DAFLex: a CEFR-graded lexical resource for German as a foreign language

Thomas François, Patricia Kerres,
Damien De Meyere, Ferran Suñer Muñoz

GR4L2 Workshop

December, 7th 2021

Plan

Plan

Vocabulary and L2 reading

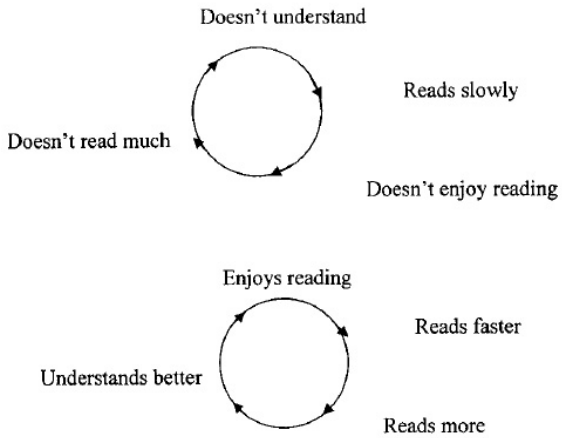
Vocabulary and text comprehension

- Vocabulary knowledge is crucial for L2 learning and a reader must know between 95% to 98% of the words in a text to adequately understand it [Hu and Nation, 2000, Laufer and Ravenhorst-Kalovski, 2010]
- In readability formulas, the lexical features have been shown to account the most for text's reading difficulty [Chall and Dale, 1995]

- Control of the level of vocabulary in a text is valuable to support reading comprehension

Vocabulary and L2 learning

- Extensive reading of understood texts can also have an impact on L2 learning and reading fluency [Grabe, 2014]
  $\rightarrow$ provided the readers do not enter into the vicious circle of reading, but instead the virtuous one [Coady, 1997].
- **Mean**: learners' should read texts with 1-5% of unknown words, but...:
  - How assess the lexical knowledge of a learner compared to the level of the text [Tack, 2021]
    $\rightarrow$ [Nation, 2006] defines a generic lexical coverage, but how apply that to a given text?
  - How decide which unknown words should be proposed to learners (zone of proximal development).



Objectives of the presentation

- Introduce a way to assess vocabulary complexity that combines CEFR levels and frequencies: the CEFRLex project
  $\rightarrow$ Can be used to address the above issues
- Describe the building of the German resource, DAFLex
- Demonstrate the dedicated interface to German

Plan

First approach: frequency lists

- 1st frequency list might be for German: *Häufigkeitswörterbuch der deutschen Sprache* [Kaeding, 1898]
- Various lists developed during 60-70s':
  [Pfeffer, 1964, Swenson, 1968, Rosengren, 1972]
- Modern computerized frequency lists appeared with Celex
  [Baayen et al., 1995]
  $\rightarrow$ corpus of 5.4 millions words (written) and 600 000 words (oral).
- the Frequency Dictionary German [Quasthoff et al., 2011]
- *dlexDB* project [Heister et al., 2011], based on 100 million words
- SUBTLEX-D [Brysbaert et al., 2011] is based on 25.4 million words from 4,610 films and TV series.

These lists are not specialized for L2 learners.

First approach: frequency lists

- [Tschirner et al., 2006] aims to define a core vocabulary of 5009 most frequent German words for learners (beginner and intermediate levels).
  $\rightarrow$ the Frequency Dictionary of German
  - Based on the TagAnt Tagger - that uses the TreeTagger tagset.
  - Names and separable verbs have been manually corrected (eg. ausmachen)!
  - They report frequency and dispersion

Frequency lists: limitations

These lists were defined from frequencies in the general language.

- Several issues are inherent to this approach:
  - Frequency estimation is not always robust ([Thorndike, 1921] : second half of the list less robust)
  - [Michéa, 1953] highlighted that some common words in language (available words) are not well estimated.
  - Not obvious how to transform frequencies into educational levels.

Frequency lists are not really educationally-graded resources!

