A Dictionary-Based Study of Word Sense Difficulty

David Alfter, Rémi Cardon and Thomas François
CENTAL
Université catholique de Louvain, Belgium
{first.last}@uclouvain.be

Abstract

In this article, we present an exploratory study on perceived word sense difficulty by native and non-native speakers of French. We use a graded lexicon in conjunction with the French Wiktionary to generate tasks in bundles of four items. Annotators manually rate the difficulty of the word senses based on their usage in a sentence by selecting the easiest and the most difficult word sense out of four. Our results show that the native and non-native speakers largely agree when it comes to the difficulty of words. Further, the rankings derived from the manual annotation broadly follow the levels of the words in the graded resource, although these levels were not overtly available to annotators. Using clustering, we investigate whether there is a link between the complexity of a definition and the difficulty of the associated word sense. However, results were inconclusive. The annotated data set will be made available for research purposes.

Keywords: word difficulty, readability, lexicography

1. Introduction

Dictionaries are used by native speakers of a language and by language learners when they want to learn or check the meaning of a word. Even though there are dictionaries that are specifically targeted at learners, other widely known online dictionaries such as Wiktionary[1] may be one of the first resources that comes to mind. Along with definitions, those resources provide the users with examples of use for words.

In this article, we want to check how useful these examples are in helping users with understanding words. The hypothesis for this research question is that non-native speakers can assess the difficulty of the meaning of a word when they see it used in a sentence, in the same way readers can infer the meaning of words based on the context (Miller et al., 1996; Miller, 1999). We observe how word difficulty is rated both by native and non-native speakers. We also assess whether a single word sense is rated differently based on the example of usage.

We also want to check whether dictionary definitions are more difficult when the word meanings themselves are more difficult. On the terminological side, it should be noted that *complexity* can be seen as an inherent property of a word or text, invariant and independent of the context (Pallotti, 2015), whereas *difficulty* can be seen as a construct that arises when a given reader interacts with a given word or text; *difficulty* varies from reader to reader. However, the distinction between these two categories is not always very clear, even in the literature on the topic.

In an effort to clarify the terminological question, we detail the operationalizations of the concepts used in this paper: (1) in the data annotation, we look at lexical *difficulty* to choose words, since our selection is based on proficiency levels from a learner-targeted vocabulary resource; (2) we also assess words in terms of lexical *difficulty*, since the ratings we collect are based on the annotators' intuition; (3) in the second experiment, the *complexity* of definitions is approximated through the lens of readability research, i.e. by characterizing the definitions using a number of linguistic variables.

The hypothesis here is that it is more difficult to explain difficult words, thus leading to a potentially more verbose explanation (that might nonetheless still be easy to understand); this is parallel to the idea that "languages encode conceptually more complex meanings with longer linguistic forms" (Lewis and Frank, 2016). In order to assess these hypotheses, we performed various experiments with resources in French.

For the first hypothesis, we asked native speakers of French and non-native speakers of French to rate word difficulty based on their use in dictionary examples. To do so, seven annotators rated word difficulty by only having access to dictionary examples. The annotated data will be made available for research purposes. We surmise that the data might be useful for the evaluation of complex word identification systems. One contribution of this part of the work is that the resource is annotated at the sense level, and not at the word level, which – to the best of our knowledge – is something that has not been done before for French.

For the second hypothesis, we explore the correlation between the ranking of the words difficulty produced by the annotators and the readability of the definitions.

In section 3, we describe the data that we used as a source for our linguistic material. Section 4 describes the experimental protocol that we put in place. Results are presented in section 5 and discussed in section 6.

