

# Swedish Lexical Profile

David Alfter

Gothenburg Research Infrastructure in Digital Humanities (GRIDH)

University of Gothenburg, Sweden



# Overview

- Single words and proficiency levels
- Multi-word expressions and proficiency levels
- Applications and demonstration



# Introduction

- “while without grammar very little can be conveyed, without vocabulary nothing can be conveyed” (Wilkins 1972: pp.111-112)
- Text > Sentence > Word



# Interest in the topic

- Shared tasks on complex word identification (2016, 2018)
- Shared task on lexical complexity prediction (2021)



# Introduction

- BA, MA in Computational linguistics
- PhD in Language Technology
  
- Exploring natural language processing for single-word and multi-word **lexical complexity from a second language learner** perspective



# Introduction

- Objective of the presentation
  - Presentation of our methodology
- Key points
  - Linking words and expressions to CEFR levels
  - Resources



# Previous work

- Manually created lists
  - Nusvensk frekvensordbok baserad på tidningstext (Allén 1971)
  - Tiotusen i top (Allén 1972)
  - Frekvensordbok över svenska elevtexter (Larsson, Rosén and Anderson 1985)
  - Talspråksfrekvenser (Allwood 1999)
  - Base Vocabulary Pool (Forsbom 2006)
  - Akademisk ordlista (Sköldberg and Johansson Kokkinakis 2012)



# Previous work

- Manually created lists
  - Paper versions
  - Copyright issues
  - Outdated





# Previous work

- Automatically created lists
  - KELLY list (Volodina and Johansson Kokkinakis, 2012)
  - SVALex (François et al., 2016)
  - SweLlex (Volodina et al., 2016)
  - NyLlex (Holmer and Rennes, 2022)



# Previous work

- Automatically created lists
  - Validity
  - Link to CEFR?



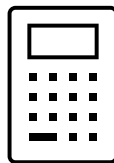
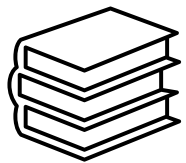
# Scope

- Focus on Swedish
- Focus on two resources:
  - SVALex as receptive vocabulary list
  - SweLLex as productive vocabulary list



# Scope

- CEFRLex (<https://cental.uclouvain.be/cefrlex/>)
  - Textbook-derived lists
  - Student essay derived list for Swedish
  - 6 languages



Lemma	POS-tag	A1	A2	B1	B2	C1	Total
bil	NN_UTR	430.2138	1234.2078	728.9847	422.283	363.5446	618.8567
överge	VB	0	0	7.3203	24.5182	39.6516	17.2695
rättvisa	NN_UTR	0	0	3.6601	25.6189	26.4344	13.6602
kilo	NN_NEU	0	302.0833	145.1229	65.0611	13.2172	89.8907
resa	VB	166.3009	375.2582	450.3526	298.4905	330.4297	356.362
låg	JJ	0	49.315	125.922	217.3103	252.1311	156.126
så klart	ABM_MWE	0	16.2635	81.6019	45.5033	13.2172	38.1738
till skillnad från	PPM_MWE	0	0	5.3395	2.409	3.6699	5.1839



