

Linking CEFR-based learner profiles to lexicographic data

Kris Heylen (Dutch Language Institute, INT)
Ilan Kernerman (Lexicala by K Dictionaries)
Carole Tiberius (Dutch Language Institute)

Acknowledgements:

Jelena Kallas (Institute for the Estonian Language)
NexusLinguarum COST Action (CA18209)

l2p-2023

University of Gothenburg

20 April 2023

Overview

- Background
- Aims
- CEFR-graded word lists
- Dutch pilot study: linking CEFR lists and lexicographic data
- Cross-lingual pilot study: comparison of CEFR-levels
- Intermediate conclusions
- Towards a linked CEFR infrastructure

Background of collaboration

- European Network of e-Lexicography ([ENeL COST](#) 2013-2017)
 - ELEXIS ([Horizon 2020 RIA](#) 2018-2022)
 - Lexicographic research infrastructure project
 - Datamodels and tools for linking lexicographic data (Dictionary Matrix)
 - NexusLinguarum COST-action ([CA18209](#), 2019-2024)
 - Network for Linguistic Linked Data Science
 - [STSM](#): Ilan hosted by INT in April 2022
 - Lexicographic resources for language learners
 - Lexicala: Global multi-layer lexical datasets
 - INT: general dictionary, verb patterns, phrases and idioms for Dutch
 - How can we make our lexicographic resources more useful for language learning applications?
- => Linking to learner profiles and CEFR levels

Aims

Dictionaries have a long tradition of supporting language learning but typically do not have learner level profiling. We want to

- upgrade the usability of our rich lexicographic data for creators and users of CEFR-based vocabulary learning materials:
 - definitions
 - examples of usage
 - multiword expressions
 - other e.g., synonyms, domain labels, register, pronunciation, media ...
- link lexicographic data and CEFR through existing CEFR-wordlists
- cross-lingualize datasets of different languages
- make the results available as Linked (Open) Data
- reach out to CEFR-communities (SLA, CALL, teaching&testing)
- initiate a new joint project for a linked CEFR infrastructure

CEFR-graded word lists

- Common European Framework of Reference for Languages
 - Council of Europe: language-independent [Companion Volume](#)
 - Reference Level Descriptions for individual languages ([approx. 11 lang.](#))
- CEFR grading of vocabulary
 - word lists for > 30 languages (notable projects: [Kelly](#) (2009-2011), [English Profile](#) (2009-2012) [CEFR-Japan](#) x28 (2008-..) [CEFRLex](#) (2014-.. fr/nl/en/sv/es/de)
 - different methodologies: learner/textbook corpus vs. expert-based
 - mostly word-based, only a handful are sense-disambiguated (EP, NT2Lex)
- Linked to lexicographic data?
 - fully: only English ([Cambridge](#) and [Oxford](#) advanced learner's dictionaries)
 - in progress: Estonian (EKI) [sonaveeb.ee](#)
 - other languages: isolated word lists without lexicographic linking
- **Pilot studies**
 - **Dutch**: assign CEFR-labels to lexicographic entries via NT2Lex (CEFRlex Dutch)
 - **Cross-lingual**: linking CEFR-lists through multilingual lexicographic data

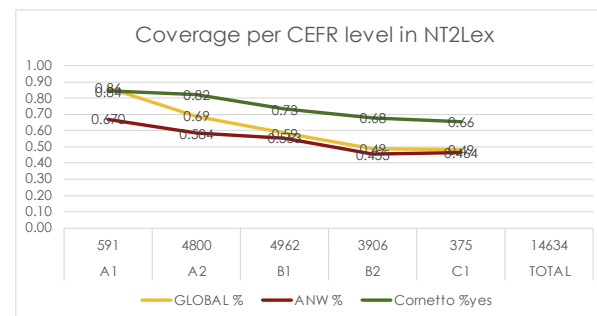
Pilot study on Dutch

- CEFR-wordlist NT2Lex for Dutch (Tack et al. 2018)
 - compiled at UCLouvain as part of CEFRLex
 - derived from corpus of graded textbook readers for Dutch learners
 - +17K entries with frequency distribution over levels A1-C1 per entry
 - level assignment based on normalized freq./first occurrence
 - senses: linked to OpenDutchWordNet (Postma et al. 2016) by automatic WSD
- Lexicographic Data for Dutch
 - General Dutch Dictionary (**ANW**) 80K entries, in development @INT
 - **Cornetto** (Vossen et al. 2008) DutchWordNet+, 92K lemmas / 120K senses
 - Lexicala **Global** (K Dictionaries) Dutch part, 31K lemmas / 35K senses
- How feasible is linking these 2 types of resources?
 1. Linking on the lemma-POS level through python scripts
 2. Manual linking on sense-level of a sample to identify potential issues