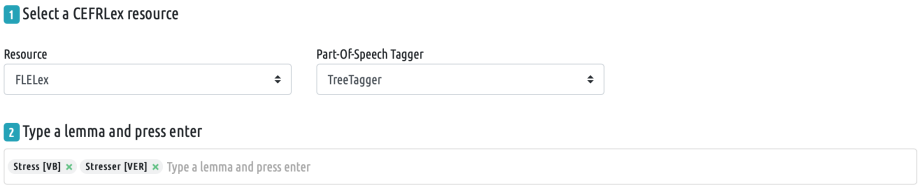
Second approach: the RLDs

Reference Level Descriptors are pedagogical references linked to the CECR

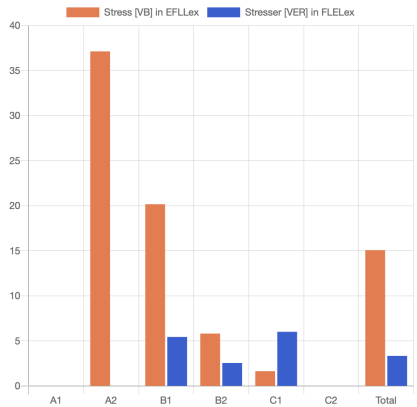
- Current references for L2 learning are the CEFR reference level descriptions:
  - French: [Beacco et al., 2004]
  - English: English Vocabulary profile [Capel, 2010, Capel, 2012]
  - German: Profile Deutsch [Glaboniat et al., 2005]
    $\rightarrow$ list of words to be learned, divided in 15 topics.
- Precise the CEFR about the specific lexical skills to learn, but...
  - No distinctions are made between words within a level
  - The format is not suitable for NLP approaches
  - Concerns has been raised as regards the validity of these referentials [Hulstijn, 2007]
  - No C1 and C2 levels for German

Plan

Objectives of the CEFRLex project



- To offer lexical resources describing word distributions in textbooks across the 6 CEFR levels.

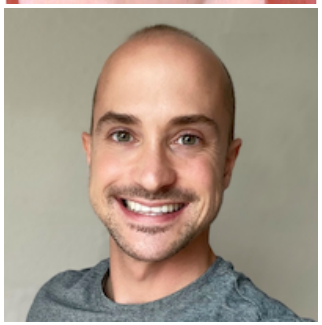


  - Possible uses :
    - Targeted vocabulary learning (which word to learn at which level)
    - Comparing the frequency of usage of synonyms
    - Using it within a language model for various iCALL tasks (readability, etc.)
    - Apply it for automatic text simplification (ATS)

The DAFLex team

DAFLex

- Available at `https://cental.uclouvain.be/cefrlex/daflex`
- Publication: to come
- Team: Thomas François, Patricia Kerres, Damien De Meyere, Ferran Suñer Muñoz
- Resource oriented towards reading (reception)



Builiding of DAFLex: methodology

1. Collect a corpus of texts intended for L2 learners (from textbooks)
   $\longrightarrow$ The texts must be labelled with a CEFR level
2. Find the lemma and the part-of-speech tag of each word in the corpus
   $\longrightarrow$ Issue : what is a word? MWE!
3. Estimate the frequency distribution of each lemma using a robust estimator
   $\longrightarrow$ dispersion index [Carroll et al., 1971] to normalize frequencies
4. Iterative process: manual postprocessing of the resource to correct NLP errors precedes a new frequency estimation step

DAFLex: the corpus

Genre	A1	A2	B1	B2
Dialogue	104 (10,113)	42 (8,566)	22 (5,093)	13 (4,092)
E-mail, mail	61 (7,112)	38 (6,364)	46 (6,563)	43 (8,349)
Sentences	306 (22,724)	176 (16,090)	291 (31,193)	323 (46,015)
Informative	124 (10,324)	121 (17,295)	146 (17,767)	208 (33,721)
Narrative	154 (21,814)	130 (34,973)	263 (66,497)	304 (74,722)
Varias	89 (9,483)	119 (15,841)	104 (15,151)	132 (22,439)
Total	838 (81,570)	626 (99,129)	872 (142,264)	1,023 (189,338)

