GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Introducing SVALex:

a corpus-based lexical resource for second language learning

Elena Volodina, Ildikó Pilán & Thomas François

COST ENeL WG3 meeting

Herstmonceux, 13 August 2015

EUROPEAN NETWORK OF e-LEXICOGRAPHY

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

To start with ...

Our task: 1. identify relevant vocabulary for second language learners
2. at different proficiency levels (acc. to CEFR)

More precisely: *Which words at which level? How many?*

Focus on: receptive vocabulary (as opposed to productive)

Alternatives to SVALex

Kelly - based on native speaker corpus (2010), linked to CEFR, no validation (Kilgarriff et al., 2014)

Lexin – many sources, no frequency information or recommendations on learner level (Hult et al., 2010)

Base Vocabulary Pool - based on native speaker corpus (Forsbom, 2006)

Academic word list – based on corpora of academic papers, for university students with good command of Swedish (Carlund et al., 2012)

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Pedagogical framework

- **CEFR** – Common European Framework of Reference
- 6 proficiency **levels**

The six proficiency levels are named as follows:

C2	Mastery	} Proficient user
C1	Effective Operational Proficiency	
B2	Vantage	} Independent user
B1	Threshold	
A2	Waystage	} Basic user
A1	Breakthrough	

Pedagogical framework

- **Can-do** statement on vocabulary range (A2):

Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.

Has sufficient vocabulary for the expression of basic communicative needs.

Has sufficient vocabulary for coping with simple survival needs.

Pedagogical framework

- **Can-do** statement on vocabulary range (A2):

*Has **sufficient vocabulary** to conduct **routine, everyday transactions** involving **familiar situations and topics**.*

*Has sufficient vocabulary for the expression of **basic communicative needs**.*

*Has sufficient vocabulary for coping with **simple survival needs**.*

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Stage 1. Keyword extraction

1. Data: corpus of coursebooks, digitized, annotated, linked to CEFR

2. Method: frequency extraction, adjusted or distributed

3. Result: list analysis & comparison

4. Manual cleaning

Digitized coursebooks

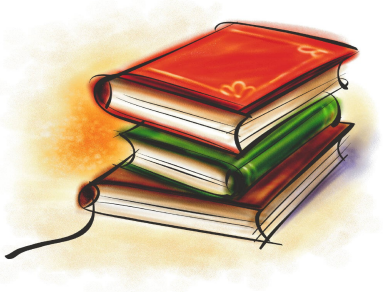
Frequency extraction

Manual cleaning

Repeat steps 2-5 over and over again

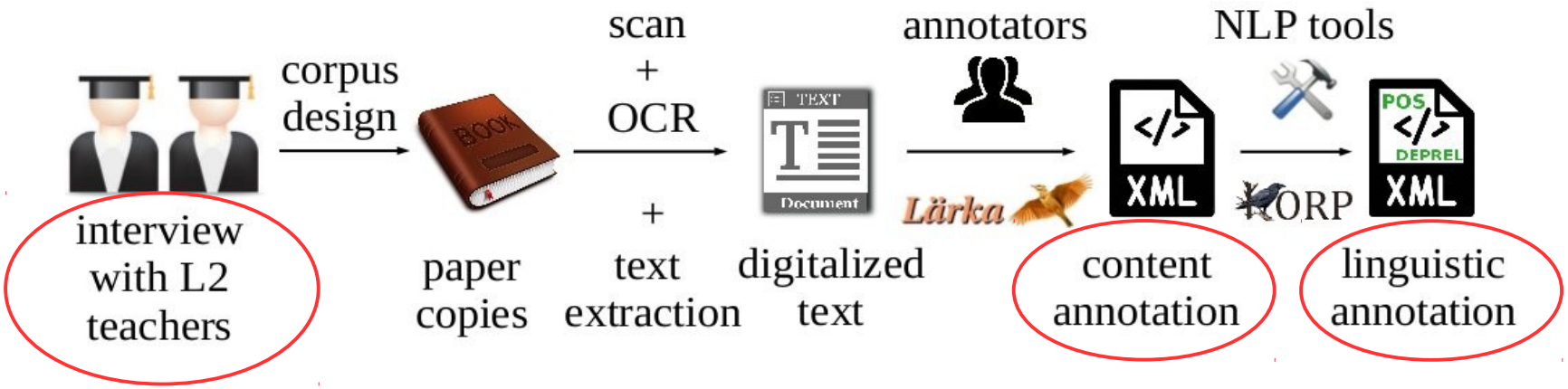
GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

