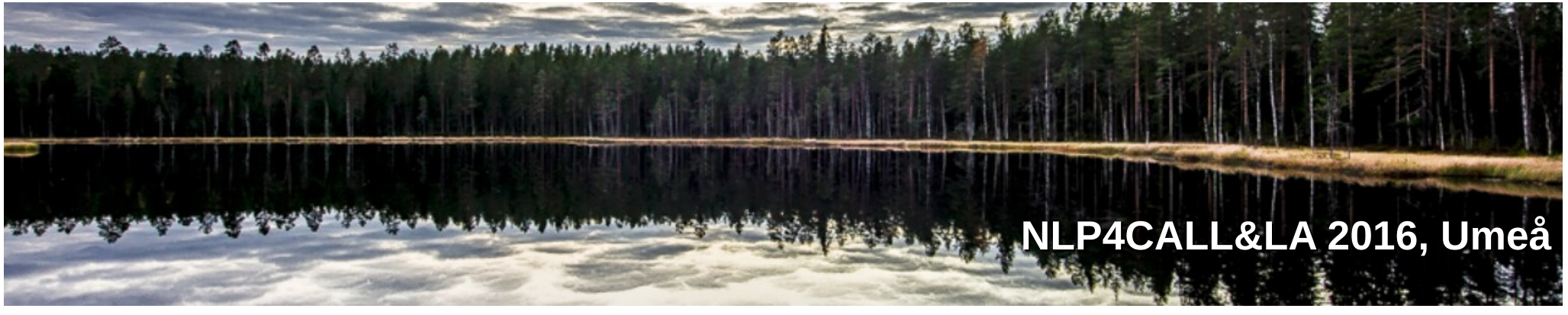
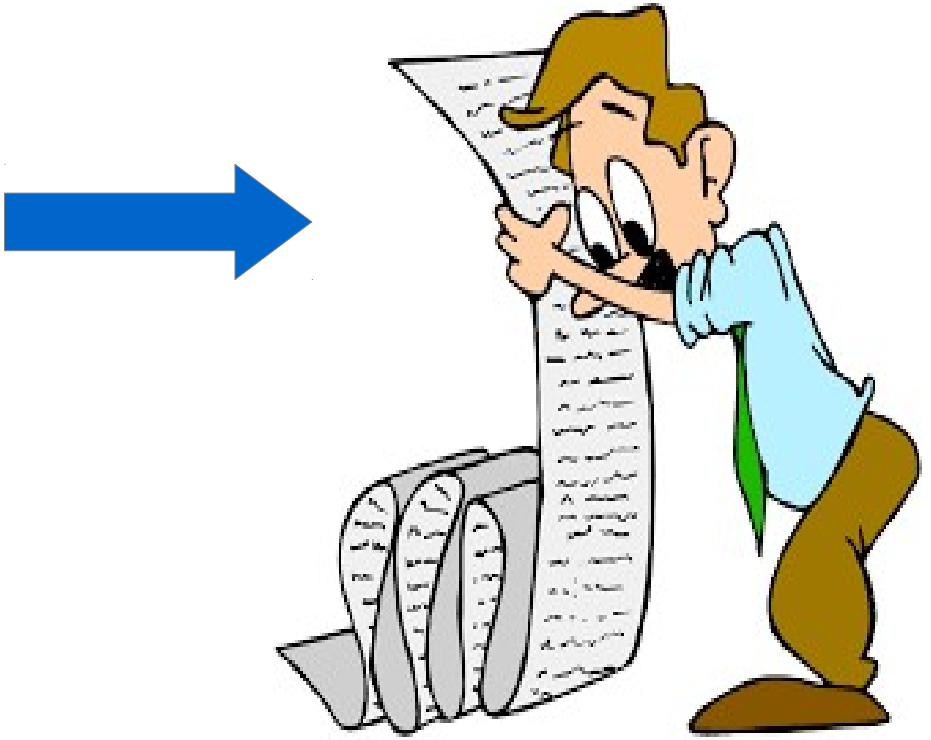
SweLLex:

second language learner's productive vocabulary

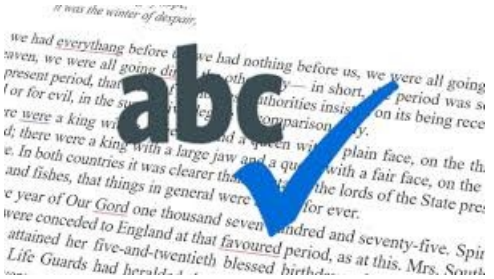
Elena Volodina, Ildikó Pilán, Lorena Llozhi,
Baptiste Degryse, Thomas François

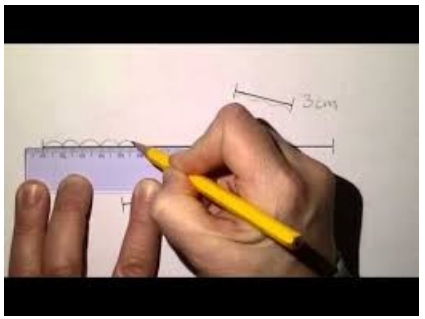


From a corpus of L2 essays to a list of words



Not “just essays” → normalized essays





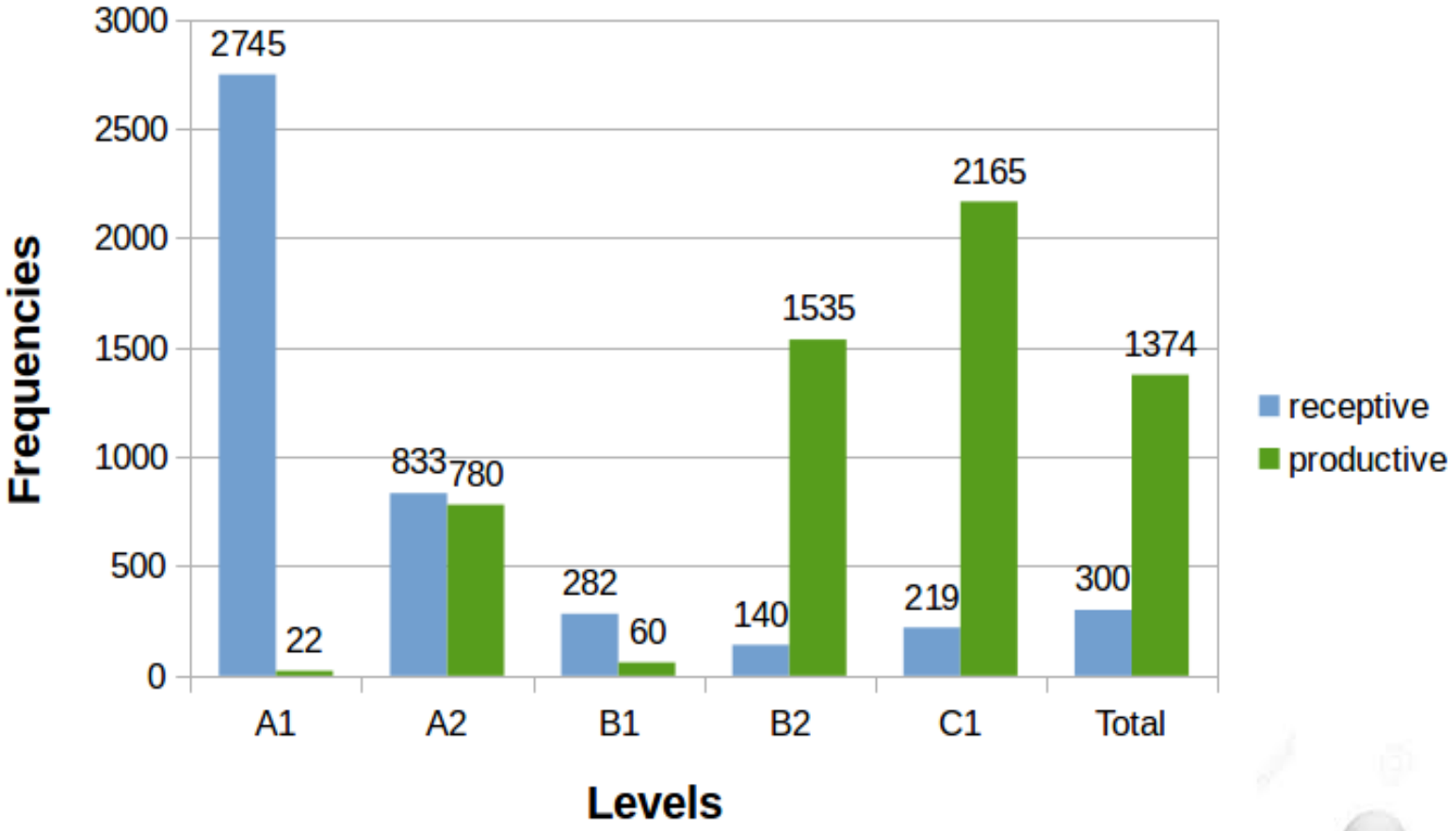
- **Levenstein distance**
  - Good for advanced levels
    (edit distance of 1)
  - Fails at lower levels (with multiple changes)