Genre	C1	C2	Total
Dialogue	11 (4,381)	/	192 (32,245)
E-mail, mail	20 (5,338)	2 (803)	210 (34,529)
Sentences	199 (28,044)	221 (37,386)	1,516 (181,452)
Informative	109 (24,640)	5 (503)	713 (104,250)
Narrative	315 (104,781)	165 (81,086)	1,331 (383,873)
Varias	38 (6,293)	19 (4,268)	501 (73,475)
Total	692 (173,477)	412 (124,046)	4,463 (809,824)

Largest corpus of the CEFRLex project (slightly bigger than FLELex)

The tagging process

- **Goal:** Obtain the lemma of every form observed in the corpus and disambiguate homographic forms with different P.O.S.
  → Using inflecting forms would imply splitting frequency density across several forms.
  → It would also imply that we consider learners unable to relate inflected forms.

- **Problem:** The tagger precision matters, otherwise we can get:
  - Entries with wrong part-of-speech tag (e.g. *adoptez* PREP or *tu* ADV);
  - Entries with a non-attested lemma (e.g. *faire partir* instead of *faire partie*);
  - Likely tags that but are erroneous in the specific context of the word.

Selected taggers

We compared 4 taggers:

- TreeTagger [Schmid, 1994]
- spaCy (`de-core-news-lg` model)
- Stanford/Stanza [Qi et al., 2020]
- Freeling [Padró and Stanilovsky, 2012]

Tagger evaluation

Performance of the 4 taggers is not known on data for learners:

Methodology to compare the taggers

- Test set = sample of 100 sentences from the corpus
- Each word has been assessed by two experts, for each tagger.
- Error annotation schemes :

  0 no error;
  1 lemma is correct, but not the POS;
  2 POS is correct, but not the lemma;
  3 error for both the lemma and the POS;
  4 segmentation error

- $0.10 \leq \kappa \leq 0.39$: annotation was followed by a discussion step to produce a gold version

Results

Results per type of error:

Catégorie	TreeTagger	spaCy	Stanford
(0) Correct	85.6%	85.1%	90.7%
(1) POS	5.4%	5.2%	3.5%
(2) Lemme	8.2%	7.8%	4.39%
(3) Lemme + POS	0.45%	0.4%	0.7%
(4) Segmentation	0.25%	1.5%	0.7%

Freeling not assessed yet, as it appears less promising for our purpose

Best tagger appears to be Stanford!

Specific issues with German taggers

Stanford

- Tagset is very generic, maybe too much for us
  → Doesn't fit perfectly in all POS-categories of German
- Various issues: Segmentation problems for compound words; some confusions between "Adverb" and "Adjective"

Freeling

- Very rich tagset (making it complex to evaluate manually)
- Various issues: Segmentation problems for compound words; lemmatizes "Noun" without upper case

Specific issues with German taggers

spaCy

- Tagset seems adequate for our purpose (has verb prefixes)
- Various issues: Wrong lemmatisation of NN and NE; some flexionnal morphemes sometimes remain after lemmatization
- Verbs “haben”, “sein”, “werden” are always classified as auxiliaries

TreeTagger

- Developed in Stuttgart, very relevant tagset for German
- Includes verb prefixes and relative pronouns as category
- Verbs “haben”, “sein”, “werden” are always classified as auxiliary

Currently, TreeTagger has been selected based on the quantitative and qualitative analysis.

Computing the distributions

- We used the dispersion index [Carroll et al., 1971]

$$D_{w,K} = [\log(\sum p_i) - \frac{\sum p_i \log(p_i)}{\sum p_i}] / \log(I) \quad (1)$$

K = CEFR level ; I = number of textbooks in level K;
p_i = word probability in textbook i.