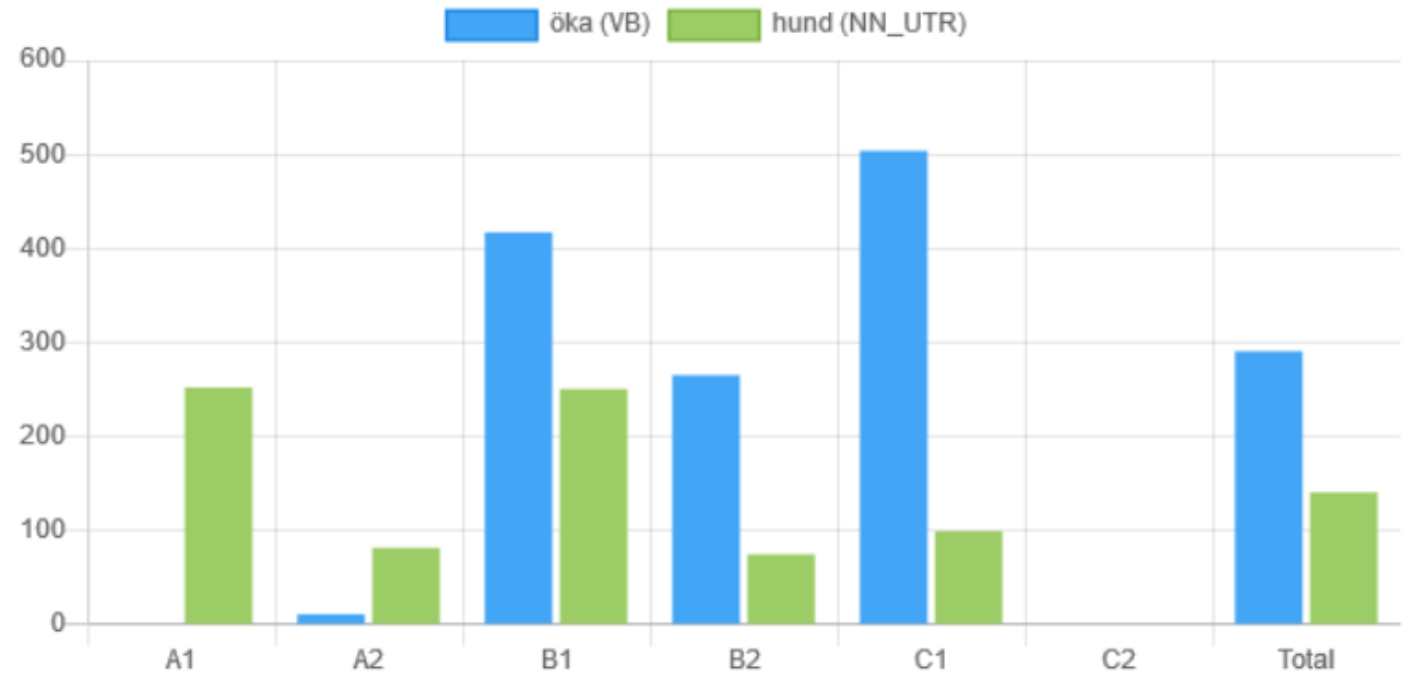
# Scope

- Based on authentic data
- Data-driven approach
- Not only frequency



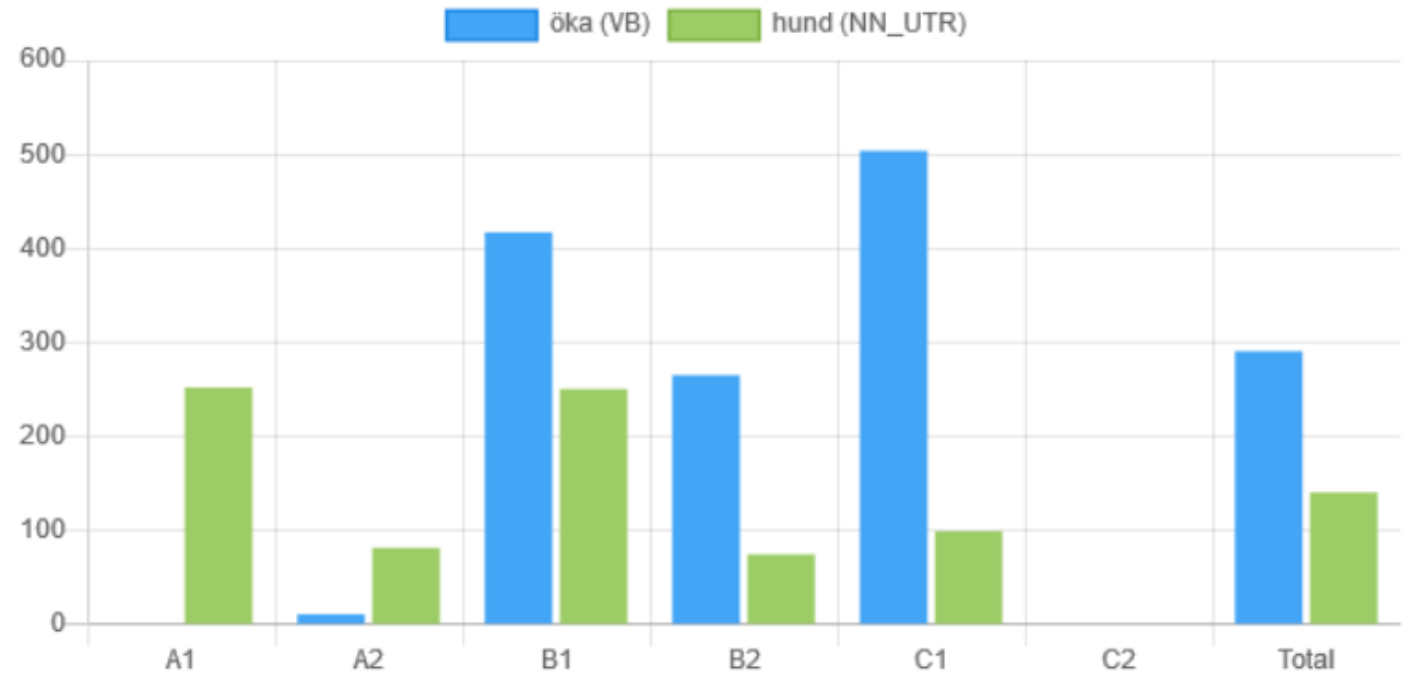
# Scope

- Based on authentic data
- Data-driven approach
- **Not only frequency**



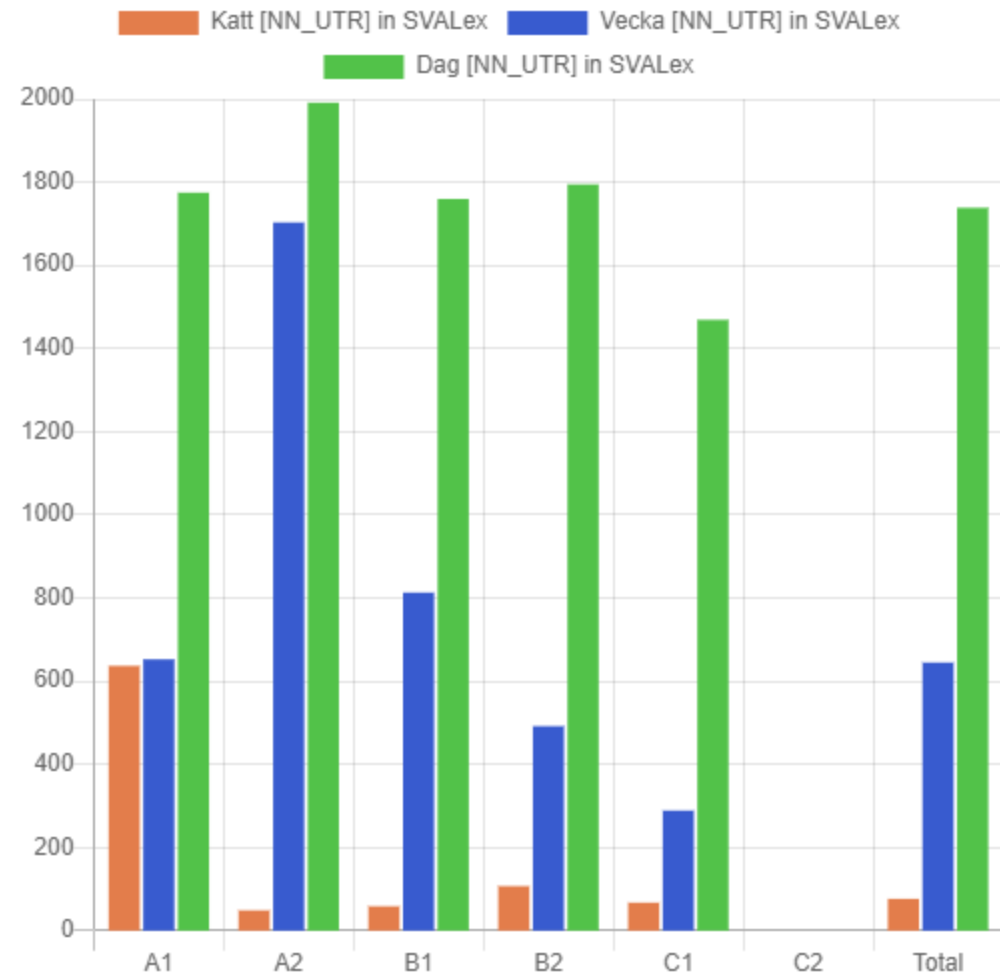
# Scope

- Based on authentic data
- Data-driven approach
- **Not only frequency**
- **Adapted to learners**



# Limitations

- No CEFR level





# Limitations

- No CEFR level
- No sense distinction

## land HOMONYM

Homonymen **land** har **4 definitioner** inlagda i ordboken.

1. fastlandet (ej till sjöss)
2. nation/stat eller rike
3. område utanför tätort (landsbygden)
4. stycke land som är till för odlingar