2. Related Work

Word lists are often used in second language learning scenarios (e.g. Laufer and Nation (1999; O'Dell

[1] en.wiktionary.org/

et al. (2000; Meara (2002; Gu (2003; Nation (2013)). However, word lists compiled from L1 material are rarely suitable for L2 purposes (Richards (1974, p.72); François et al. (2014, p.3767)). There are some resources such as the CEFRLex family[2] that are based on L2 textbooks, thus directly targeting second language learners. However, even such resources generally use lemmas as primary entries, conflating different word senses. Especially from a language learning perspective, it is to be argued that not all word senses are learned at once, and thus basing vocabulary knowledge on lists where word senses are conflated is potentially misleading. Further, more frequent words (that are generally taught early, since frequency is often taken as a proxy for complexity, e.g. Rayner and Duffy (1986)) also tend to have more senses than less frequently used words (Crossley et al., 2010).

For English, there exists a dataset that is annotated both for lexical complexity and word senses, SeCoDa (Strohmaier et al., 2020), leveraging the Cambridge Advanced Learner's Dictionary[3].

However, even the shared tasks organized on the topic of Complex Word Identification (CWI; cf. (Paetzold and Specia, 2016; Yimam et al., 2018)) and Lexical Complexity Prediction (LCP; cf. (Shardlow et al., 2021)) do not explicitly distinguish between the complexity of different word senses; while some data sets do indeed present words in context, thus disambiguation of the words. That said, this information is not directly operationalized.

To the best of our knowledge, there is no work on using dictionaries to disambiguate vocabulary lists for French. Regarding the investigation of the complexity of definitions with regards to the complexity of their head words, we did not find any systematic study. We only found one article working on the complexity of dictionary definitions (Gross, 2018). However, the author states that a true semantic definition of a word (especially nouns) should not be conceptual but contain the whole set of appropriate predicates for this noun. This is not directly in line with our approach, as we work with conceptual definitions.

3. Data

3.1. Source

The data for our experiment comes from two different resources.

As we want to relate the outcome of the experiment to second language learning, we base ourselves on the French textbook-derived vocabulary list FLELex (François et al., 2014). FLELex lists words as well as their frequencies observed across different proficiency levels. In order to divide FLELex into six discrete levels, we use the machine learning based level assignment proposed by Pintard and François (2020) which is freely available through the CEFRLex webpage[4]. It contains 14,236 rated words.

As we work at the word sense level, we rely on a dictionary resource. The resource we use is GLAWI (Sajous and Hathout, 2015). GLAWI is an XML version of the French Wiktionary[5]. GLAWI's senses are strongly fine-grained : in the list we extracted, the average number of definitions per lemma is 13. We also calculated the median value, which is 2.

We filter GLAWI to extract the words that are found in FLELex. Every lemma, sense, definition or example that we mention throughout the article is extracted from the resulting subset.

3.2. Anchor Words

In order to "anchor" the relative difficulty rankings obtained with the methodology (cf. Section 4), we included "anchors", i.e. words that have a reliable fixed difficulty level. Anchor words were chosen among monosemous words that show a strong centrality for their respective level, i.e. words that are likely to be representative of a given level, based on the continuous numerical score N_c introduced in Gala et al. (2013):

$$N_c = N_i + e^{-r}, r = \frac{\Sigma_{k=1}^{i} U_k}{\Sigma_{k=i+1}^{N} U_k} \quad (1)$$

N_c is calculated for a given level N_i which corresponds to the level of first occurrence, i.e. the first level at which a word is observed with a frequency greater than 0, and modifies the level score by a score $e^{-r} \in [0, 1[$. r is calculated as the ratio of frequencies U_k, with U_k the frequency at level $k \in [1, N]$. In other words, r indicates the cumulative frequency up to level i divided by the remaining cumulative frequencies after level i. High values of e^{-r} indicate that there exists a non-negligible frequency mass outside of level N_i, while low values of e^{-r} indicate that the main frequency mass is located at level N_i.

