that (e.g. `go, bygone`), compounds (e.g. `light, light-hearted`), derived forms (e.g. `teach, teacher`) and multi-word expressions (e.g. `day off, day by day`)

(2) the same meaning (**X** `sunlight, light-hearted`)

How did we do it?

CoDeRooMor dataset:

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



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')

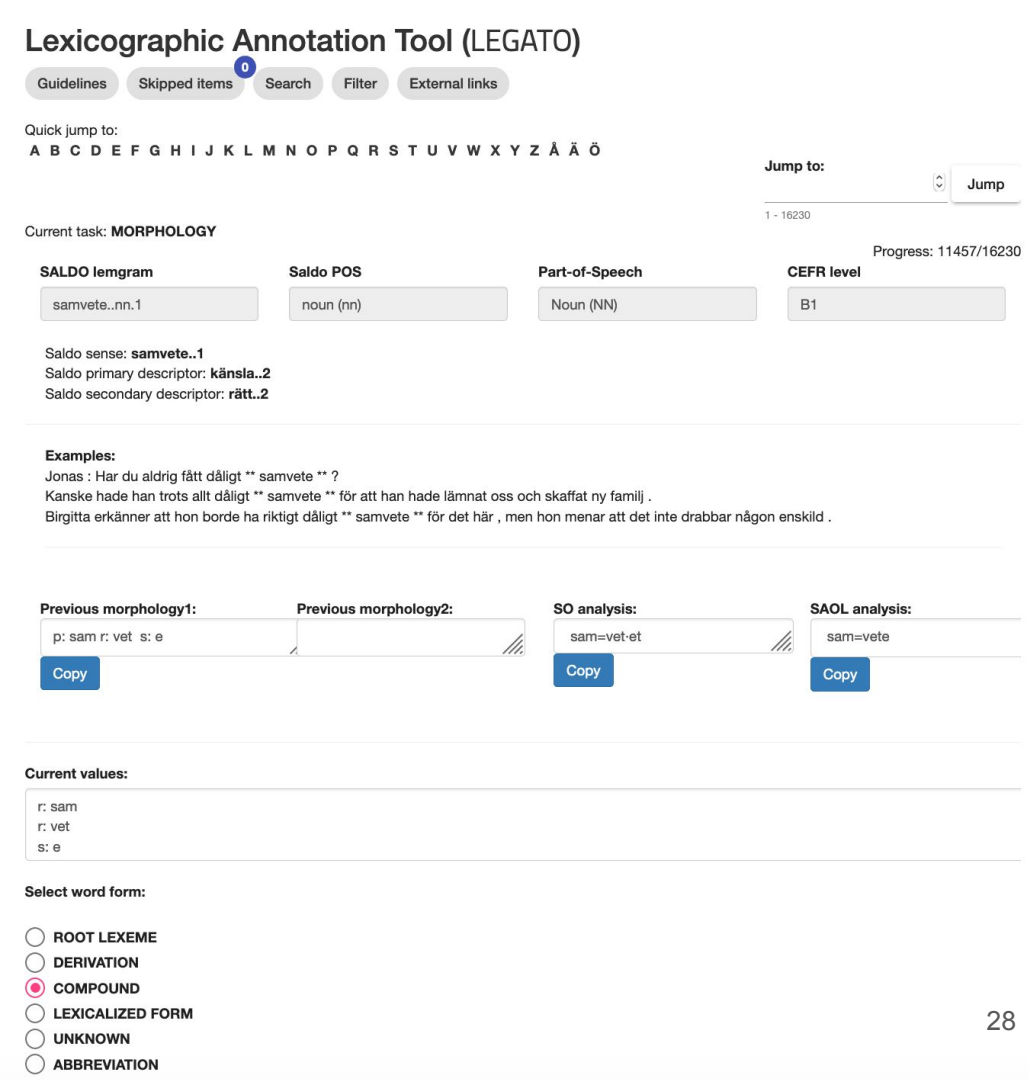
Guidelines: https://tinyurl.com/fbe73ran

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*	kedj**e**brev
i	inflectional suffix	i_höst**as**
?	unknown	ir**on**i

Annotation tool: Legato

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



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)