Mainland

Nation/state or kingdom

Area outside of urban areas

piece of land that is used for cultivation

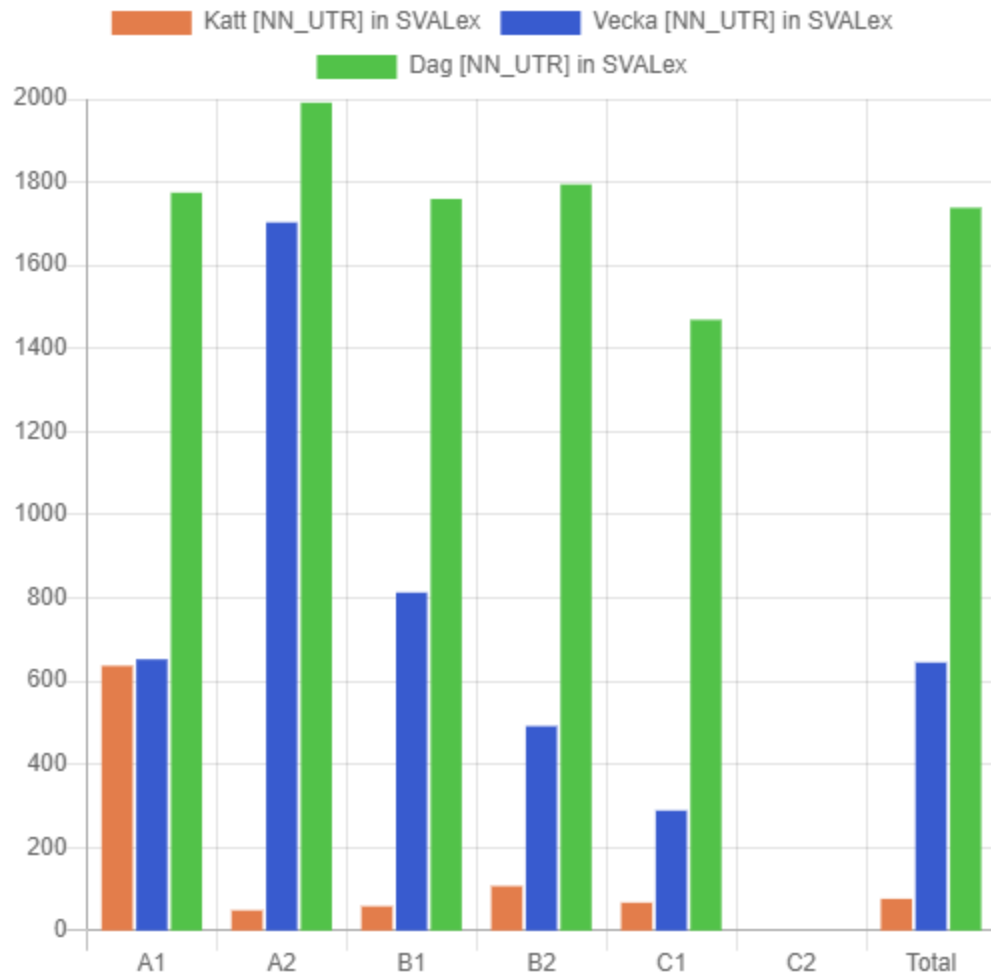


# Methodology

- Transform distributions into labels
- Train machine learning algorithm
- Word sense disambiguation



# Mapping distributions to labels

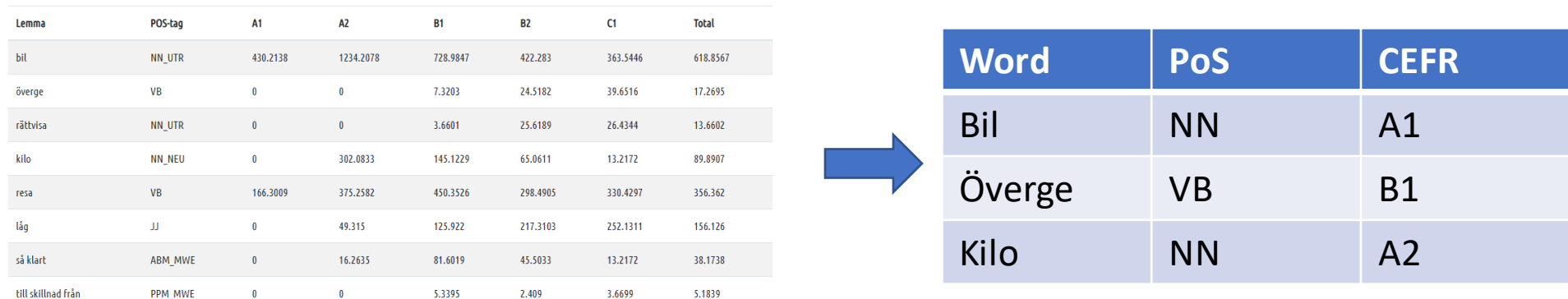


- First occurrence
- Threshold
- Maximum



# Mapping distributions to labels

- First occurrence and threshold agree largely
- First occurrence is easier
- First occurrence  $>$  threshold for textbook data



Lemma	POS-tag	A1	A2	B1	B2	C1	Total
bil	NN_UTR	430.2138	1234.2078	728.9847	422.283	363.5446	618.8567
överge	VB	0	0	7.3203	24.5182	39.6516	17.2695
rättvisa	NN_UTR	0	0	3.6601	25.6189	26.4344	13.6602
kilo	NN_NEU	0	302.0833	145.1229	65.0611	13.2172	89.8907
resa	VB	166.3009	375.2582	450.3526	298.4905	330.4297	356.362
låg	JJ	0	49.315	125.922	217.3103	252.1311	156.126
så klart	ABM_MWE	0	16.2635	81.6019	45.5033	13.2172	38.1738
till skillnad från	PPM_MWE	0	0	5.3395	2.409	3.6699	5.1839

Word	PoS	CEFR
Bil	NN	A1
Överge	VB	B1
Kilo	NN	A2



# Learning levels

- Words and levels
- Feature representations
- Machine learning model
- Learn association between features and levels

---

## Count features

---

Length (number of characters)

Syllable count

Contains non-alphanumeric character

Contains number

Is MWE

Character bigrams

n-gram probabilities

---

## Morphological features

---

Part-of-speech

Suffix length

Compound count

Compounds

Gender

---

## Semantic features

---

Degree of polysemy

Degree of homonymy

---

## Context features

---

Topic distributions

---



# Predicting levels

- New word
- Feature representation
- Machine learning model
- Level prediction

---

## Count features

---

Length (number of characters)