For the selection of anchor words, we calculated N_c for all words in FLELex, excluded all words that did not fullfil the criterion $e^{-r} < 0.1$, and manually selected 5 words per level for a total of 30 anchor words.[6]

4. Experiments

4.1. Data Combination

For this experiment, we use a similar setup as in (Alfter et al., 2021), i.e. we arrange the example sentences into sets of four and ask annotators to select the easiest

[2] https://cental.uclouvain.be/cefrlex

[3] https://dictionary.cambridge.org/dictionary/english/

[4] https://cental.uclouvain.be/cefrlex/flelex/

[5] https://fr.wiktionary.org/

[6] While this methodology of level assignment differs from the methodology by (Pintard and François, 2020), it allows for a more fine-grained assessment of "centrality", and given that we work on a restricted subset of items where $e^{-r} < 0.1$, the items under scrutiny are of comparable quality with regards to automatic level assignment.

and the hardest of the words, a technique called best-word scaling (Louviere et al., 2015). A set of four items constitutes one *task* (see figure 1).
We use the combinatorial redundancy reducing algorithm (Alfter et al., 2021) for calculating the optimal number of tasks with minimal redundancy. This number, for 120 examples, comes to 1300. Each example is shown between 40 and 49 times in different combinations with other examples.
Care was taken to arrange the examples in such a manner that the four examples illustrating the same word but different senses end up as one task each.

4.2. Data Selection

For the experiment, we work at the definition (i.e. sense) level of words. We use example sentences from definitions as illustrations of a certain word sense.
In order to explore the different hypotheses, data was automatically selected according to the following criteria:

- 5 anchor words per level ($5 * 6 = 30$)
- 4 examples (= 1 task) per level that illustrate the same word but different senses ($4 * 6 = 24$)
- 4 examples (= 1 task) per level that illustrate the same word and same sense but with different examples ($4 * 6 = 24$)
- 3 paired examples per level, i.e. two examples of the same word but with different senses, chosen among words with at least two senses ($3 * 2 * 6 = 36$)
- 6 randomly chosen examples that have at least two senses

Thus, the total number of examples in the experiment is $30 + 24 + 24 + 36 + 6 = 120$. While this may seem like a relatively small number of items, we surmise that it is a sufficient amount for an exploratory work with an acceptable trade-off between quantity of items and annotation time.

4.3. Annotation

For the experimental design, we use a custom graphical user interface shown in Figure 1.[7] The user interface presents four sentences with one or more word(s) marked in bold and in purple, which is the word to be judged. After each sentence, we also display the lemma of the word. On the left side and right side of the examples are buttons to choose the easiest and hardest expressions. After selecting a word as being the easiest or hardest, the color of the lemma respectively changes to green and red in order to also reflect the choice visually.

[7]The interface shows a translated mockup. Note that the English Wiktionary indicates *years* (plural) as a lemma for the second example: `https://en.wiktionary.org/wiki/years`.

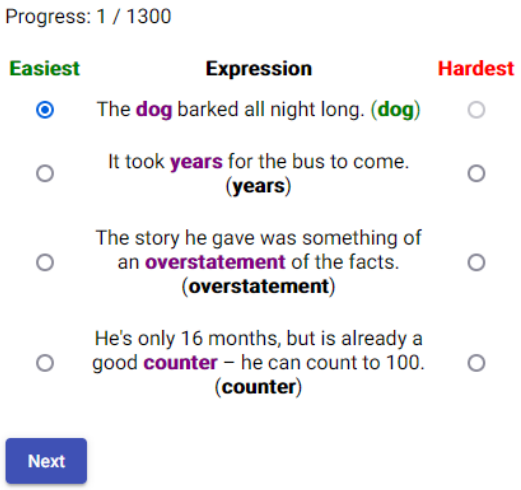


Figure 1: Graphical user interface with 'dog' being selected as easiest sense.