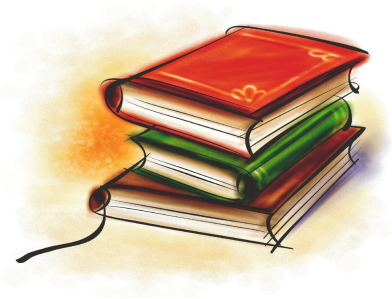


Data: COCTAILL

Corpus **o**f CEFR-based **T**extbooks **a**s **I**nput for **L**earner **L**evels' modelling -



CLT Språk-BANKEN

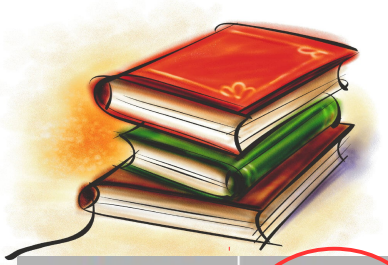


COCTAILL content “ingredients”

ELEMENTS	ATTRIBUTES
Coursebook	title, author etc.
└ Extras	
└ Lesson	CEFR level, (title)
└ Text	topic (e.g. arts), (title)
└ Genre	genre family (e.g. facts), macrogenre (e.g. dialogue)
└ Subheadings	
└ Activity instruction	skills/competences, activity type/format
└ Task	
└ List	skills/competences, linguistic unit
└ Language Example	
Common attributes:	id, ref

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN



COCTAILL

CEFR level	Nr. of books	Nr. of authors	Nr. of lessons	Nr. of texts	Nr. of tasks	Nr. of sentences (texts)	Nr. of tokens (texts)
A1	4	10	37	101	160	1581	11132
A2	4	10	105	232	244	4217	37259
B1	4	12	83	345	389	6510	79402
B2	4	8	31	314	368	8527	101583
C1	2	2	22	115	333	5085	71991
Total	18 (12 titles)	42 (26 different names)	278	1106	1494	25920	301367

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Frequency measure

FLELex-inspired - François et al. (2014)

Lemgram-based - baseform + part-of-speech; including multi-word expressions

Dispersion-based - raw frequencies normalized by dispersion index (Carroll et al., 1971; François et al., 2014)

Statistics granularity – raw versus dispersed statistics kept for each textbook so that further analysis can be more refined

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Resulting list

Level	# items	# new items	# MWE	# doc.hapax	Doc.hapax examples	# EVP
A1	1,157	1,157	92	99	*postnummer* ”zip code”	601
A2	3,327	2,432	300	635	*jurist* ”lawyer”	925
B1	6,554	4,332	617	1,868	*öga mot öga* ”face to face”	1,429
B2	8,728	4,553	880	3,051	*snigelfart* ”snail speed”	1,711
C1	7,564	3,160	783	2,709	*inom synhåll* ”within eyesight”	N/A
Total	15,681	15,681	1,426	8,362		

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Manual cleaning

Initial manual inspection - of items without assigned lemma

Manual cleaning

Automatic matching - to lemgrams in Base Vocabulary Pool, Lexin and Kelly

Resource	# items	# overlap	# missing
SVALex	15,681	N/A	N/A
Swedish Kelly	8,425	5,757	9,924
Base Vocabulary	8,220	4,964	10,717
Lexin	30,684	9,039	6,642