Syllable count

Contains non-alphanumeric character

Contains number

Is MWE

Character bigrams

n-gram probabilities

---

## Morphological features

---

Part-of-speech

Suffix length

Compound count

Compounds

Gender

---

## Semantic features

---

Degree of polysemy

Degree of homonymy

---

## Context features

---

Topic distributions

---



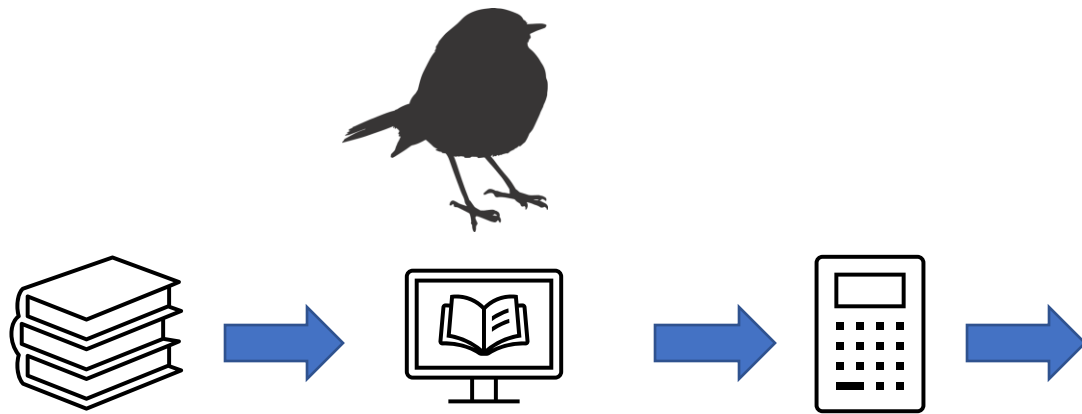
# Predicting levels

- Performance in line with previous research for different languages



# Towards senses

- Corpus re-annotation with sense distinctions



Lemma	POS-tag	A1	A2	B1	B2	C1	Total
bil	NN_UTR	430.2138	1234.2078	728.9847	422.283	363.5446	618.8567
överge	VB	0	0	7.3203	24.5182	39.6516	17.2695
rättvisa	NN_UTR	0	0	3.6601	25.6189	26.4344	13.6602
kilo	NN_NEU	0	302.0833	145.1229	65.0611	13.2172	89.8907
resa	VB	166.3009	375.2582	450.3526	298.4905	330.4297	356.362
låg	JJ	0	49.315	125.922	217.3103	252.1311	156.126
så klart	ABM_MWE	0	16.2635	81.6019	45.5033	13.2172	38.1738
till skillnad från	PPM_MWE	0	0	5.3395	2.409	3.6699	5.1839





# Towards senses

- SVALex → SenSVALex
- SweLLex → SenSweLLex



Sen\*Lex



# Limitations?

- Sparse data
- Sense splitting



Even sparser data



# Limitations?

- Use of external resources for sense disambiguation, e.g., dictionaries (cf Alfter et al., 2022)



Multi-word expressions

# Multi-word expressions

Expression	Automatic prediction		Word list level	
	Receptive	Productive	SVALex	SweLLex
kort sagt 'in brief'	C1	C1	A2	C1
just det 'exactly'	B2	B2	A1	C1
ingen fara 'no worries'	C1	C1	C1	A2
god natt 'good night'	C1	C1	A1	n/a
god morgon 'good morning'	C1	C1	A1	n/a
pommes frites 'french fries'	B2	C1	B2	n/a
på pin kiv 'for spite'	C1	C1	C1	n/a



# Multi-word expressions

- Crowdsourcing for difficulty estimation
- Rank words in relation to each other



# Multi-word expressions

Lättast	Uttryck	Svårast
<input type="radio"/>	bita i det sura äpplet	<input type="radio"/>
<input type="radio"/>	av ondo	<input type="radio"/>
<input type="radio"/>	betala för kalaset	<input type="radio"/>
<input type="radio"/>	för det mesta	<input type="radio"/>

Spara

*bita i det sura äpplet*

**Definition:** (idiomatiskt) tvingas göra något som man inte vill, t.e.x något obehagligt; vara tvungen att foga sig (Wiktionary)

Nuvarande uppgifts-id-nummer: 3288 .