The user interface is designed to be simple and intuitive to use. It is possible to stop annotation at any time and resume later at the point where one left off. Further, the interface can be accessed from different devices such as laptops, computers, or smartphones, and one can freely switch between devices. The interface automatically registers the time elapsed between the completion of two tasks.
The user interface also attempts to enforce valid answers by disallowing clicking next if no choice has been made or if only one side contains a choice.
Internally, the example chosen as easiest is assigned a score of 1, the example chosen as hardest a score of 3 and the two examples not selected a score of 2. In the end, all votes v_{ij} for item i are aggregated into a single score s_i, with $i \in [1, 120]$ and $j \in [1, n]$, n being the number of votes for item i:

$$s_i = \frac{\Sigma_{j=1}^{n} v_{ij}}{n} \quad (2)$$

In order to see whether annotators are consistent in their annotation, we duplicated one task as control task. After shuffling of the data, the control task was inserted at positions 4 and 1299 (of 1300).
For this experiment, we recruited two student helpers who were paid 12€ per hour as well as three colleagues. In total, including two of the authors, seven people contributed to the experiment.
Each time they access the interface (unless requested otherwise, using an opt-out option in the form of a checkbox), users see a page displaying the instructions. Here is a translation (from French) of the detailed instructions displayed for the task (we leave out the instructions related to the interface):

> You will see sets with 4 sentences followed by the lemma of the word in bold. We ask you, for each set, to indicate which of the emphasized word's meanings seems to be

the most difficult to understand for you, and which one seems to be the easiest.

Don't think for too long, use your intuition.

Judge only the item in bold, with reference to its dictionary form (e.g., verbs in the subjunctive mood should be judged as equal to verbs in the indicative mood).

Context is given to indicate the sense of the word and should not have an influence on the judgment.

The annotators were orally asked to avoid discussing the tasks between themselves before completion.

Table 1 gives an overview over the demographic information of the participants.

Gender	
Male	4
Female	3
Mother tongue	
French	4
Japanese	1
Spanish	1
Luxembourgish	1

Table 1: Demographic information of participants

Each annotator was asked to complete all 1300 tasks.

4.4. Definition Complexity Evaluation

To assess definition complexity, we randomly selected 30,000 definitions from words that appear in FLELex. We processed them through FABRA, the French Aggregator-Based Readability Assessment Toolkit (Wilkens et al., 2022) and performed clustering on the data. FABRA calculates in excess of 4,000 features for each sentence. We restricted FABRA features to surface (e.g. word length) and lexical (e.g. frequency in different word lists) features, as the other classes of features (syntactic and discourse features) are more suitable to full text, and definition texts are not necessarily complete sentences. Before clustering, we perform dimensionality reduction (to 100 dimension) using Principal Component Analysis (PCA) (Pearson, 1901) as implemented in scikit-learn (Pedregosa et al., 2011). For clustering, we use the KMeans (Lloyd, 1957; MacQueen, 1967) implementation also available in scikit-learn, with the number of clusters set to 6 in order to match the number of CEFR levels, which are used by the FLELEX resource.

We then look at the cluster to which every definition linked to word senses that were annotated correspond, to check whether we can observe a correlation between the clusters and the difficulty level.

5. Results

5.1. Time per Annotator

Using a preliminary (randomly chosen) set of items, the authors tested the platform in order to estimate the time investment necessary. We found that it took about 12 seconds to complete one task. Based on this estimation and adding some margin (20 seconds per task), we estimated the total time needed to complete the experiment at around 7 hours.

Table 2 shows the average time taken for a single task, per annotator, in seconds. As the interface counts the number of seconds between the completion of two tasks, as long as annotators leave the interface open before continuing, time is being counted. By excluding outlier values – outliers being values of more than 90 seconds (an arbitrarily chosen threshold corresponding to an implausible time for a single task completion) – we obtain the average time per task as shown on the right side in Table 2 (Avg time excl. outl.), which is much closer to the originally predicted time per task.

