NT2Lex

A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet

Anaïs Tack [1,2] • **Thomas François** [1] • **Piet Desmet** [2] • **Cédrick Fairon** [1]

[1] CENTAL, Université catholique de Louvain, Louvain-la-Neuve, Belgium

[2] ITEC, imec, KU Leuven Kulak, Kortrijk, Belgium

CEFR-GRADED LEXICONS

a **graded lexicon** is a lexical database that includes lexical frequencies observed in texts graded along a difficulty scale

Foreign language (L2) materials

- textbooks and readers / learner texts
- CEFR scale [A1 > A2 > B1 > B2 > C1 > C2] (Council of Europe, 2001)

CEFRLex cental.uclouvain.be/cefrlex/

French - **FLELex** (François et al., 2014)

English - **EFLLex** (Dürlich & François, 2018)

Swedish - **SVALex** (François et al., 2016)

Swedish - **SweLLex** (Volodina et al., 2016)

NT2LEX — Resource

Corpus of reading materials

- corpus of 461,088 tokens
- 5 CEFR levels (A1, A2, B1, B2, C1)

Preprocessing

- part-of-speech tagging with **Frog** (van den Bosch et al., 2007)
- **SVM WSD** tool trained on DutchSemCor (Vossen et al., 2012)
- linkage to **Open Dutch WordNet** (Postma et al., 2016)

Lexical frequencies

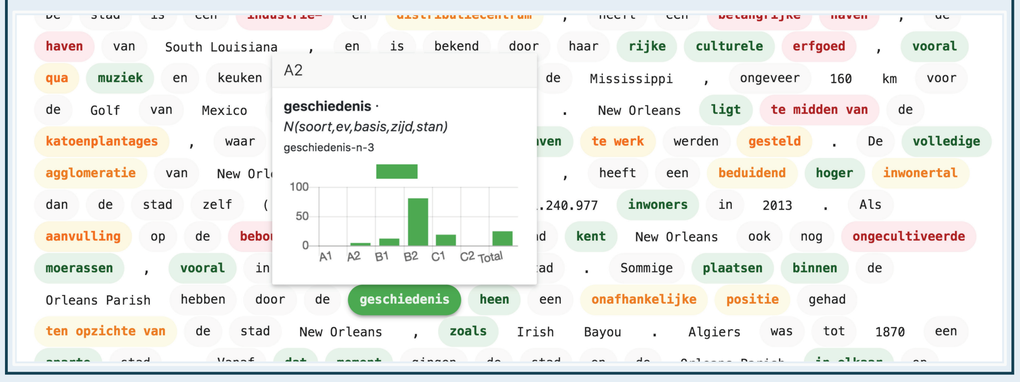
- lexical entries with per-level observed frequency
- normalised for lexical dispersion (Carroll et al., 1971)

lemma	pos	sense	synset	A1	A2	B1	B2	C1
pakken *to grab*	WW()	pakken-v-1	odwn-10-101230891-v	35	117	101	5	-
pakken *to defeat*	WW()	pakken-v-10	eng-30-01100145-v	-	51	12	-	-
zijn *to exist*	WW()	zijn-v-1	eng-30-02603699-v	2,094	1,647	1,423	1,253	1,335

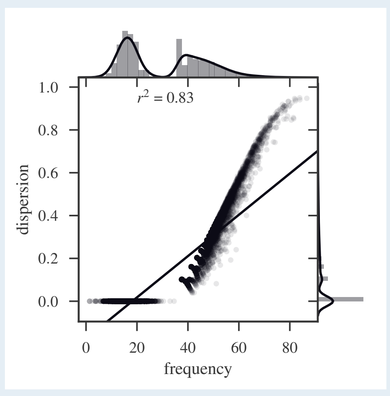
NT2LEX — Tools

Online tools for lexical complexity analysis

- database **search**
- CEFR-based **complex word identification** (Tack et al., 2016)



ANALYSIS — Frequency



frequency

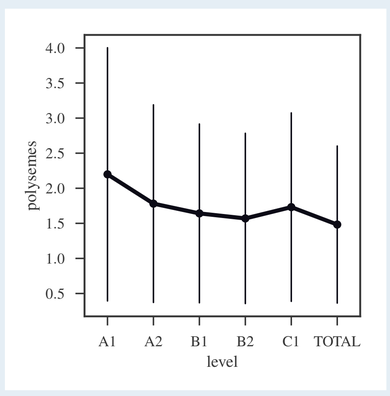
- correlation Subtlex-NL (Keuleers et al., 2010)
- Zipfian effects

  shorter = more frequent

dispersion

- theoretical familiarity
- more dispersed = basic voc

ANALYSIS — Semantics



semasiology

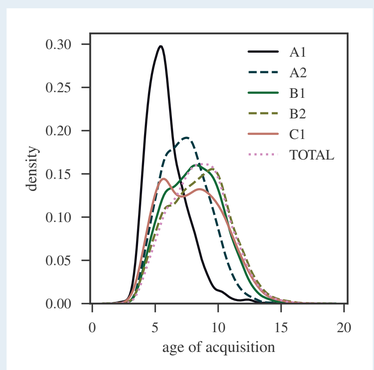
- form > meaning mappings
- easy = more polysemous

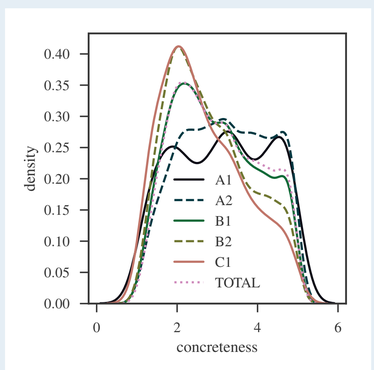
onomasiology

- meaning > form mappings
- lower degree of synonymy
- L2-specific lexicalisations

ANALYSIS — Psycholinguistics

interplay of **psycholinguistic norms** (Brysbaert et al., 2014)





KEY TAKEAWAYS

NT2Lex

- a new resource for **Dutch as a foreign language** (NT2)
- **17,743 entries** with graded frequency distributions
- measure of **receptive word difficulty**
- measure of **word sense complexity** through linkage to **Open Dutch WordNet**

cental.uclouvain.be/nt2lex/