Pilot study on Dutch

1. Linking on the lemma-POS level (adjectives, nouns, verbs)

NT2LEX		Lexicala Global			INT - ANW				Cornetto		
level	words	yes	no	cov.	full	minimal	no	cov.	yes	no	cov.
A1	591	511	80	0.86	119	396	76	0.67	499	92	0.84
A2	4800	3305	1495	0.68	886	2802	1112	0.58	3957	843	0.82
B1	4962	2903	2059	0.58	625	2746	1591	0.55	3643	1319	0.73
B2	3906	1910	1996	0.48	410	1776	1720	0.45	2651	1255	0.67
C1	375	182	193	0.48	41	174	160	0.46	246	129	0.65
TOT	14634	8811	5823		2081	7894	4659		10996	3638	



- Coverage decreases significantly for higher CEFR levels
 - from ±87% at A1 to ±50% at B2 and C1
 - Unexpected in dictionaries for intermediate to advanced learners
- Causes: problematic CEFR level assignments in NT2Lex
 - Many idiosyncrasies of the chosen corpus (textbooks)
 - Many low frequent words with unreliable CEFR level
 - Complex mapping of frequency distribution to CEFR label

⇒ **NLP-compiled dataset VS curated lexicographic database**
methodological consistency VS end product consistency

Pilot study on Dutch

2. Manual linking on sense-level: e.g. *aardig* [nice] A1

word	tag	Transl.	D@A1	sense_id	Cornetto	Global	ANW
aardig	ADJ	nice	0,34	r_a-8687	sympathiek	NL_SE00000472	1.0
aardig	ADJ	lovely	0,14	r_a-9035	leuk	NL_SE00000472	2.0 3.0 4.0
aardig	ADJ	quite	A2	r_a-8688	behoorlijk	NL_SE00000473	5.0 6.0

Annotations: A red arrow labeled "lumping" points from the Global cell of the first row to the Global cell of the second row. A red arrow labeled "splitting" points from the ANW cell of the second row to the ANW cell of the third row.

<Definition> als [A1] je [A1] iemand [A2] of iets [A1] aangenaam [B2] of [A1] vriendelijk [A2] vindt [A1]
if you someone or something likeable or friendly find

<Synonym> sympathie [B1]

<Translation lang="fr" > sympathique [A1]

<Translation lang="fr" > gentil [A1]

<Translation lang="en" > friendly [A1]

<Translation lang="en" > pleasant [A2]

<Example> De buurvrouw [A2] is een aardig mens. [A1]
The neighbour is a nice person

[A1] same CEFR level as headword

[A2] one CEFR level higher

[B1] two CEFR levels higher

[B2] three CEFR levels higher



- Differences in granularity of sense distinctions, as expected for WordNet / learner's dictionary / scholarly dictionary
- Words used within definitions, examples and **translations** are not always situated on the appropriate CEFR levels (= same or lower)

Cross-lingual pilot study

Common framework, based on the same can-do statements

- Same CEFR-label for lexical equivalents across languages?
- Similar amount of words per CEFR-level in each language?
 - quite variable but more agreement on lower levels

Milton & Alexiou 2009

CEFR level	XLex (5000 max)	
	 English	 French
A1	<1500	1160
A2	1500 - 2500	1650
B1	2750 - 3250	2422
B2	3250 - 3750	2630
C1	3750 - 4500	3212
C2	4500 - 5000	3525











Capel 2010: English profile

Words at A1	601
New words at A2	925
New words at B1	1,429
New words at B2	1,711

CEFR-J Wordlist [Tono 2019](#)