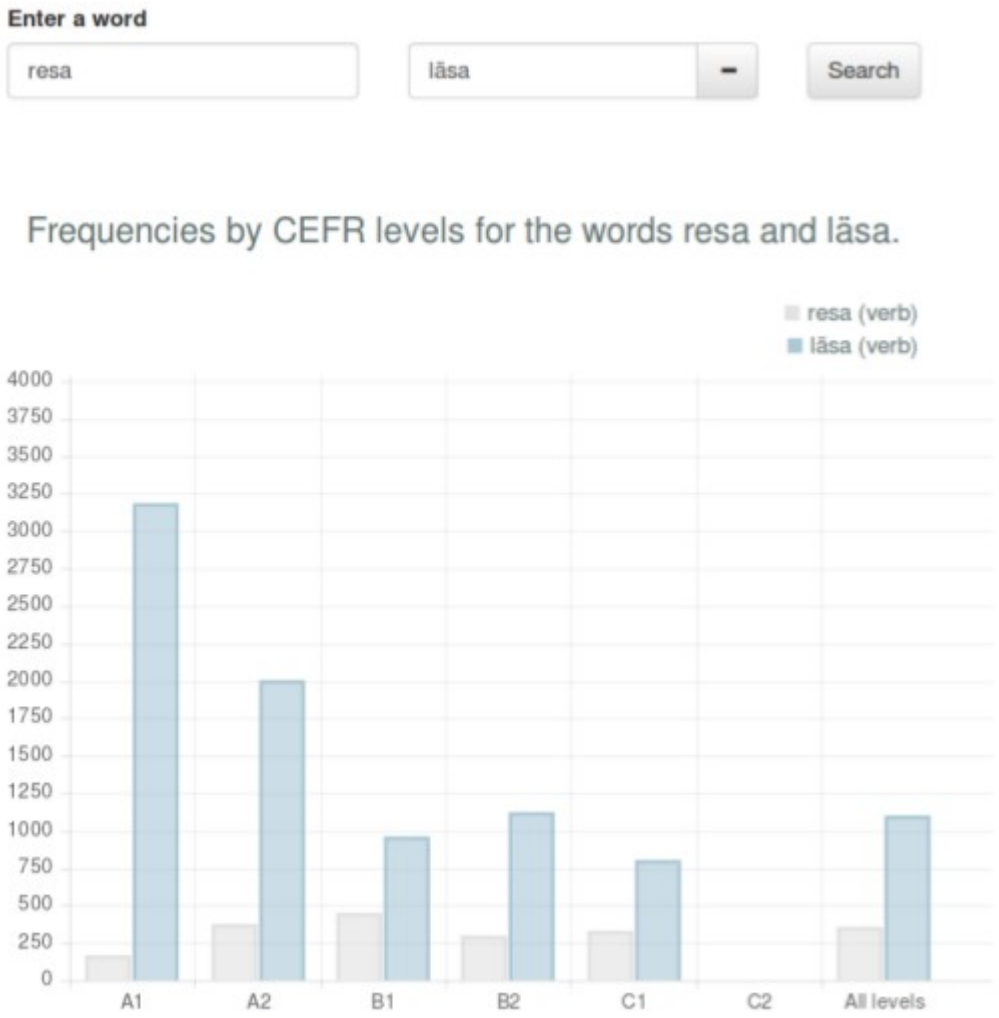
Manual inspection and result analysis – focus on SVALex items missing from the above resources

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Browsing SVALex

- travel (verb)
- read (verb)



Future

(1. Keyword extraction and cleaning)

2. Linking to other resources, aligning by sense

3. Filtering (central-peripheral) and extending

4. Visualization

5. Validation

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Stage 2. Linking to other resources