- Then, raw frequencies are normalized as follows:

$$U = (\frac{1\ 000\ 000}{N_k})[RFL * D + (1 - D) * f_{min}] \quad (2)$$

where N_k = number of tokens at level k;
$f_{min} = \frac{1}{N} \sum f_i s_i$ with f_i = word frequency in textbook i and s_i = number of words in textbook i

The DAFLex resource

ID card

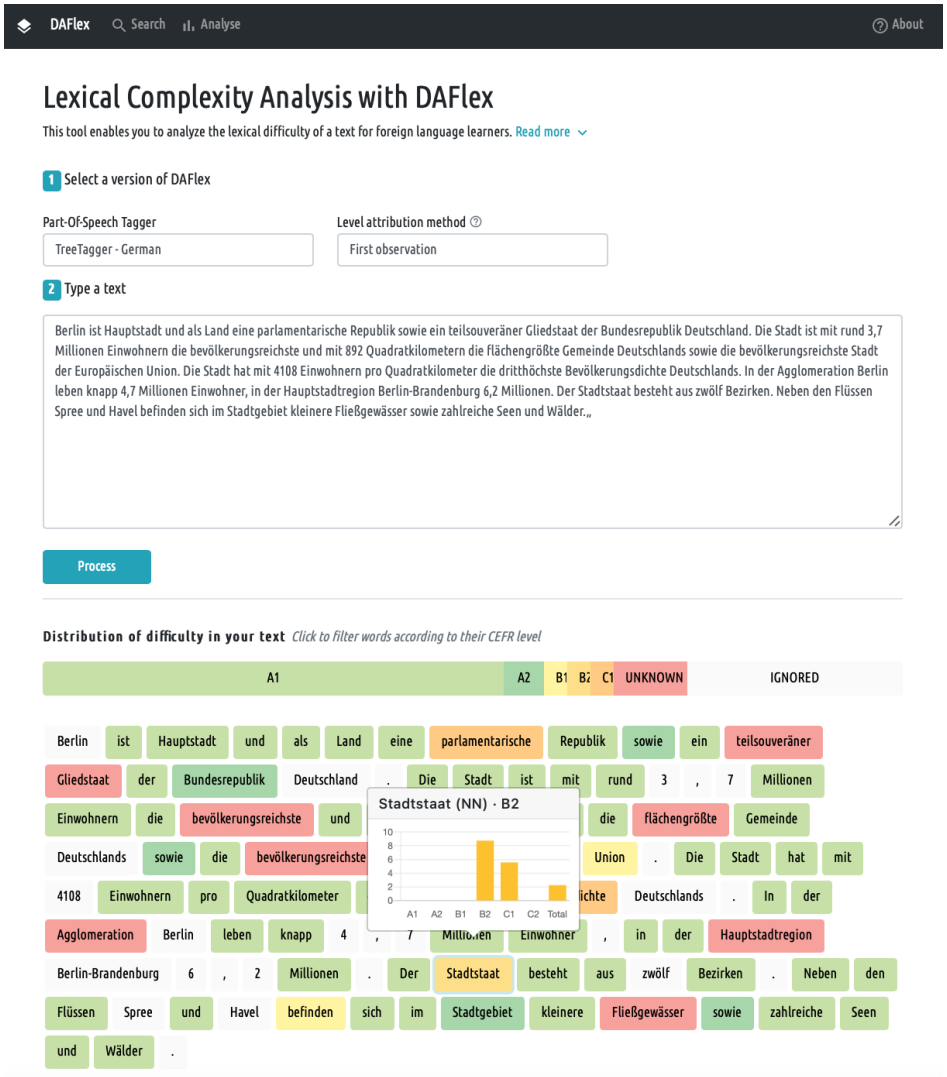
- DAFLex currently includes 41,646 entries, i.e. a pair of a lemma and a POS.
- It is based on the TreeTagger [Schmid, 1994] and is therefore easy to use within NLP applications
  $\longrightarrow$ Not able to detect split MWE (but rare in German) nor to reunite verbs and particules (e.g. rund ... ab for abrunden).
- The resource still needs to be manually checked.