- **LanguageTool + candidate ranking**
  - 73% correct variant selection
  - Failed to identify 30% of spelling errors

L2 Swedish productive vocabulary in numbers

Lev	#items	#new	#MWE	#hapax	Doc.hapax examples	#SVALex	#EVP
A1	398	398	15	0	-	1,157	601
A2	1,327	1,038	82	12	*i kväll* "tonight"	2,432	925
B1	2,380	1,542	206	36	*fylla år* "have birthday"	4,332	1,429
B2	2,396	959	264	58	*fatta beslut* "make a decision"	4,553	1,711
C1	3,566	1,545	430	152	*sätta fingret* "put a finger on sth"	3,160	N/A
C2	145	7	12	1	*i bakhuvudet* "in mind"	N/A	N/A

Productive vs receptive vocabulary



Distributions of the verb *arbeta* (Eng *work*)

http://cental.uclouvain.be/svalex/#tab-search-svalex

SweLLex: second language learners' productive vocabulary

E. Volodina[1] I. Pilán[1] Lorena Llozhi[1] B. Degryse[2] T. François[2,3]

(1) Språkbanken, University of Gothenburg (2) CENTAL, IL&C, UCLouvain (3) Post-doc FNRS, CENTAL, IL&C, UCLouvain

1. Objectives

SweLLex is a lexical resource for Swedish as a foreign language that ...

- is aimed at learners and teachers of foreign and second language (L2) Swedish, lexicographers, L2 curriculum developers, or ICALL researchers
- describes the *frequency distributions* of 6,965 words and expressions across the Common European Framework of Reference (CEFR)
- describes the *productive vocabulary*, as the resource is based on a corpus of essays written by second language learners of Swedish

Ask SweLLex:
At which level should I know the word "tendens"?

2. Similar recent studies

- **CEFR reference level descriptors** are available for various languages:
  - Description of the competence expected from an L2 learner by CEFR levels.
  - Lists of words, structures, and expressions associated with functions or themes.
  - No reference level for Swedish + issues about their conception (Allkrem, 2007; Hultgren, 2007) .
- **Kelly list**: frequency-based word list with CEFR levels (Kilgarriff et al., 2014)
  → Shortcomings: frequencies collected on a L1 web corpus + level division based on frequency threshold.
- **SVALex**: frequency-based word list with CEFR levels (Francois et al., 2016)
  → Describes *receptive vocabulary* since it is based on reading comprehension texts graded for CEFR levels.
- **Other lexical resources for Swedish** (but not related to the CEFR-scale):
  - **Base Vocabulary Pool** (BaseVoc): 8,215 lemmas composed of stylistically neutral and general-purpose simple words (Forsbom, 2003)
  - **Swedish Academic word list**: vocabulary used for writing academic papers (Carlund et al., 2012)
  - **Lexin**: series of lexicons aimed at immigrants (Huk et al., 2010)
- Other languages: **English Vocabulary Profile** (Capel 2012) ; **FLELex** (François et al., 2014)