Du har löst 0 uppgift(er) av totalt 326 . Du förväntas lösa 84 uppgifter.  
Du kan fylla i [feedbackformuläret](#) för att beskriva hur du fattade dina beslut.

Progress: 1 / 1300

Easiest	Expression	Hardest
<input checked="" type="radio"/>	The <b>dog</b> barked all night long. ( <b>dog</b> )	<input type="radio"/>
<input type="radio"/>	It took <b>years</b> for the bus to come. ( <b>years</b> )	<input type="radio"/>
<input type="radio"/>	The story he gave was something of an <b>overstatement</b> of the facts. ( <b>overstatement</b> )	<input type="radio"/>
<input type="radio"/>	He's only 16 months, but is already a good <b>counter</b> – he can count to 100. ( <b>counter</b> )	<input type="radio"/>

Next

# Ranking

- Aggregate votes into ranks
- Most difficult: 3
- Easiest: 1
- Other two: 2





# Ranking

- Aggregate votes into ranks

CEFR level	Word	CEFR level	Word
A1	bonjour	A1	bonjour
A1	copine	A1	copine
A2	vite	A2	vite
A2	frigo	A2	frigo
A1	confiture	A1	autocar
A1	vendeur	A1	vendeur
A1	autocar	B2	grille-pain
A2	guitariste	A2	guitariste
B1	apéro	A1	confiture
B2	grille-pain	B1	apéro
A2	bricoleur	B2	revolver
B2	revolver	A2	bricoleur
B1	festif	A2	clignotant
A2	clignotant	B1	festif
B1	corridor	B1	corridor
B1	vouvoyer	B2	enquêter
B2	enquêter	B1	vouvoyer
B1	vacarme	C2	haltérophilie
C1	arrachage	C2	chèvrefeuille
C2	chèvrefeuille	B1	vacarme
C1	discriminatoire	C1	surcoût
C1	surcoût	C1	discriminatoire
C2	haltérophilie	C1	arrachage
B2	perspicacité	B2	perspicacité
B2	lugubre	B2	lugubre
C1	affligeant	C1	affligeant
C2	inexorable	C2	inexorable
C2	enhardir	C2	protéiforme
C1	achoppement	C2	enhardir
C2	protéiforme	C1	achoppement



# Multi-word expressions

- Intuitive ratings similar for native and non-native speakers
- Intuitive ratings highly correlated with CEFR levels



# Manual annotation

## Lexicographic Annotation Tool (LEGATO)

[Guidelines](#)[Skipped items](#) <sup>0</sup>[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 - 8080

Current task: **NOMINAL TYPE**

Progress: 4809/8080

**SALDO lemgram**

ord..nn.1

**Part-of-Speech**

Noun (NN)

**CEFR level**

A1

Saldo sense: **ord..1**

Saldo primary descriptor: **språk..1**

Saldo secondary descriptor: **PRIM..1**

### Examples:

Ibland använder man de mer internationella \*\* orden \*\* addition , subtraktion , multiplikation och division .  
Skriv nya \*\* ord \*\* på lappar med svenska på en sida och ditt språk på andra sidan .