Some entries from DAFlex

Lemma	Tag	A1	A2	B1	B2	C1	C2	Total
Sonne ('sun')	NN	411.5	84.87	81.63	70.5	83.6	72.23	111.72
Abendessen ('supper')	NN	127.89	64.19	30.23	41.05	1.55	56.01	46.2
Abholzung ('logging')	NN	0	0	0	0	5.56	0	0.75
beliebt ('popular')	ADJA	19.32	40.34	74.28	94.56	61.87	64.99	68.03
beliebt ('popular')	ADJD	6.86	33.06	22.42	20.47	7.77	9.84	20.71
wollen ('want')	VM	1676	2948	2328	1878	1772	1287	1942
erforschen ('seek')	V	0	0	7.88	4.45	44.8	64.3	17.6
absehen ('cheat')	V	0	0	0	9.83	27.08	12.42	7.98
vorsehen ('foresee'):	V	0	0	0	2.186	2.77	36.11	2.87

- "vorsehen" and "absehen": only the infinitive forms have been captured!
- It also shows that smaller corpus component produces larger frequencies.

Demo (https://cental.uclouvain.be/cefrlex/daflex/analyse/)



A few figures about the resources

Detailed figures for DAFLex per level (+ comparisons)

Level	# entries	# new entries	Hapax	EVP	FLELex	SVALex
A1	5,157	5,157	2,498	601	4,976	1,157
A2	7,821	4,973	4,056	925	3,516	2,432
B1	11,789	6,840	6,533	1,429	4,970	4,332
B2	17,024	9,663	9,687	1,711	1,653	4,553
C1	17,646	8,511	10,000	N/A	2,122	3,160
C2	15,699	6,526	9,319	N/A	N/A	/

Lot of new words at advanced levels (compared to French): due to the high compositionality of German!

Plan

Perspectives

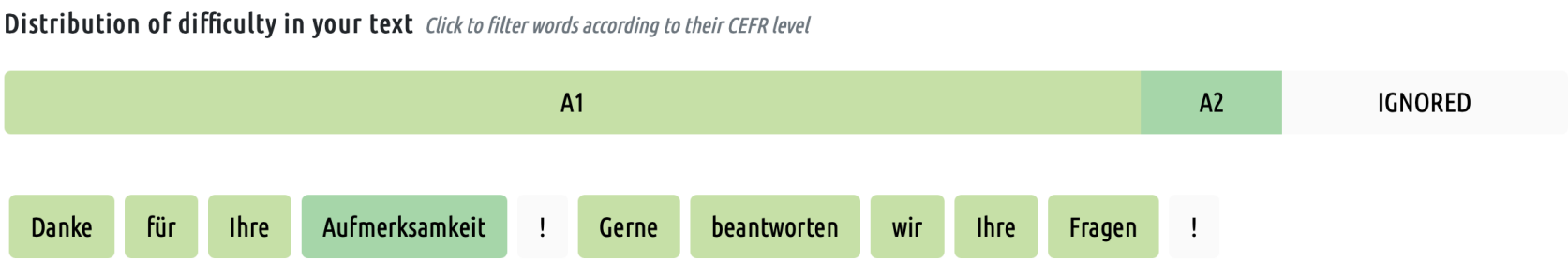
- Various uses of DAFLex can be conceived:
  - helping the teacher, standardization of textbook materials, etc.
  - automatic generation of lexicon-based exercises [Graën et al., 2020]
- Applications to automatic language difficulty assessment:
  - automatic prediction of complex words for learners [Tack et al., 2016]
  - used as features within a readability model [Yancey et al., 2021])
- Develop a disambiguated version of DAFLex [Tack et al., 2018]
- Process the particle verbs to merge them for all tenses

Conclusion

- The CEFRLex project (and DAFLex) proposes a frequency map of the use of lemmas across the six levels of the CEFR scale;
- DAFLex is freely available through a web site and will be available for download once the manual correction has been completed.

- Major issue: how to extract a core vocabulary from the DAFLex distributions?

Thank you for your attention!



References I

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995).
The celex lexical database (release 2).
Distributed by the Linguistic Data Consortium, University of Pennsylvania.

Beacco, J.-C., Bouquet, S., and Porquier, R. (2004).
Niveau B2 pour le français : un référentiel : utilisateur-apprenant indépendant.
Didier, Paris.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A., Bölte, J., and Böhl, A. (2011).
The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in german.
Experimental psychology.

Capel, A. (2010).
A1-b2 vocabulary: Insights and issues arising from the english profile wordlists project.
English Profile Journal, 1(1):1–11.

References II


Capel, A. (2012).
Completing the english vocabulary profile: C1 and c2 vocabulary.
English Profile Journal, 3:1–14.