1. Kelly: grammar info, style, spelling variants, examples

→

* att klä (el. kläda) *put on (clothes)*
  e.g. klä av sig/klä på sig

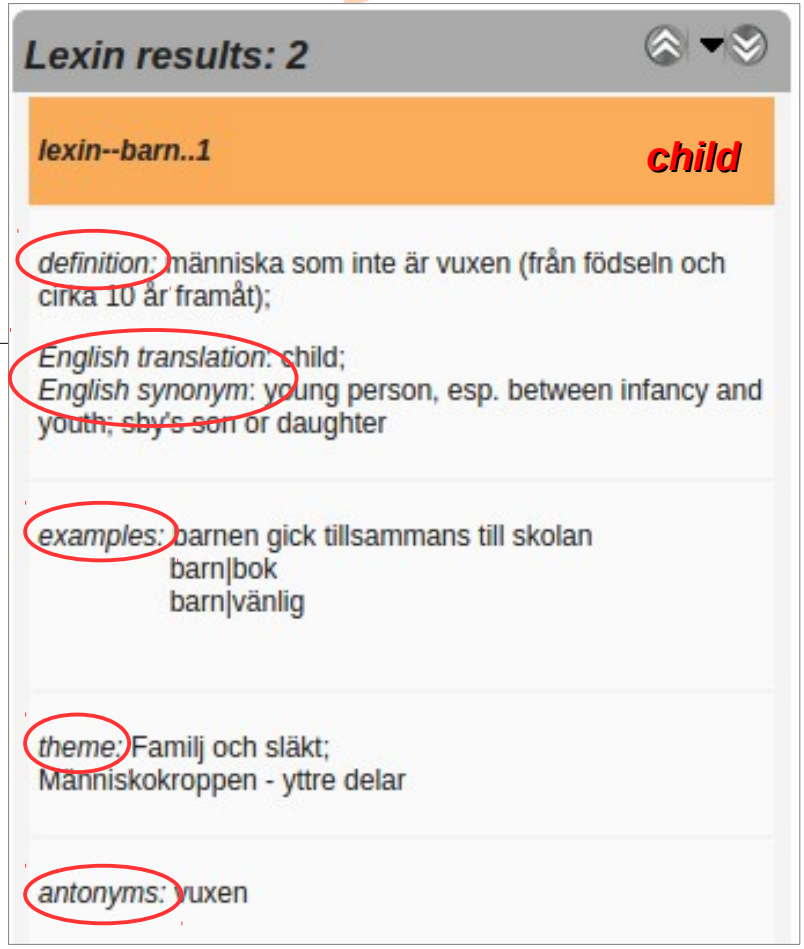
* en far (el. fader; vardagl. farsa) *father*

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Stage 2. Linking to other resources

2. Lexin: examples, definitions, patterns, translations, domains, antonyms...



GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Stage 2. Linking to other resources

3. Bring, FrameNet, Swesaurus, Simple+:
domains, synonyms, related words

Relaterade ord (SWE-FN)					
Kinship					
lillasyster	bror	syrra	halvsyster	halvsyskon	niece
fyrling	åttling	sjuling	lillebror	brorsa	brorsbarn
syster	småsyskon	storasyskon	tsardotter	tvilling	systerdotter

… + more

4. Manual control and word sense alignment

Stage 3. Filtering and extension

1. Central vs peripheral vocabulary:

define frequency threshold, decide on the overlap between coursebooks, document hapaxes

Which words are central at which level? When should they be introduced?

A2

4 (of 4) >>	december (**December**);	RF=2;	DF=326	=> central?
3 (of 4) >>	glass (**ice cream**);	RF=2;	DF=201;	=> central?
2 (of 4) >>	väldig (**huge**, adj);	RF=1;	DF=104;	=> peripheral?
1 (of 4) >>	hämta sig (**to recover**);	RF=0,8;	DF=17;	=> peripheral?

Stage 3. Filtering and extension

2. **Additional evidence?** From where? Essays? Other parts of the COCTAILL corpus (lists, exercises, examples)?

3. **Additions from Kelly:**
   evidence from L1 corpora in 9 languages. All Kelly items not yet in SVALex – to peripheral vocabulary? At which level? The one that is assigned in Kelly?