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

freely available from https://spraakbanken.gu.se/en/resources/coderoomor

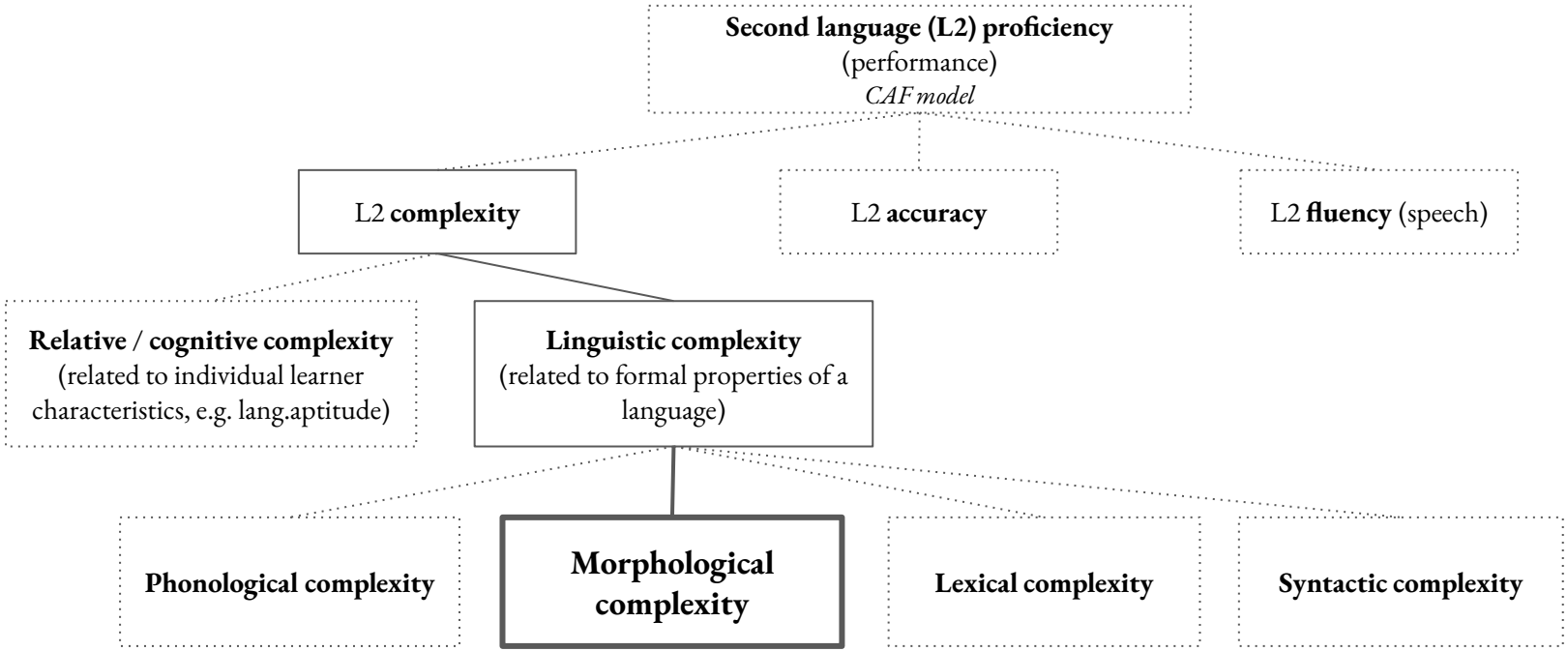
A few facts

- 4 429 word families
- smallest: 1 member (e.g. **adren**alin)
- biggest: 313 members (e.g. **ut**bildning)
- most numerous: function words consisting of root lexelmes (e.g. preposition ut)