Annotator	Avg time (s)	Avg time excl. outl.
1	15	12
2	101	9
3	12	9
4	45	18
5	39	16
6	78	8
7	11	11

Table 2: Average time per task per annotator

5.2. Rankings

In order to compute a ranking, we calculate the score of each item according to equation 2 in three different ways: taking into account all annotators, only native speakers, and only non-native speakers. This gives us three rankings: the global ranking, the native ranking, and the non-native ranking. Due to space limitations, the full results are not included here but can be retrieved at `https://github.com/daalft/dicomplex`.

Overall, the three rankings are very similar, the most dissimilar being the ranking between native and non-native speakers. However, even the most dissimilar rankings are highly correlated (Pearson's rank correlation coefficient of 0.90) as detailed in Section 5.3. A qualitative analysis reveals that there are mainly differences in rank for the words *haltérophilie* 'weight lifting', on rank 73 in the native speaker ranking and rank 115 in the non-native speaker ranking (a difference of 42 ranks), and *chèvrefeuille* 'honeysuckle', rank 78 vs 101 (difference of 23 ranks). These words seem to be relatively well known by native speakers, but introduced very late for non-native speakers. Two other notable differences can be found between *bricoleur* 'tin-

kerer', rank 28 in the native ranking versus rank 43 in the non-native ranking, and *clignotant* 'indicator/turn signal', rank 33 versus rank 49.

As regards the influence of context on the perceived difficulty of a word sense, we can see that the different examples of the same word sense end up rather close together on the ranking, with a maximum span (i.e. the difference between the maximum and minimum rank) of 19 ranks and an average span of about 14, as illustrated in Table 3. Furthermore, one can see that the words follow the progression of CEFR levels, except for *guérison* 'recovery' which ends up closer to B1 level yet shows a very narrow clustering.

Word	Level	Ranks	Span
connaître 'to know'	A1	13, 17, 23, 27	14
fixer 'to fix'	A2	31, 39, 40, 46	15
joindre 'to join'	B1	48, 52, 56, 67	19
prétention 'pretention'	B2	92, 96, 100, 105	13
guérison 'recovery'	C1	47, 49, 50, 54	7
attirail 'paraphernalia'	C2	95, 99, 107, 110	15

Table 3: Ranks of examples of the same word sense

The biggest rank span is found for *joindre* 'to join'. Upon closer inspection, we can see that the example sentence at rank 67 is *Ces planches, cette porte, ces fenêtres ne **joignent** pas bien.* 'These planks, this door, these windows do not **join** well', which is indeed a rather rare use of *joindre*.

We can see that the different senses of a word end up at quite different ranks, as illustrated in Table 4. The maximum observed span is 66 for 'point', with an average span of about 34, a significantly higher span than for examples of the same sense. An exception to the wider spread is the word 'old'. Upon closer inspection, we can see that the example sentences that were automatically selected were very short and thus did not convey the fine-grained meaning distinctions ('old' as pertaining to a certain age of a person, ancient, a derogatory term, a term of veneration). On the other hand, *point* 'point/dot/stitch pattern' ranged from *point* 'dot' (e.g., a *dot* ends a sentence) to *point* 'stitch pattern', and the meaning of stitch pattern was ranked as hardest of the senses by a large margin (rank 74, the closest rank of other meanings being rank 28). Again, one can also see that the ranking order follows the CEFR levels from FLELex in broad terms.

Table 5 shows the rank positions of all anchor words for the global ranking, the native speaker ranking and the non-native speaker ranking. As can be observed, there is a clear progression from A1 to C2, with expected overlaps between adjacent levels. Further, the rankings are quite similar, although the non-native ranking seems to follow FLELex levels a bit more closely, which is to be expected, since FLELex is a second language learner oriented resource.