- ABSTRACT
- COUNTABLE
- COLLECTIVE
- ANIMATE
- PROPER NAME
- UNKNOWN

- CONCRETE
- UNCOUNTABLE
- NON-COLLECTIVE
- INANIMATE
- UNIT OF MEASUREMENT

Exit

Skip

Previous

Next

G

# Applications

- CEFRTools (<https://spraakbanken.gu.se/larkalabb/cefrtools>)

Input      Results      Explanations

---

Enter lemma  
hund

Select all models you want to include in the evaluation

- Word list lookup
- CEFR mapping techniques
- Threshold 1 0.3 Threshold value (0-1)
- Threshold 2 (1-to-10)
- COCTAILL 5-gram language model
- Indexed embedding space
- SiWoCo
- Lemma is a noun



## CEFR mapping results

Threshold 1 (value: 0.3)

	Lemma	Part-of-speech	CEFR level
SVALex	hund	NN_UTR	A1
SweLLex	hund	NN_UTR	A1

Threshold 2 (1-to-10)

	Lemma	Part-of-speech	CEFR level
SVALex	hund	NN_UTR	A1
SweLLex	hund	NN_UTR	A1

## COCTAILL 5-gram language model predictions

	A1	A2	B1	B2	C1
Probabilities	1.14e-7	4.30e-8	6.06e-8	6.73e-8	5.81e-8
Log probabilities	-15.98	-16.96	-16.62	-16.51	-16.66

Highest probability at: A1

Highest log probability at: A1

## SiWoCo predictions

Lemma	Receptive prediction	Productive prediction
hund	A2	B1



# Applications

- Text evaluation



Vad är egentligen laktosintolerans? Att vara laktosintolerant betyder att man är överkänslig för laktos (mjölksocker). Laktos är en kolhydrat som finns naturligt i mjölk och andra mejeriprodukter. Till exempel grädde och yoghurt. Laktosintolerans orsakas egentligen av laktasbrist, det vill säga brist enzymet laktas som bryter ner laktos i tunntarmen. Utan laktasenzym förblir laktosen ospjälkad i tunntarmen och går vidare till tjocktarmen där den mjölksockret bryts ner av bakterierna som finns där och gaser bildas. Detta gör att man kan få magknip, gasbildning, diarré och/eller en känsla av upplåsthet. Symptomen är individuella och kan variera. En del får väldigt ont medan andra får lindriga besvär. Man kan uppleva att man tål mycket laktos ena dagerna och bara lite en annan. Tolerans av laktos kan också till exempel bero på måltidens sammansättning.

What do you want to assess? ⓘ

Learner essay Text readability

Show all words of the following CEFR level(s) ⓘ

- A1
- A2
- B1
- B2
- C1

Additional options ⓘ

- Mark all potentially incorrect words
- Use Spellchecker

Edit text Reset

## Evaluation

**Suggested overall level:** C1

*Given the limited amount of underlying data, this CEFR level should be considered as a suggestion and its use as a basis for decisions in high-stakes assessment is discouraged.*

### Detailed evaluation

Number of sentences	9
Number of tokens	146
Non-lemmatized forms	2
Average sentence length	16.22
Average token length	4.99
Average dependency length	2.52
LIX score	42 (normal)
Nominal ratio	1.09
Pronoun-to-noun ratio	0.35

# Applications

- Exercise generation

## Vocabulary Multiple Choice

B1 ▾

Change level

Click to generate!

4	Tillsätt lite mjölk i taget medan du fortsätter_____.	vallfärda ▾	⊕
3	" Vi behöver inte ta_____.	vallfärda idrotta deklarera tillföra vispa	✗
2	Hon kom ihåg att hon hade varit här en gång med Brigie och Mary och plockat björnbär och senare hade stugan varit fylld av den stickande lukten av kokande_____och de hade fått sylt till teet i flera veckor efteråt .	vispa	✓
1	Själv ska jag handla för att göra en_____i ugnen som räcker till hela familjen .	kaka ▾	✗

## Word guess

Tries: 1/7


Score: 0

**Definition:**

blir röd i ansiktet (ofta för att man är generad)

**Help:**

Show translation

R  D N 

A	B	C	E	F	G	H	I	J	K	L	M	O	P	Q	S	T	U
V	W	X	Y	Z	Ä	Å	É										



# Open questions

- Large Language Models
- Learn CEFR descriptors automatically



# Demonstration