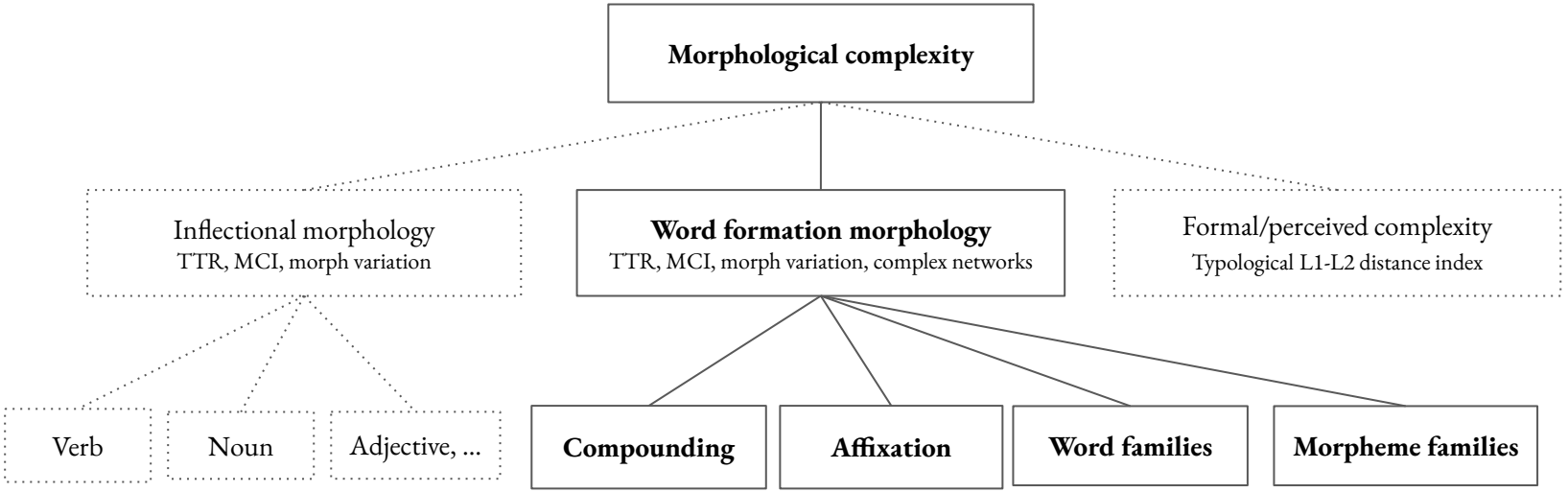
Total rows : 22553 Unique roots/level: A1: 849, A2: 1527, B1: 2168, B2: 2355, C1: 1870,

Why did we do it?

Complexity as a proxy of learner language development



Our focus


Morphological complexity
Inflectional morphology
TTR, MCI, morph variation
Word formation morphology
TTR, MCI, morph variation, complex networks
Formal/perceived complexity
Typological L1-L2 distance index
Verb
Noun
Adjective, ...
Compounding
Affixation
Word families
Morpheme families

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 & Pallotti, 2016):
  - http://corpora.lancs.ac.uk/vocab/process_text_morph.php

Complexity measures (deriv. morphology)

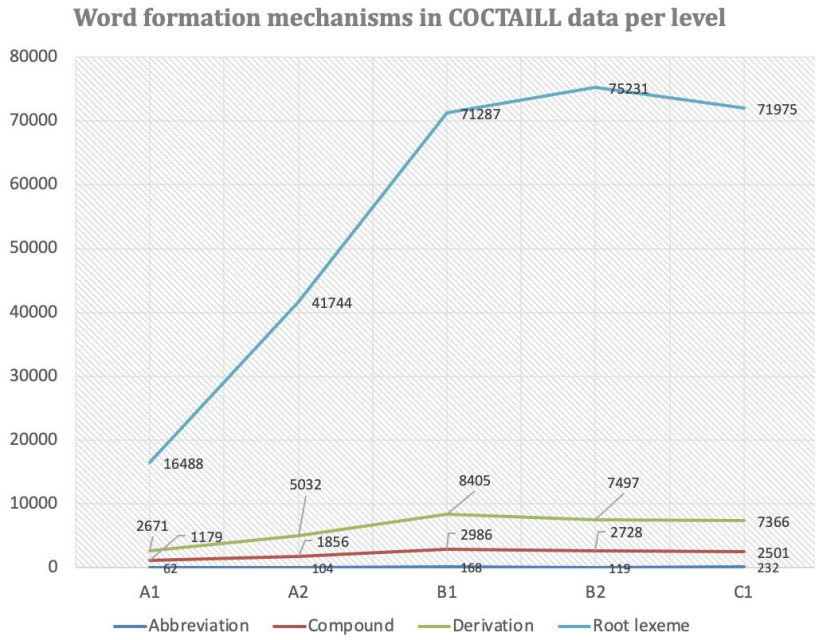
- ...yet to be developed

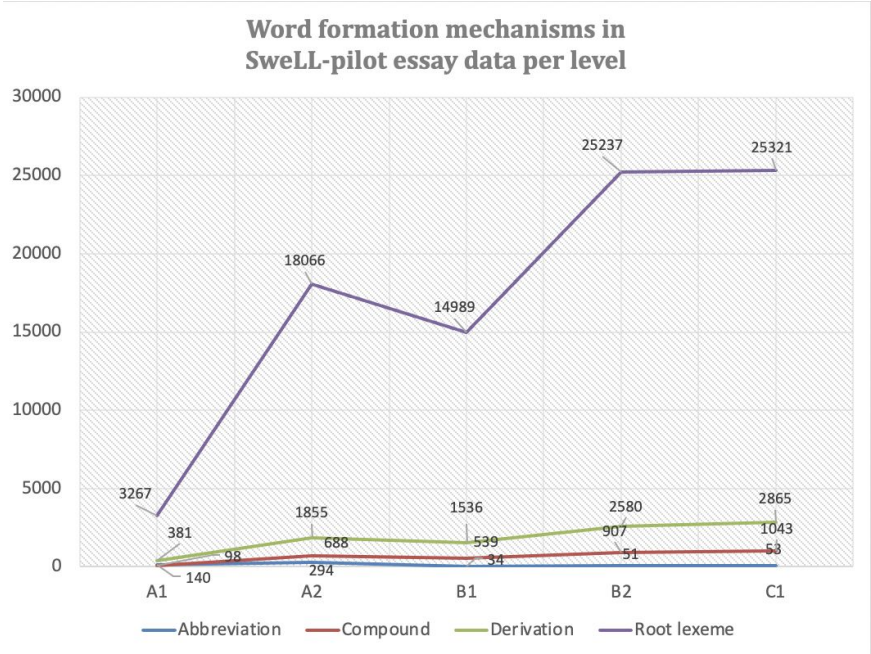
What can we do with it?

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





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

Within L2 context

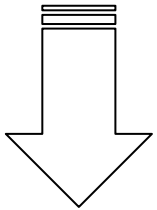
- Relation of morphological awareness and language proficiency
- Awareness of derivational morphemes
- Priming effect of roots/affixes on learning of new vocabulary (cf word families)
- Vocabulary testing
- Pedagogical applications / ICALL (e.g. exercises)
- ...

Outside L2 context

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

Automatic tools for

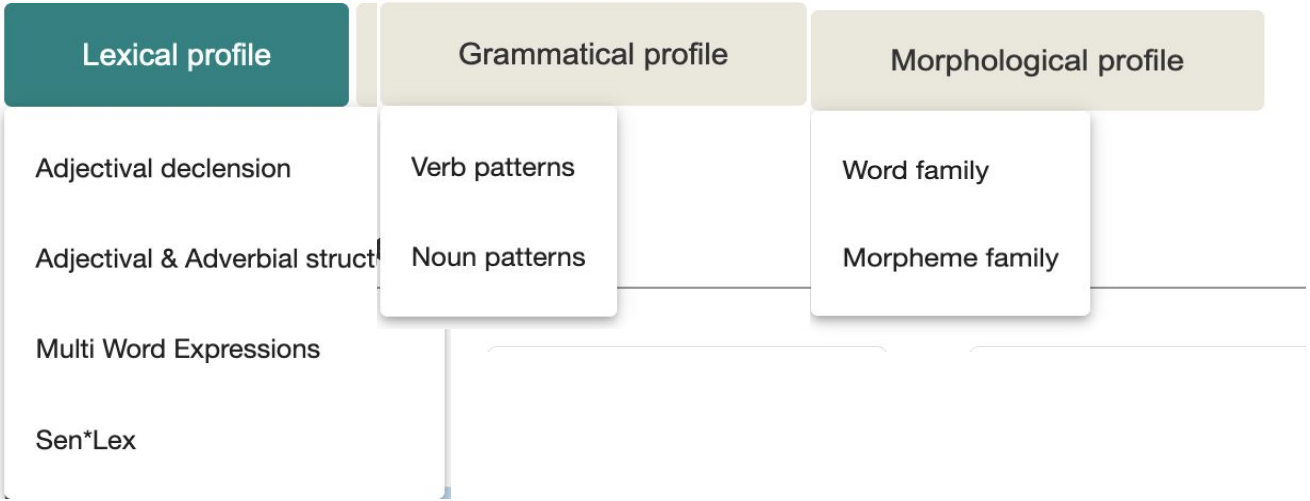
- ...detection of morpheme boundaries
- ...labeling morphemes for their types
- ...labeling words for their word formation type



- ...with the ultimate aim/hope to add this type of analysis to automatic pipelines, primarily to Sparv

Our major principles

- research oriented, in addition to teacher- & learner focus
- descriptive, transparent, empirically based, provides access to all corpus hits
- splits receptive and productive knowledge
- statistic analysis and graphical representation



GÖTEBORGS UNIVERSITET

SPRÅKBANKENTEXT

RIKSBANKENS JUBILEUMSFOND
STIFTELSEN FÖR HUMANISTISK OCH SAMHÄLLSVETENSKAPLIG FORSKNING

UNIVERSITY OF HELSINKI

Thank you!

Comments? Questions?

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