Carroll, J., Davies, P., and Richman, B. (1971).
The American Heritage word frequency book.
Houghton Mifflin Boston.


Chall, J. and Dale, E. (1995).
Readability Revisited: The New Dale-Chall Readability Formula.
Brookline Books, Cambridge.


Coady, J. (1997).
L2 vocabulary acquisition through extensive reading.
In Coady, J. and Huckin, T., editors, *Second language vocabulary acquisition*, pages 225–237. Cambridge University Press, Cambridge.

References III


Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., and Wertenschlag, L. (2005).
Profile deutsch. gemeinsamer europäischer referenzrahmen; lernzielbestimmungen, kannbeschreibung, kommunikative mittel. niveau a1, a2, b1, b2.


Grabe, W. (2014).
Key issues in l2 reading development.
In *Proceedings of the 4th CELC Symposium for English Language Teachers-Selected Papers*, pages 8–18.


Graën, J., Alfter, D., and Schneider, G. (2020).
Using multilingual resources to evaluate cefrlex for learner applications.
In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 346–355.


Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., and Kliegl, R. (2011).
dlexdb–eine lexikalische datenbank für die psychologische und linguistische forschung.
Psychologische Rundschau.

References IV


Hu, M. and Nation, P. (2000).
Unknown vocabulary density and reading comprehension.
Reading in a foreign language, 13(1):403–30.


Hulstijn, J. (2007).
The shaky ground beneath the cefr: Quantitative and qualitative dimensions of language proficiency.
The Modern Language Journal, 91(4):663–667.


Kaeding, F. (1898).
Häufigkeitswörterbuch der deutschen Sprache.
Self-Published, Steglitz.


Laufer, B. and Ravenhorst-Kalovski, G. (2010).
Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension.
Reading in a foreign language, 22(1):15–30.

References V


Michéa, R. (1953).
Mots fréquents et mots disponibles. un aspect nouveau de la statistique du langage.
Les langues modernes, 47(4):338–344.


Nation, I. (2006).
How large a vocabulary is needed for reading and listening?
Canadian Modern Language Review, 63(1):59–82.


Padró, L. and Stanilovsky, E. (2012).
Freeling 3.0: Towards wider multilinguality.
In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.


Pfeffer, J. A. (1964).
Basic (spoken) German word list: Grundstufe.
Prentice Hall.

References VI

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020).
Stanza: A Python natural language processing toolkit for many human languages.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Quasthoff, U., Fiedler, S., and Hallsteinsdóttir, E. (2011).
Frequency Dictionary German/Häufigkeitswörterbuch Deutsch.
Leipziger Universitätsverlag, Leipzig.

Rosengren, I. (1972).
Ein frequenzwörterbuch der deutschen zeitungssprache: Lunder germanistische forschungen: Vol. 41.

Schmid, H. (1994).
Probabilistic part-of-speech tagging using decision trees.
In *Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.

References VII


Swenson, R. N. (1968).
A frequency count of contemporary german vocabulary based on three current leading newspapers.


Tack, A. (2021).
Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers.
PhD thesis, UCLouvain and KULeuven.
Thesis Supervisors : Piet Desmet, Cédrick Fairon and Thomas François.


Tack, A., François, T., Desmet, P., and Fairon, C. (2018).
NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet.
In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL 2018)*.

References VIII


Tack, A., François, T., Ligozat, A.-L., and Fairon, C. (2016).
Evaluating lexical simplification and vocabulary knowledge for learners of french: possibilities of using the flelex resource.
In *Proceedings of the Tenth conference on International Language Resources and Evaluation (LREC'16)*, pages 230–236.


Thorndike, E. (1921).
Word knowledge in the elementary school.
The Teachers College Record, 22(4):334–370.


Tschirner, E., Möhring, J., and Muntschick, E. (2006).
A frequency dictionary of German: Core vocabulary for learners.
Routledge.


Yancey, K., Pintard, A., and François, T. (2021).
Investigating readability of French as a foreign language with deep learning and cognitive and pedagogical features.
Lingue e Linguaggio, 46(2):229–258.