3. Methodology

3.1. The corpus

- Building SweLLex requires a CEFR-graded corpus of L2 essays in Swedish !
- We used **SweLL**, a corpus of L2 essays graded by teachers (Volodina et al., 2016)
- Essays written by various first language groups, on multiple topics, ranging from A1 to C2, were used
- All texts were lemmatized, POS-tagged with the Korp pipeline (Borin et al., 2012)
- Multi-word expressions (MWE) were detected and checked with the SALDO lexicon (Borin et al., 2013)

3.2. L2 text normalization

- First lemmatization of SweLL-corpus rendered 4,308 non-lemmatized items, categorized as *misspellings*, *compounds*, *hyphenation*, *foreign words*, *acronyms*. To tackle those cases we experimented with two normalization approaches on single-word level

Approach 1: Levenstein distance based on Saldo morphology lexicon. Works well with edit distance 1, which is seldom the case for lower levels. Table below shows number of correctly returned suggestions per level

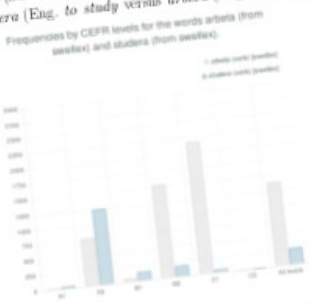
Level	Correct/total
A1	7/20
A2	13/20
B1	13/20
B2	15/20
C1	16/20

4. Description of the resource

SweLLex contains 6,965 entries, of which 1,490 are not lemmatized. Calculations below exclude erroneous items from counts. The distribution of SweLLex items across the 6 CEFR levels looks like this:

Lev	#items	#new	#MWE	#hapax	Dis:hapax examples	#SVALex	#EVP
A1	998	998	15	0	-	1,157	601
A2	1,327	1,038	82	12	*i kväll* "tonight"	2,432	925
B1	2,380	1,542	206	36	*fylla år* "have birthday"	4,332	1,429
B2	2,366	959	264	58	*fatta beslut* "make a decision"	4,553	1,711
C1	3,566	1,545	430	152	*sätta fingret* "put a finger on sth"	3,160	N/A
C2	145	7	12	1	*i bakhuvudet* "in mind"	N/A	N/A

SweLLex can be searched online, comparing items from the same resource (e.g. SweLLex) or items from two resources (SweLLex and SVALex, i.e. productive and receptive vocabulary), e.g. Swedish verbs *studera* (Eng. *to study* versus *arbeta* (Eng. *to work*):



The resource is freely available at http://cental.uclouvain.be/svalex/

5. Comparisons and insights

We compared SweLLex with two L2 resources: the English Vocabulary Profile (EVP) (productive L2 English vocabulary) and SVALex (receptive L2 Swedish vocabulary).

Swedish versus English productive vocabulary

- Patterns for new items per level are comparable on A1-B1 levels, but deviate at B2.
  → Hypothetically, it depends upon limited topic variety in the SweLL corpus
- Number of multi-word expressions (MWE) is, however, steadily growing in SweLLex, despite limited topic variety.
  → Hypothetically, MWEs can be used as predictors of level/complexity development in L2 learner writing

Swedish productive versus receptive vocabulary

- Similarly to EVP, patterns for new items per level are comparable (see SweLLex distribution table) on A1-B1 levels, but deviate at B2.
- It is possible to compare patterns in vocabulary learning for the same item using

For details, visit our poster