Word	Level	Ranks	Span
vieux 'old'	A1	3, 4, 5, 6	3
point 'point/dot/stitch pattern'	A2	8, 25, 28, 74	66
repasser 'to iron/pass again/redo'	B1	30, 58, 65, 79	49
perte 'loss/ruin'	B2	32, 33, 53, 61	29
pétiller 'to fizz/sparkle/crackle'	C1	73, 101, 104, 111	38
fausser 'to falsify/forge/fake'	C2	76, 83, 85, 97	21

Table 4: Ranks of examples of different word senses

5.3. Intra- and Inter-Annotator Agreement

Based on the control task that was annotated twice by each annotator, we can see that five out of seven annotators were completely consistent in their annotation. For the remaining two annotators, one person chose a different "easiest" word while the other person chose a different "most difficult" word. As the control task was randomly chosen, there was no expectation regarding which word should be considered the easiest or the most difficult. Furthermore, the two annotators in question still remained consistent in their other choice, hence we neither discard the these annotators nor proceed in any kind of remediation.

Inter-annotator agreement (Pearson's rank correlation coefficient) shows a high agreement of 0.90 between the ranking of native speakers and the ranking of non-native speakers. This seems to confirm that non-native speakers can produce native-like rankings. This is consistent with what has been found in a similar study (cf. Alfter et al. (2021)).

5.4. Clustering

Figure 2a shows a visualization of the clustering using t-distributed stochastic neighbor embedding (t-SNE; (Van der Maaten and Hinton, 2008), a popular tech-

Global ranking		Native ranking		Non-native ranking	
CEFR level	Word	CEFR level	Word	CEFR level	Word
A1	bonjour	A1	bonjour	A1	bonjour
A1	copine	A1	copine	A1	copine
A2	vite	A2	vite	A1	confiture
A2	frigo	A2	frigo	A2	frigo
A1	confiture	A1	autocar	A2	vite
A1	vendeur	A1	vendeur	A1	vendeur
A1	autocar	B2	grille-pain	B1	apéro
A2	guitariste	A2	guitariste	A1	autocar
B1	apéro	A1	confiture	A2	guitariste
B2	grille-pain	B1	apéro	B2	grille-pain
A2	bricoleur	B2	revolver	B1	corridor
B2	revolver	A2	bricoleur	B1	vouvoyer
B1	festif	A2	clignotant	B1	festif
A2	clignotant	B1	festif	A2	bricoleur
B1	corridor	B1	corridor	A2	clignotant
B1	vouvoyer	B2	enquêter	B2	enquêter
B2	enquêter	B1	vouvoyer	B2	revolver
B1	vacarme	C2	haltérophilie	C1	arrachage
C1	arrachage	C2	chèvrefeuille	C1	discriminatoire
C2	chèvrefeuille	B1	vacarme	B1	vacarme
C1	discriminatoire	C1	surcoût	C2	chèvrefeuille
C1	surcoût	C1	discriminatoire	C1	surcoût
C2	haltérophilie	C1	arrachage	B2	lugubre
B2	perspicacité	B2	perspicacité	B2	perspicacité
B2	lugubre	B2	lugubre	C1	affligeant
C1	affligeant	C1	affligeant	C2	haltérophilie
C2	inexorable	C2	inexorable	C1	achoppement
C2	enhardir	C2	protéiforme	C2	inexorable
C1	achoppement	C2	enhardir	C2	enhardir
C2	protéiforme	C1	achoppement	C2	protéiforme

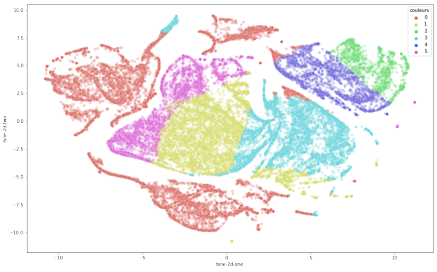
Table 5: Anchor words: comparison of ranking order positions for global, native and non-native rankings

nique for dimensionality reduction and visualization of high-dimensional data, and Figure 2b shows a visualization of the clustering using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP; McInnes et al. (2018)). UMAP is a fast and scalable dimensionality reduction algorithm that is said to be "better at preserving some aspects of global structure of the data than most implementations of t-SNE" (McInnes et al., 2018).

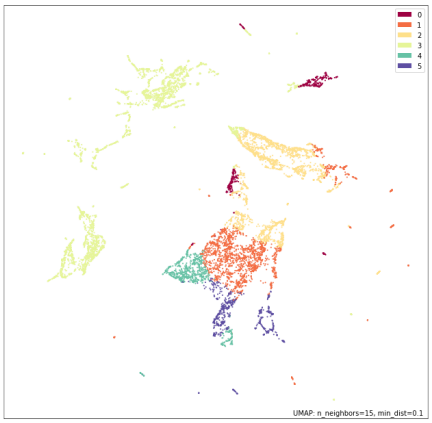
The 102 definitions corresponding to the 120 annotated examples are only found in three clusters out of six. Using the cluster numbers from the t-sne visualization, 48 were found in cluster 0 (red), 16 in cluster 2 (green) and 38 in cluster 4 (navy blue). The average scores by cluster, based on the global annotation ranking, are respectively 1.97 (that would be rank 59/120), 2.05 (that would be rank 65/120), and 2.01 (that would be rank 60/120). What we can draw from this observation is that no correlation between the difficulty of a word and the readability of its definition is found in the data we used for our experiments.

6. Discussion and Future Work

While we found the clustering not to correlate with the ranking, we still see a clear delineation of clusters in both visualizations. Further studies should investigate these clusters in more detail to find out what they represent and whether this information might be useful in future studies. From our observations we can get insights about the writing process of a dictionary. For example, the definition of the easy word *pêche* with the meaning of 'peach' is *Fruit du pêcher, parfumé et d'un goût savoureux, dont le très dur noyau est enrobé par une chair jaune ou blanche et une fine peau veloutée de teinte jaune et rouge-orange.* ("Fruit of the peach tree, fragrant and tasty, whose very hard pit is coated by a yellow or white flesh and a thin skin mixed with yellow and red-orange"), while the difficult word *perspicacité* (perceptiveness) is defined with *Pénétration d'esprit* ("Spirit penetration"). Peach is defined in a quite detailed way, while the definition of perceptiveness is abstract and vague. Those are two extreme examples, but it contributes to illustrate why we did not



(a) Visualization of clusters using t-SNE



(b) Visualization of clusters using UMAP

find correlations between the readability of definitions and the difficulty of word senses. We believe it would be beneficial to perform more studies on this very aspect, namely with comparing learners' dictionaries and more traditional dictionaries, so as to identify gaps between the phrasing of the definitions and the need of the targeted audience and systematically prevent them.
It would be beneficial to fine-tune the methodology of annotation; the current methodology covers all relations between examples. However, it is possible to drastically reduce the number of comparisons needed by inferring relations based on annotated relations. Thus, for example, – and for simplicity's sake with a simple comparison between two items – if one finds that A is easier than B and that B is easier than C, one could infer that A is easier than C. Thus, one would not need to annotate the relation between A and C. Preliminary experiments have shown that for a binary classification, i.e. choosing the "easiest" of two items, with 100 items, this would require about 700 comparisons between two items. In contrast, using the current methodology and adapting it to the case of binary classification, one would need 4950 comparisons.
It would also be interesting to further explore the differences between annotators, since the question of rater subjectivity has recently become a topic of interest in research on lexical complexity (Gooding et al., 2021; Shardlow, 2022).

7. Conclusion

In this article, we have presented a study on word sense complexity using a graded lexicon with CEFR levels for French and linguistic material (definitions and examples) extracted from the French Wiktionary. We asked seven annotators to rate the complexity of word senses based on their usage in a sentence. The resulting dataset will be made available upon publication. It consists 1,300 sets of four dictionary examples, along with the annotation of which one is the most difficult and which one is the easiest. Those 1,300 sets are found seven times, produced by four French native speakers and three non-native speakers. We have found that native speakers and non-native speakers agree to a quite large extent. However, the clustering was found not to correlate with rankings.
We compared the word senses ranking information to the corresponding definitions in order look for a correlation between a definition's readability and the difficulty of a word. We found no such correlation. Though, by examining closely the data we may argue that assessing the readability of definitions when writing a dictionary could improve its effectiveness.

Acknowledgments

We would like to thank the reviewers for their appreciation of the paper and their numerous detailed comments and suggestions. We also would like to thank our co-raters Nils Bouckaert, Alba Garcia Prades, Angela Kasparian, Hubert Naets and Nami Yamaguchi. David Alfter is supported by the Fonds de la Recherche Scientifique - FNRS under the grant MIS/PGY F.4518.21. Rémi Cardon is supported by the FSR Incoming Postdoc Fellowship program of the FSR - Université Catholique de Louvain.

8. Bibliographical References

Alfter, D., Tiedemann Lindström, T., and Volodina, E. (2021). Crowdsourcing Relative Rankings of Multi-Word Expressions: Experts versus Non-Experts. *Northern European Journal of Language Technology*.

Crossley, S., Salsbury, T., and McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, 60(3):573–605.

François, T., Gala, N., Watrin, P., and Fairon, C. (2014). FLELex: a graded lexical resource for French foreign learners. In *LREC*, pages 3766–3773.

Gala, N., François, T., and Fairon, C. (2013). Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *eLex-Electronic Lexicography*.

Gooding, S., Kochmar, E., Yimam, S. M., and Biemann, C. (2021). Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449.

Gross, G. (2018). Complexité lexicale: le substantif débat (s). *Neophilologica*, (30):9–24.

Gu, P. Y. (2003). Vocabulary learning in a second language: Person, task, context and strategies. *TESL-EJ*, 7(2):1–25.

Laufer, B. and Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language testing*, 16(1):33–51.

Lewis, M. L. and Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, 153:182–195.

Lloyd, S. (1957). Least square quantization in pcm. bell telephone laboratories paper. published in journal much later: Lloyd, sp: Least squares quantization in pcm. *IEEE Trans. Inform. Theor.(1957/1982)*, 18:11.

Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, 18(4):393–407.

Miller, G. A., Oakhill, J., and Garnham, A. (1996). Contextuality. *Mental models in cognitive science: Essays in honour of Phil Johnson-Laird*, pages 1–18.

Miller, G. A. (1999). On knowing a word. *Annual Review of Psychology*, 50(1):1–19.

Nation, P. (2013). *Learning Vocabulary in Another Language*. Cambridge University Press.

O'Dell, F., Read, J., McCarthy, M., et al. (2000). *Assessing vocabulary*. Cambridge university press.

Paetzold, G. and Specia, L. (2016). SemEval 2016 Task 11: Complex Word Identification. In *SemEval at NAACL-HLT*, pages 560–569.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of machine Learning research*, 12:2825–2830.

Pintard, A. and François, T. (2020). Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 85–92.

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Richards, J. C. (1974). Word lists: Problems and prospects. *RELC journal*, 5(2):69–84.

Shardlow, M., Evans, R., Paetzold, G. H., and Zampieri, M. (2021). SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online, August. Association for Computational Linguistics.

Shardlow, M. (2022). Agree to Disagree: Exploring Subjectivity in Lexical Complexity. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*.

Strohmaier, D., Gooding, S., Taslimipoor, S., and Kochmar, E. (2020). SeCoDa: Sense complexity dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5962–5967.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K., and François, T. (2022). FABRA: French Aggregator-Based Readability Assessment toolkit. In *Proceedings of the thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, United States. Association for Computational Linguistics.

9. Language Resource References

François, Thomas and Gala, Núria and Watrin, Patrick and Fairon, Cédrick. (2014). *FLELex: a graded lexical resource for French foreign learners*. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), ISLRN 742-240-876-017-1.

Sajous, F. and Hathout, N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, pages 405–426, Herstmonceux, England, august.