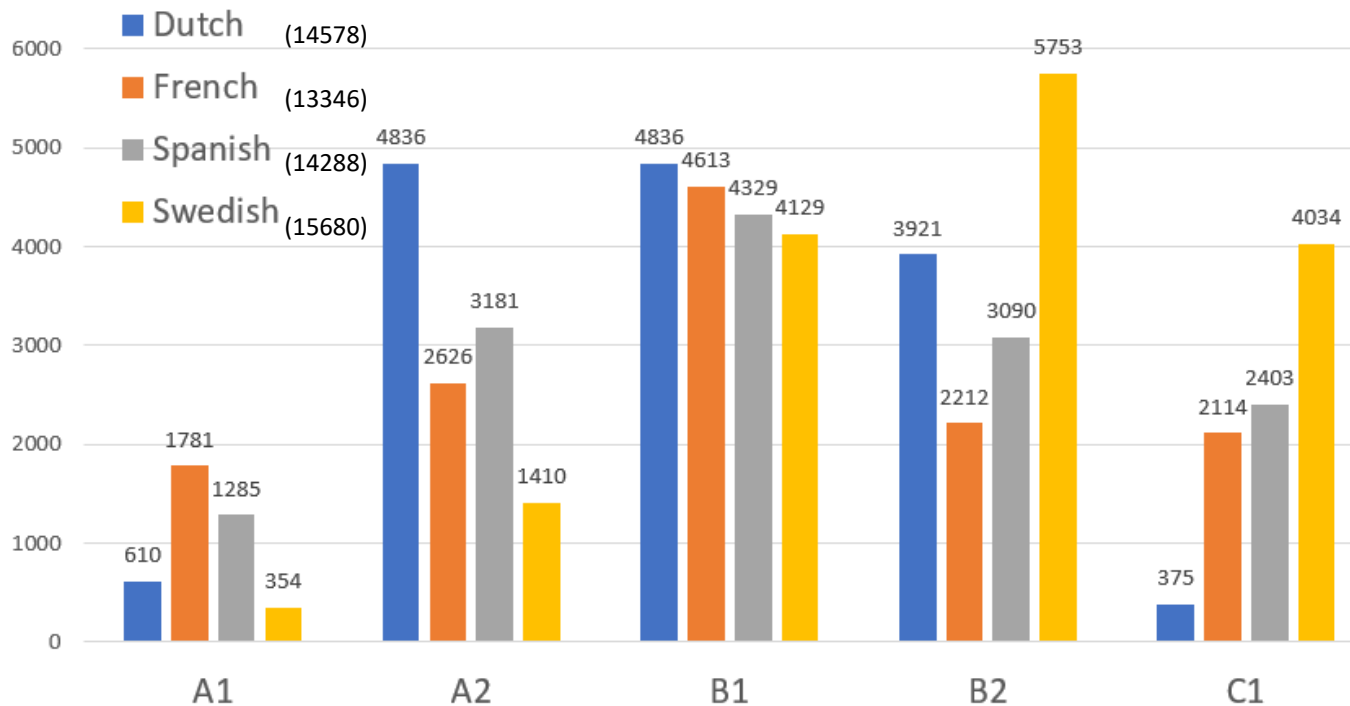
A1	A2	B1	B2	Total
1,068	1,358	2,359	2,785	7,570

Decoo 2011

Authors	A1	A2	B1	B2	C1	C2
 Van Ek & Alexander 1980		700	1,100-1,500			
 Van Ek 1976			1,600			
 Meara & Milton 2003	<1,500	1,500-2,500	2,750-3,250	3,250-3,750	3,750-4,500	4,500-5,000
 Schmitt 2008, see also Nation 2006						15,000
 Coste a.o. 1976			3,000			
 Beacco a.o. 2004	1,000	1,700	(4,000)	6,800		
 Rolland & Picoche 2008	3,357					
 Milton 2006	(400)	800-1,000	800-1,000	2,000		3,300
 Instituto Cervantes 2006	1,300	3,000	7,000	14,000	21,000	30,000
 Bergan 2001		850	1,500	4,500		

Cross-lingual pilot study

CEFRLex-wordlists for Dutch, French, Spanish, Swedish
(words assigned to first level of occurrence)



- Similar total number of words but **stark differences per level**.
- Comparable compilation methodology for all languages (graded text book corpus) does not lead to cross-lingual consistency.

Cross-lingual pilot study


Do translational equivalents have the same CEFR-label?

CEFR Companion	Dutch (NT2Lex)	French (FLELex)	Spanish (ELELex)	Swedish (SVALex)	English (EP)	English (Oxf.)
<i>"make and accept offers"</i>	accepteren aanvaarden	accepter	aceptar	acceptera anta	accept	accept
A2	B1	B2	A1	B2 / B1	B1	A2

Also no agreement in the Reference Level Descriptions
(bodyparts in [Italian](#) and [Spanish](#))

Profilo della lingua italiana

Home » Contenuti linguistici » Nozioni specifiche » Dominio pubblico » Parti del corpo (categoria aperta)

Parti del corpo (categoria aperta) 

Livello A1

Nessun descrittore per questo livello

Livello A2

- baffi
- barba
- bocca
- braccio
- corpo
- dente
- dito
- faccia
- gamba
- mano
- naso
- occhi
- orecchio
- piede
- testa

Centro Virtual Cervantes

Biblioteca del profesor > Plan curricular > Índice > 9. Nociones específicas. Inventario A1-A2

ENSEÑANZA

Plan curricular

Nociones específicas. Inventario A1-A2

1. Individuo: dimensión física

1.1. Partes del cuerpo [v. Nociones específicas 13.]	
A1	A2
<ul style="list-style-type: none"> pelo, ojo, nariz 	<ul style="list-style-type: none"> cabeza, cara, brazo, mano, dedo, pierna, pie oído, muela, garganta, estómago, espalda

Intermediate conclusions

- Dutch: difficult to link lexicographic data through CEFR word list
 - Overlap/coverage low, even at intermediate levels
 - Limited word sense disambiguation to link with lexicographic entries
 - Words from higher CEFR level within lexicographic components
- Cross-lingual linking thanks to a common framework?
 - Inconsistencies in vocabulary size per level between languages
 - Translational equivalents are not at the same level
 - Difficult to create new vocab resources based on existing ones
- Limited collaboration between communities
 - SLA/CALL/NLP seems separate cluster from e-lexicography/NLP/Linked Data
 - Integrating CEFR resources with other language resources requires expertise in (computer assisted) language learning and assessment, lexicography, natural language processing, data linking and artificial intelligence

Towards a linked CEFR infrastructure?

- Although CEFR aims to systematize L2 learning across languages, practical resources are mostly isolated both per language and relative to other learning resources (even other CEFR-resources)
 - RLDs \neq CEFR descriptors; Vocab lists \neq RLDs; Vocab between languages
 - Mostly textual descriptions of can-do statements or functions/notions
 - No explicit datamodel with standardized categories for creating structured data
 - No inventory of best practices for creating resources
- If CEFR is a language-independent framework, can it be used to cross-lingually link language-specific resources? Feasibility of
 - Linking to *language independent* descriptors
 - Linking between *language specific* Reference Level Descriptors
 - Linking between **grammar** and **vocabulary resources**
 - => **Building on existing expertise of linked lexicographic data?**

Why a linked CEFR infrastructure?

TEACHING, LEARNING & ASSESSMENT MATERIALS:

- combined availability of all relevant language resources for developers of learning and assessment applications
- comparable and portable across languages

RESOURCE CREATION:

- Bootstrapping new resource (for an under-resourced language) from existing resources (of higher-resourced languages)

RESEARCH PURPOSES:

- comparing systematicity of L2 learner profiling across languages
- ground truth data for generative AI testing/prompting

QUESTIONS:

- What are the actual needs for the communities around CEFR?
- Is linking of lexicographic data to CEFR interesting for the language learning/teaching/assessment communities?

Steps towards a CEFR infrastructure

LINKING OF RESOURCES:

- ELEXIS tools and data (dictionary matrix => CEFR matrix?)
- Semi-automatic translation (CEFR-J x 28 project, Tono et al. 2017)
- CALL-project collaborations (CEFRLex)

PARTNERS IN A PROJECT CONSORTIUM

- language teaching, learning and assessment community including CEFR-expertise
- lexicography (including constructions/idioms)
- linked data community (linking of multilingual resources)
- computational linguistics and artificial intelligence

FUNDING SCHEMES:

- ERASMUS+ / Horizon Europe / bilateral schemes / COST / CLARIN

Thank you!

Questions?

- *Is there a need for a linked data CEFR infrastructure?*
- *Is there an interest for a joint project around this?*

Other workshops planned

- [eLex 2023: *Lexicography and Cefr: Linking Lexicographic Resources and Language Proficiency Levels* \(29 June\)](#)
- [LDK 2023: *Linking Lexicographic and Language Learning Resources* \(13 Sep. / deadline 19 May\)](#)

Contact: kris.heylen@ivdnt.org ; ilan@lexicala.com ;
carole.tiberius@ivdnt.org; jelena.kallas@eki.ee