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Stage 4. Visualization

Choose level

- [ ] A1 [ ] *new items only*
- [x] A2 [x] *new items only*
- [ ] B1 [ ] *new items only*
- [ ] B2 [ ] *new items only*
- [ ] C1 [ ] *new items only*

Search item:

Select

Filters

Central/peripheral

Category *(abbr, words, phrases, idioms, compounds)*

Parts of speech

Grammar *(???)*

Domains

Usage/style

Hapaxes

Clear filters

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Stage 4. Visualization

Choose level

- [ ] A1 [ ] *new items only*
- [x] A2 [x] *new items only*
- [ ] B1 [ ] *new items only*
- [ ] B2 [ ] *new items only*
- [ ] C1 [ ] *new items only*

Search item:

Select

Filters

Central/peripheral

Category *(abbr, words, phrases, idioms, compounds)*

Parts of speech

Grammar *(???)*

Domains

Usage/style

Hapaxes

Clear filters

Search results

A2, new items (2432 hits)

fastighet *noun* A2
fattig *adjective* A2, B1, B2, C1
favoritblomma
favoritcitat
favoritdjur
favoritfärg
favoritförfattare
favoritmusiker
favoritnamn
favoritord
feber *noun* A2, B1, B2
feberfri
febertermometer
februari
fem
femtio
festmat
fet
figur
fika
fika
fikapaus

Save list

feber

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Stage 4. Visualization

Choose level

- [ ] A1 [ ] *new items only*
- [x] A2 [x] *new items only*
- [ ] B1 [ ] *new items only*
- [ ] B2 [ ] *new items only*
- [ ] C1 [ ] *new items only*

Search item:

Select

Filters

Central/peripheral

Category *(abbr, words, phrases, idioms, compounds)*

Parts of speech

Grammar *(???)*

Domains

Usage/style

Hapaxes

Clear filters

Search results

A2, new items (2432 hits)

fastighet *noun* A2
fattig *adjective* A2, B1, B2, C1
favoritblomma
favoritcitat
favoritdjur
favoritfärg
favoritförfattare
favoritmusiker
favoritnamn
favoritord
feber *noun* A2, B1, B2
feberfri
febertermometer
februari
fem
femtio
festmat
fet
figur
fika
fika
fikapaus

Save list

feber *noun* A2, B1, B2
feber, febers, febern, febrar, febrarna

1. hög temperatur i kroppen
   English: fever, temperature
   Examples:
   * febern steg framåt kvällen
   * barnet hade 39 graders feber
   * på morgonen hade febern gått ner
   * feber|fri
   * feber|nedsättande
   * feber|sjukdom

COCTAILL examples:

** Jag har ganska hög feber och ont i halsen.[A2]*
** Några av dem blir mycket sjuka med feber, värk i lederna och ibland ansiktsförlamning. [B1]*

DOMAIN:

** Health status, Ailment, Symptom, Experiencer [FrameNet], Medicine [Simple]*

Related words [Health status]:

aidssjuk, allmänbefinnande, anemisk, ansiktssmärta, ...

2. upphetsat tillstånd
   Examples:
   * guld|feber
   * res|feber

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Stage 5. Validation with students and teachers

Lärka-based - exercises (indirect evaluation)

List browsing – tasks for students and teachers (questionnaire to follow)

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

Thank you!

GÖTEBORGS UNIVERSITET

CLT Språk-BANKEN

From FLELex towards CEFR-Lex family?

FLELex | Home | Search FLELex | Download FLELex

Introduction to FLELex

What is FLELex? FLELex is a lexicon for French as a foreign language (FFL) that reports the normalized frequencies of words (lemmas) across each level of the CEFR (Common European Framework of Reference for Languages).

The frequencies have been estimated on a corpus of FFL textbooks and FFL simplified readers. More details on the corpus, the computation and normalization of the word frequencies, and the ressource itself can be found in:

> François, T., Gala, N., Watrin, P. et Fairon, C. FLELex: a graded lexical resource for French foreign learners. In *the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland, 26-31 May.

What's in FLELex? For every word in FLELex, you can find his part-of-speech (P.O.S.) along with his normalized frequency for each level of the CEFR, and his total normalized frequency in our corpus. Here are some of entries from FLELex: