# NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet

**Anaïs Tack**[1,2,3a] **Thomas François**[1,3b] **Piet Desmet**[2] **Cédrick Fairon**[1]
[1] CENTAL, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[2] ITEC, imec, KU Leuven Campus Kulak, Kortrijk, Belgium
[3] F.R.S.-FNRS [a] Research Fellow, [b] Postdoctoral Researcher
{anais.tack,thomas.francois,cedrick.fairon}@uclouvain.be
{anais.tack,piet.desmet}@kuleuven.be

## Abstract

In this paper, we introduce NT2Lex, a novel lexical resource for Dutch as a foreign language (NT2) which includes frequency distributions of 17,743 words and expressions attested in expert-written textbook texts and readers graded along the scale of the Common European Framework of Reference (CEFR). In essence, the lexicon informs us about what kind of vocabulary should be understood when reading Dutch as a non-native reader at a particular proficiency level.

The main novelty of the resource with respect to the previously developed CEFR-graded lexicons concerns the introduction of corpus-based evidence for L2 word sense complexity through the linkage to Open Dutch WordNet (Postma et al., 2016). The resource thus contains, on top of the lemmatised and part-of-speech tagged lexical entries, a total of 11,999 unique word senses and 8,934 distinct synsets.

## 1 Introduction

In the recent years, a number of graded lexical resources have been developed to further research on first (L1) or second (L2) language complexity. Such a graded lexicon can be defined as a lexical database describing the graded frequency distributions of lexemes as they are attested in authentic pedagogical material along the successive grade levels of a particular language curriculum. The graded lexicons that have been built on these learning scales therefore either specifically pertain to the educational programme of (elementary) school children (Lété et al., 2004) or to the curriculum of foreign language learners (François et al., 2014).

As for the L2 language curriculum in particular, one of the most widespread learning scales which has been used to date is the Common European Framework of Reference for Languages (Council of Europe, 2001) or CEFR scale. The CEFR scale is a general framework that aims to provide a comprehensive description of the types of (written or spoken) discourse a learner at a particular proficiency level[1] should be able to understand or produce. Based on the CEFR scale and as part of the CEFRLex[2] project, a number of graded lexical resources have been developed for French (FLELex, François et al., 2014), Swedish (SVALex, François et al., 2016; SweLLex, Volodina et al., 2016) and English (EFLLex, Dürlich and François, 2018) as a foreign language. These lexicons were compiled from a corpus of L2 learning materials graded per level of the CEFR scale. The materials either include reading activities in textbooks or simplified readers (receptive graded lexicons; François et al., 2014, 2016; Dürlich and François, 2018) or texts written by learners (productive graded lexicons; Volodina et al., 2016). As a result, they inform us about what kind of vocabulary should be understood or produced when reading or writing in a foreign language at a particular proficiency level.

The lexical resources cited above have also found their purpose as components of NLP-driven educational applications. Up to date, we have seen some of the resources being integrated as features of a complex word identification system for French (Tack et al., 2016a,b), as components in a readability-driven learning platform for Swedish (Pilán et al., 2016a) or as part of an automated essay grading system for Swedish as well (Pilán et al., 2016b). It is therefore clear to say that the scope of relevance of the graded lexical resources goes well beyond their apparent usefulness to gain didactic insights into the complexity of the L2 curriculum.

[1] The CEFR scale includes six levels ranging from the elementary (A1/A2), to the intermediate (B1/B2) and advanced (C1/C2) levels. See Council of Europe (2001) for more details on the specific learning objectives per level.

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



*Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146
New Orleans, Louisiana, June 5, 2018. 

The principal aim of this paper is to augment the CEFRLex project by introducing a novel graded receptive lexicon for Dutch as a second or foreign language (*Nederlands tweede taal*, NT2), viz. the NT2Lex resource. Moreover, through the linkage of NT2Lex to Open Dutch WordNet (ODWN) (Postma et al., 2016), our additional objective is to expand upon and to advance the current methodology by introducing the first lexicon with graded frequency distributions for word senses.

The paper is structured as follows. The following section (Section 2) presents a bird's eye review of the literature on L2 receptive vocabulary and on the importance of measuring word sense complexity. In the subsequent sections, we will describe the revised methodology used to generate NT2Lex (Section 3) and we will compare the resource to the other CEFR-graded lexicons (Section 4). In the last section (Section 5), we will analyse the distribution of lexical entries in NT2Lex in light of standard indices of lexical complexity.

## 2 Background

The construct of receptive vocabulary knowledge has been an important factor when it comes to determining successful reading comprehension in a foreign language. We know that the input conveyed to foreign language learners through reading or listening should be sufficiently comprehensible not only for the message to be understood, but also for subsequent implicit or incidental acquisition to occur (Krashen, 1989). The notion of breadth of vocabulary knowledge (or vocabulary size) in particular plays an important role in predicting adequate comprehension of the L2 input. For reading comprehension, we know that 98% of the running words in the text should be known, which amounts to a vocabulary size of 8,000 word families (Laufer and Ravenhorst-Kalovski, 2010). However, the extent of vocabulary size is also heavily conditioned on the well-known variability in the interlanguage. It is therefore vital to obtain correct and relevant estimates of vocabulary knowledge when defining the lexical adequacy of a specific reading activity.

Various lexicon-based approaches have been considered to estimate the vocabulary knowledge that should be covered when learning or teaching a foreign language.[3] The first approach consists in measuring the vocabulary size based on frequency bands attested in academic word lists drawn from a reference corpus of the target language (Nation and Waring, 1997). A second approach resides in the use of L2-specific pedagogical vocabulary lists, which can be either expert-written such as the CEFR reference level descriptors (Marello, 2012; Milton, 2010) or corpus-based such as the English Vocabulary Profile (Capel, 2010, 2012). Finally, the CEFRLex project proposes a third approach to lexicon-driven evidence of vocabulary knowledge through the use of graded and corpus-based receptive lexicons (François et al., 2014).

An important aspect of vocabulary knowledge that has mostly been overlooked in the lexicon-based approaches concerns the distinction of word senses. Yet, the importance of taking into account form-meaning mappings has been well-evidenced in L2 reading comprehension. Qian (1999), for instance, hightlighted that in the interplay between vocabulary size and reading comprehension, the notion of depth of vocabulary knowledge also plays a significant role. The essential requirements for deep vocabulary knowledge include – besides the surface-level (i.e. spelling and phonetics) and morphological features – a thorough mastery of the various semantic, collocational, discursive and other contextual aspects of the word. Zooming in on the first two aspects in particular, he observed a significant addition of depth of vocabulary to explain the variability in comprehension scores.

The need to account for this semasiological variation when estimating word difficulty can be traced as far back as to Tharp (1939). Indeed, Tharp highlighted the drawbacks of defining word difficulty estimates by tallying the frequency of occurrence of similar word forms that are inherently polysemous. Subsequently, various studies have sought to parameterise the extent of semasiological and onomasiological variation in text-level readability assessment using polysemic, hypernymic and other features based on WordNet (Fellbaum, 1998), the most notable contribution of which relates to Coh-Metrix (Graesser et al., 2004). As for word-level readability assessment, a number of studies on lexical simplification have made advances in the ranking of the difficulty of synonyms based on contextual factors (Jauhar and Specia, 2012) or based on a lexical database of synonyms ranked according to elementary grade levels (Gala et al., 2013).

[3] For a more detailed overview on these lexicon-based approaches, we refer the reader to François et al. (2014).

| level | A1 | A2 | B1 | B2 | C1 | total |
|---|---|---|---|---|---|---|
| # tokens | 17,878 | 205,035 | 153,537 | 78,439 | 6,199 | 461,088 |
| # readers | 5 | 22 | 11 | 6 | 1 | 45 |
| # documents | 53 | 447 | 306 | 110 | 10 | 926 |

Table 1: Corpus statistics

## 3 Methodology

In view of the need for estimating word sense complexity in L2 learning addressed hereabove, we developed a graded lexical resource for learners of Dutch L2 which includes lexical entries linked to Open Dutch WordNet (ODWN) (Postma et al., 2016). The methodology for compiling a graded lexical resource can be found in Lété et al. (2004); François et al. (2014). Here, we will briefly summarise the method of estimating graded lexical entries, focusing on the particularities for Dutch.

### 3.1 Data

We used a 461,088-token corpus of CEFR-graded readers and textbooks for Dutch as a foreign language, ranging from the A1 to the C1 levels and with a mixture of writings in Netherlandic and Belgian Dutch (Table 1).

**Preprocessing** Typographical and language errors as well as other idiosyncracies observed in the OCR-ised texts were manually corrected. Tonic diacritics commonly used to indicate stress in written Dutch (e.g. *veel/véél*, 'many') were also manually removed, excluding the mandatory diaereses (e.g. *efficiënt*, 'efficient') and accents in loan words (e.g. *café*, 'pub'). All texts were lemmatised and part-of-speech tagged and multi-word units were automatically identified with the Frog tagger (van den Bosch et al., 2007).

The tagged texts were then fed to a word-sense disambiguation (WSD) tool[4]. The tool is based on a one-vs.-rest SVM classifier trained on the DutchSemCor (Vossen et al., 2012) and includes a dictionary of 92,617 lexemes and 117,225 senses, of which 52,430 (45%) seem to be matched to ODWN synsets. To increase the tool's coverage with ODWN, we also included all monosemous ODWN entries which were not included in the tool. In total, 76% of all distinct lexical units (adjectives, adverbs, nouns and verbs) were disambiguated for word senses.

[4]The tool was created by Rubén Izquierdo and is available on http://github.com/cltl/svm_wsd.

### 3.2 Definition of Lexical Entries

To define the list of lexical entries which make up the resource, we proceeded to some extra correction, simplification and filtering of the previously tagged and word-sense disambiguated texts.

**Lemmata** We first ruled out all non-alphanumeric entries such as punctuation marks, Arabic numerals, as well as non-standard word forms and abbreviations commonly found in Dutch chatspeak. We also simplified similar alphanumeric numbers (e.g. *4de, 5de*, '4th, 5th') as belonging to the same lexical entry *[digit]de*.

We then resolved some specificities of the Dutch compounding system. On the one hand, we decided to split a number of compounds with an optional parenthesised stem. For a lemma such as *(studie)keuze* ('(study) choice') for instance, we counted the occurrence as two separate lexemes: *keuze* ('choice') and *studiekeuze* ('study choice'). On the other hand, we also resolved the omission of shared stems in a number of coordinated compounds (e.g. *binnenland* vs. *binnen- en buitenland*, 'home and abroad').[5]

**Parts of speech** The Frog part-of-speech tagger is based on the CGN (*Corpus Gesproken Nederlands*) tagset (Van Eynde, 2004). The CGN tagset is quite extensive in that it counts over 320 tags and thus accounts for a number of detailed lexical and morphological features. However, we found it irrelevant to keep all of these precise features in the resource. We therefore decided to simplify the tagset to a set of 37 tags (Table 2). Consequently, all other special symbols not covered by the tagset were filtered from the resource.

We should note that the multi-word units detected by Frog tagger were not tagged with a specific part of speech, but with a "multi-tag" part of speech (e.g. *door en door*, VZ(fin)_VG(neven)_VZ(fin), 'through and through'). For all of these multi-word units, we also subsequently transposed each one of the individual tags according to our simplified tagset.

**Word senses** Finally, all lexical entries which were disambiguated for word senses were thus supplemented with a tuple of ODWN sense and

[5]To this end, automatic compound splitting was performed using the publicly available rule-based compound splitter for Dutch (http://ilps.science.uva.nl/resources/compound-splitter-nl/)

| simplified tag | *part of speech* | # 37 |
|---|---|---|
| N(soort/eigen) | *noun (common/proper)* | # 2 |
| ADJ() | *adjective* | # 1 |
| WW() | *verb* | # 1 |
| TW(hoofd/rang) | *numeral (card./ord.)* | # 2 |
| VNW(...) | *pronoun* | # 20 |
| LID(bep/onbep) | *article (def./indef.)* | # 2 |
| VZ(init/fin/versm) | *preposition (initial/final/fused)* | # 3 |
| VG(neven/onder) | *conjunction (coord./subord.)* | # 2 |
| BW() | *adverb* | # 1 |
| TSW() | *interjection* | # 1 |
| SPEC(deeleigen) | *part of proper noun* | # 1 |
| LET() | *punctuation* | # 1 |

Table 2: List of simplified CGN tags.

corresponding synset ids (e.g. *lezen-v-1*, eng-30-00625119-v, 'to read'). The synset ids include either the WordNet 3.0 offset (*eng-30*) or the ODWN 1.0 offset (*odwn-10*) otherwise. However, we should note that not all WordNet and ODWN synsets included in Open Dutch WordNet have a corresponding lexical entry. We therefore completed those entries with their corresponding sense number (e.g. *overduidelijk*, obvious.a.01) in NLTK's Open Multilingual WordNet (Bird et al., 2009). Finally, in the absence of an ODWN equivalent to DutchSemCor, we decided to keep the original sense id obtained through WSD for the sake of completeness and for future compatibility (e.g. *overbodig*, d_a-415574, 'superfluous').

### 3.3 Lexical Frequencies and Weighting

After having defined the set of lexical entries, we computed their graded frequency distributions across the five CEFR levels attested in the corpus. The following statistics were computed for each lexical entry and per each level.

**Raw frequency** The frequency $F_{\text{entry,level}}$ is simply computed as the number of times the entry occurs in the level, which amounts to summing up the vector $\mathbf{f}$ of the entry's frequencies of occurrence $f$ in document $i$ for all $d$ documents in that level (see Table 1 on the preceding page).

$$F_{\text{entry,level}} = \sum \mathbf{f} = \sum_{i=1}^{d} f_i \qquad (3.3.1)$$

**Dispersion and adjusted frequencies** The exclusive use of raw frequencies to observe lexical distributions has been subjected to much debate in corpus linguistics and especially when applied to mining corpora to futher L2 research. Indeed, Gries (2008) previously stated that the extent of written language proficiency in learner corpora appears to be closely linked to the scope of lexical dispersion: the more dispersed the use of a word, the better it is mastered.

For written language comprehension on the other hand, we could also state that the extent of lexical dispersion in readers and textbooks gives us a better view on what kind of vocabulary is subject to being well-understood at a particular proficiency level. As a consequence, the lexical frequencies we want to use to gain insights in non-native language comprehension should be adjusted to take into account lexical dispersion as well. The following dispersion ($D_{\text{entry,level}}$) and adjusted frequency indices ($U_{\text{entry,level}}$ and $SFI_{\text{entry,level}}$) were computed following Carroll et al. (1971). In the following formulae, $N_{\text{level}}$ denotes the number of words in the level and $n_i$ denotes the number of words in document $i$ of all documents $d$ in that level.

$$D = \left[ \ln(\sum \mathbf{f}) - \left( \frac{\sum_{i=1}^{d} f_i \cdot \ln(f_i)}{\sum \mathbf{f}} \right) \right] \cdot \frac{1}{\ln(d)} \qquad (3.3.2)$$

$$U = \frac{10^6}{N_{\text{level}}} \left[ F \cdot D + (1 - D) \cdot \left( \frac{1}{N_{\text{level}}} \sum_{i=1}^{d} f_i \cdot n_i \right) \right] \qquad (3.3.3)$$

$$SFI = 10 \cdot [\log_{10}(U) + 4] \qquad (3.3.4)$$

## 4 Resource Description

We compiled two separate versions of NT2Lex. A first version contains only the lemmatised and part-of-speech tagged entries (**NT2Lex-CGN**) and is thus similar to the other graded lexicons previously developed in the CEFRLex project (cf. *supra*). The word-sense disambiguated entries, on the other hand, have been added to a second version of the resource (**NT2Lex-CGN+ODWN**) (see Table 3 for an example). A comparative overview of the number of entries in both versions of NT2Lex and in the other resources can be found in Table 5 on page 6. A more detailed description of the resource is given hereinbelow and in the following section.

**NT2Lex-CGN** The first version of NT2lex counts 15,227 entries. The total number of entries in the resource is therefore similar to the resources developed for English and Swedish, although slightly lower than for French. Not surprisingly, the majority of the entries contain lexical

| *lemma* | pos | sense-id | synset-id | gloss | U@A1 | U@A2 | U@B1 | U@B2 | U@C1 | U@total |
|---|---|---|---|---|---|---|---|---|---|---|
| *in zwang* | VZ(init) N(soort) | in_zwang-n-1 | eng-30-14411884-n | 'in vogue' | - | - | - | - | 22 | 0 |
| *omgangstaal* | N(soort) | omgangstaal-n-1 | eng-30-07157123-n | 'vernacular' | - | - | - | 26 | - | 3 |
| *pakken* | WW() | pakken-v-1 | odwn-10-101230891-v | 'grab' | 35 | 117 | 101 | 5 | - | 99 |
| *pakken* | WW() | pakken-v-10 | eng-30-01100145-v | 'defeat' | - | 51 | 12 | - | - | 28 |
| *zijn* | VNW(bez,det) | - | - | 'his' | 3,349 | 7,900 | 4,124 | 3,479 | 4,308 | 5,798 |
| *zijn* | WW() | zijn-v-1 | eng-30-02603699-v | 'exist' | 2,094 | 1,647 | 1,423 | 1,253 | 1,335 | 1,601 |

Table 3: Example of entries in NT2Lex-CGN+ODWN with their graded adjusted frequencies $U$. A column with glosses was added for illustrative purposes.

words and the number of grammatical entries also remains strongly comparable across all resources.

There is a striking difference however in the number of multi-word entries that are included in NT2Lex and in the other resources. Only 459 of the entries are multi-word units, contrary to 2,038 for French and 1,450 for Swedish. The multi-word units that are included in the resource mostly pertain to well-known named entities (e.g. *Olympische Spelen*, 'Olympic Games') and other phrasal verbs (e.g. *voorzien van*, 'to provide'), adverbs (e.g. *om het even*, 'all the same'), etc.

This difference could be explained by the fact that the majority of the compound words which are multi-word units in other languages (such as in French or English) are one-word units or agglutinative compounds in Dutch (e.g. *afvalverwijderingsstructuur*, 'waste disposal structure'). We observe that 4,431 (31%) of the single-word entries in NT2Lex are in fact compounds. As for the Swedish language, where the compounding system is similar to Dutch, we could attribute this disparity to the fact that different taggers were used to detect multi-word units. Indeed, the recall and precision of multi-word identification depend heavily on the assumptions made by the tagger to resolve the sequential ambiguity, contrary to the agglutinative compounds, which do not need to be disambiguated in this case.

**NT2Lex-CGN+ODWN** The word-sense disambiguated version of NT2Lex counts 17,743 entries in all, with an extra 2,516 lexical entries and with 1,454 polysemous entries (with at least two senses). Table 4 shows the distribution of polysemous entries across all levels. Although all of these polysemous entries are lexical ones, we should note that some multi-word entries have also been disambiguated for word senses, but none of them are polysemous. The most polysemous entry in the resource is the entry *pakken* (verb, 'to take / grab / defeat / hinder / etc.') which has a total of 10 different senses attested in the resource.

| | A1 | A2 | B1 | B2 | C1 | total |
|---|---|---|---|---|---|---|
| # entries | 1,189 | 7,630 | 10,160 | 9,366 | 1,841 | 17,743 |
| # senses | 849 | 5,705 | 7,272 | 6,517 | 1,302 | 11,999 |
| # polysemes | 139 | 828 | 979 | 771 | 118 | 1,451 |
| # synsets | 658 | 4,450 | 5,465 | 4,936 | 1,046 | 8,934 |

Table 4: The number of word senses, polysemes (entries with >1 sense) and unique synsets in NT2Lex-CGN+ODWN

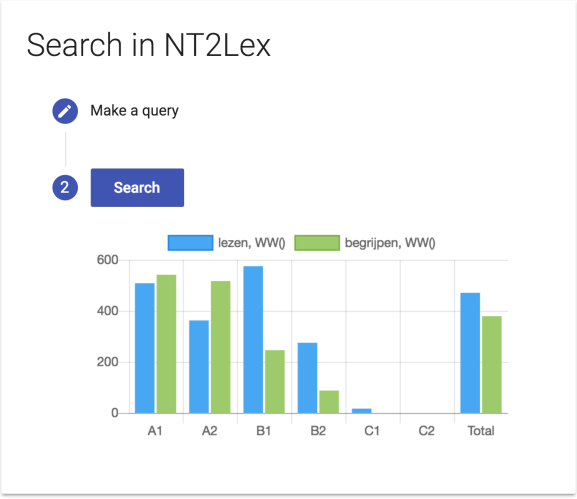



Figure 1: Screenshot of an online query in NT2Lex for the verbs *lezen* ('to read') and *begrijpen* ('to understand')

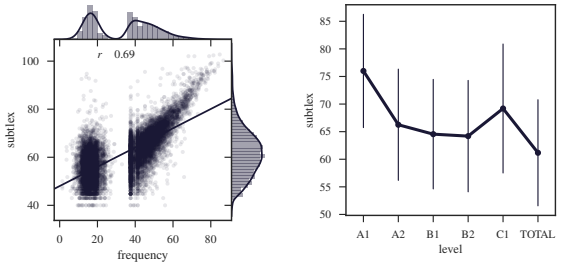
Table 4 also shows the number of unique concepts (# synsets) included per level and in total. We observe that the resource includes a high variety of concepts, with 8,934 distinct synsets out of 11,999 word senses.

**Online query and annotation tools** Both versions of the resource will be made available for non-commercial use in the CEFRLex project.[6] Similar to the previous resources, a number of online tools will be made available for teachers and/or researchers to query the lexical database and to annotate a text using NT2Lex (Figure 1).

[6] http://cental.uclouvain.be/nt2lex/

| *resource* | NT2Lex | | | | | | | | | FLELex | | SVALex | | EFLLex | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *version* | CGN | | | | | CGN+ODWN | | | | CRF | | / | | / | |
| **# entries** | **15,227** | | | | | **17,743** | | | | **17,871** | | **15,681** | | **15,281** | |
| lexical | 14,368 | | | | | 16,884 | | | | 17,404 | | 15,291 | | 14,857 | |
| grammat. | 400 | | | | | 400 | | | | 467 | | 390 | | 424 | |
| multi-w. | 459 | | | | | 459 | | | | 2,038 | | 1,450 | | 3,852 | |
| levels | # | new (%) | compound | hapax | >10 | # | new (%) | hapax | >10 | # | new (%) | # | new (%) | # | new (%) |
| A1 | 953 | 953 (1.00) | 70 | 313 | 225 | 1,189 | 1,189 (1.00) | 427 | 228 | 4,976 | 4,976 (1.00) | 1,157 | 1,157 (1.00) | 2,395 | 2,395 (1.00) |
| A2 | 6,220 | 5,383 (0.87) | 1,224 | 2,482 | 1,231 | 7,630 | 6,580 (0.86) | 3,073 | 1,386 | 6,995 | 3,516 (0.50) | 3,327 | 2,432 (0.73) | 4,205 | 2,478 (0.59) |
| B1 | 8,559 | 4,879 (0.57) | 1,997 | 3,936 | 1,081 | 10,160 | 5,571 (0.55) | 4,739 | 1,128 | 10,780 | 4,970 (0.46) | 6,554 | 4,332 (0.66) | 5,607 | 2,740 (0.49) |
| B2 | 8,172 | 3,641 (0.45) | 1,861 | 4,362 | 638 | 9,366 | 3,998 (0.43) | 5,092 | 619 | 7,349 | 1,653 (0.22) | 8,728 | 4,553 (0.52) | 8,228 | 3,935 (0.48) |
| C1 | 1,680 | 371 (0.22) | 252 | 1,127 | 63 | 1,841 | 405 (0.22) | 1,282 | 62 | 8,348 | 2,122 (0.25) | 7,564 | 3,160 (0.41) | 9,232 | 3,733 (0.40) |
| C2 | - | - - | - | - | - | - | - - | - | - | 7,433 | 634 (0.09) | - | - - | - | - - |

Table 5: A comparative overview of NT2Lex and the other lexicons in terms of the number (#) of entries per level (including new entries, compounds, hapaxes and entries with a frequency greater than 10), as well as the number of lexical (adjectives, adverbs, nouns and verbs), grammatical and multi-word entries.




(a) correlation between NT2Lex and Subtlex (b) median of Subtlex frequencies per level

Figure 2: Comparison of NT2Lex and Subtlex-NL standardised frequencies

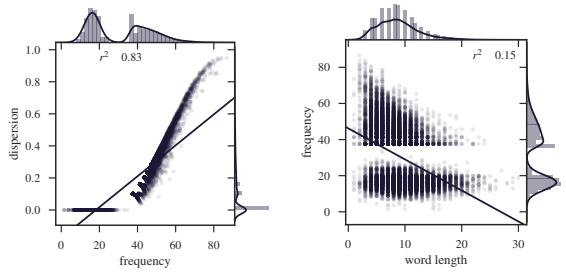



Figure 3: Zipfian effects for adjusted frequencies (*SFI*), dispersions (*D*) and word lengths for all entries in NT2Lex

## 5 Analysis

In the next sections, we will compare the distribution of lexical entries in NT2Lex in light of a number of standard indices of lexical complexity. We will only report statistics for the most complete version of our resource, i.e. NT2Lex-CGN+ODWN.

### 5.1 Frequency Effects

As a first means of analysis, we aim to examine the coherence of the frequency distributions in the resource with respect to the word frequency effect in second language processing, which states that words that are more frequent and more familiar are more easily processed by a learner (Ellis, 2002).

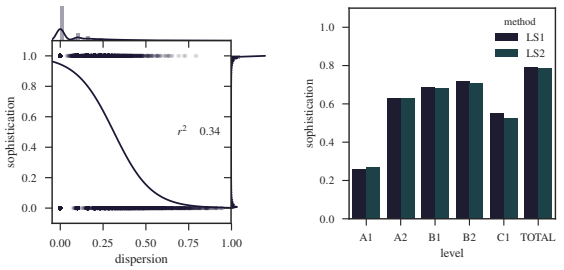
**Lexical frequency** To compare the frequency distributions in the resource, we use the standard frequency index (*SFI*, Formula 3.3.4), which might be best suited to measure the desired effects: a value of 100, 90, 80, ..., 40 on the standard scale indicates that the entry respectively occurs once every $10^0$, $10^1$, $10^2$, ..., $10^6$ entries, and so forth.

When comparing the standardised frequency distributions with Subtlex-NL (Keuleers et al., 2010), we observe a positive value for the Pearson correlation coefficient ($r = .69$, $p < .001$[7]; Figure 2a). This shows us that even though the adjusted frequencies were estimated on a relatively small corpus, they are still very much coherent with the frequencies estimated for the same entries on a reference corpus. Moreover, Figure 2b also illustrates that the average of Subtlex-NL frequencies also decreases per level, with the exception of the C1 level. A possible reason for this is that the C1 subcorpus is the most restricted in size due to the limited availability of C1-level readers.

**Dispersion and word familiarity** Because of lacking experimental data on actual word familiarity in Dutch L2, we make a simplifying assumption here and use our dispersion metric as a measure of *theoretical* word familiarity: the more the word is dispersed across the L2 documents, the more familiar it should be to a learner in general.

We observe from Figure 3 that this theoretical word familiarity accounts for about 83% in the distribution of the adjusted frequencies. In this respect, we also observe an interesting split in

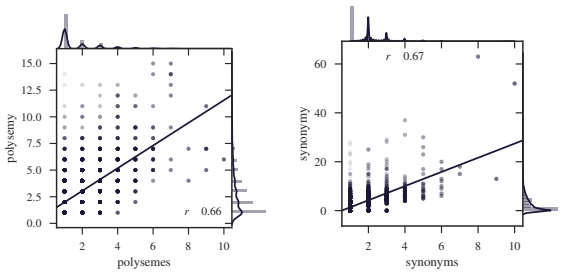
[7]For reasons of comparability with Subtlex, which does not include frequencies for word senses, we report the correlation coefficient for the non-WSD version of the resource.




(a) degree of lexical dispersion and lexical sophistication

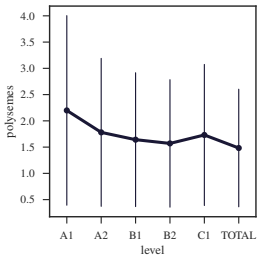
(b) lexical sophistication per level based on Lu (2012)'s ratios

Figure 4: The interplay between lexical dispersion and lexical sophistication in NT2Lex




(a) number of polysemes per entry in function of ODWN polysemy

(b) number of synonyms per entry in function of ODWN synonymy




(c) average polysemes / level

Figure 5: Polysemy and synonymy in NT2Lex

the normalised frequency distribution which originates from the way the adjusted frequencies were computed ($U$; Formula 3.3.3). When $D > 0$, the influence of the raw frequency $F$ increases between $40 < SFI < 100$. Conversely, when $D = 0$, the raw frequencies are not taken into account, but are drawn from a weigthed (Gaussian) frequency distribution ($0 < SFI < 40$) instead.

**Zipfian effects** From these associations between dispersion and standardised frequencies, we can also observe a number of Zipfian effects. Indeed, for all entries which have a non-zero dispersion (range $40 < SFI < 100$), Zipf's distribution (Zipf, 1949) applies in the standard frequency index. Moreover, these frequencies are in turn negatively correlated with word length: the shorter the word, the more frequent ($r = -.39$; $p < .001$). We take these results as a proof for the consistency of the resource.

**Lexical sophistication** As a final note on the issue of word frequency, we compare the lexical sophistication ratio of the entries with a basic word list[8] of the 2,000 most frequent Dutch words according to the *Basiswoordenboek Nederlands* (Kleijn and Nieuwberg, 1993).[9] Figure 4a shows that the more dispersed (and hence the more familiar) the entry in the corpus, the least sophisticated the entry is. Moreover, we also observe that the proportion of sophisticated entries (i.e. that go beyond the 2,000 most frequent words) increases per level (Figure 4b), except for the C1 level where fewer sophisticated words have been attested due to the limitations highlighted earlier.

## 5.2 Semasio-onomasiological Indices

In addition to the word frequency effect of L2 vocabulary learning, we also investigated the interplay of form-meaning mappings in the resource.

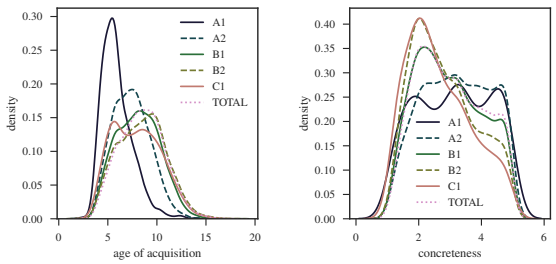
We observe on the one hand that the degree of polysemy and synonymy attested in the resource is strongly correlated to the degree of synonymy and polysemy that is expected in the Open Dutch WordNet (ODWN) (Figure 5). We can therefore conclude that in addition to the correlation between the estimated frequencies and Subtlex-NL (cf. *supra*), the word senses included in the resource are also consistent with the structure of a general semantic network.

However, the lower extent of onomasiological variation (i.e. meaning-to-form mappings) in NT2Lex compared to ODWN synonymy (Figure 5b) might be indicative of the specialised nature of the resource in that for a defined set of concepts it includes a limited range of lexicalizations, which are likely to be specific to the L2.

As for semasiological variation (i.e. form-to-meaning mappings) in NT2Lex, we observe an interesting decreasing trend in the degree of polysemy per level (Figure 5c). This highlights the fact that the lexical stock of elementary L2 texts con-

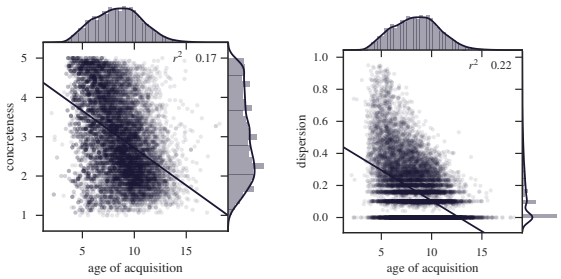
[8] http://www.dikverhaar.nl/wp-content/uploads/Basiswoordenlijst_2000_frequente_meest_woorden.pdf

[9] We should note that 61 of the 2,000 basic word forms are not attested in NT2Lex.




(a) age of acquisition per level (b) concreteness per level

Figure 6: Psycholinguistic norms in NT2Lex




(a) with word concreteness (b) with word familiarity

Figure 7: Age of acquisition in NT2Lex

tains more ambiguous entries, which in turn tend to be more easily processed (Millis and Bution, 1989). However, no other significant effects in terms of synonymy, polysemy or hypernymy were observed.

### 5.3 Psycholinguistic Norms

Finally, to investigate the interplay of different psycholinguistic norms in the resource, we use a lexical database of age of acquisition and concreteness norms for Dutch (Brysbaert et al., 2014). Figure 6 shows the distribution of the norms per each attested level in the resource.

**Age of acquisition** We observe that the vast majority of the lexical stock in the elementary levels (i.e A1/A2) contain words which are acquired the earliest by native speakers as well, approximately around the age of five, whereas the entries in the intermediate levels (i.e. B1/B2) levels are acquired later, approximately between ages 5 and 10. Moreover, the more concrete the word, the earlier it is acquired (Figure 7a), which is consistent with previous observations (Crossley et al., 2009). The earlier the word is acquired, the more familiar it is according to its dispersion in the resource (Figure 7b). As for the C1 level, we observe a similar trend, except for a smaller proportion of entries that are acquired earlier as well (with a higher concentration around the age of 5), which might also explain the higher average of Subtlex frequencies and the lower degree of sophistication attested at this level (cf. *supra*).

**Concreteness** As regards word concreteness, we observe on the one hand that the highest levels (i.e. B2/C1) contain a considerably higher proportion of abstract (less concrete) words. This observation highlights the fact that, even though the C1 level includes some outliers on the level of lexical frequency and sophistication, the distribution of the concreteness norms at C1 are similar to what is expected. The most basic levels, on the other hand, contain a lower proportion of abstract words, but the difference between the number of concrete and abstract words appears to be proportionally less clear-cut.

## 6 Conclusion

In this paper, we presented a new graded lexical resource for Dutch as a foreign language (NT2) based on the proficiency scale of the Common European Framework of Reference (CEFR). Similar to the previous CEFR-graded lexicons for French, Swedish and English, the NT2Lex resource contains graded frequency distributions per lexical entry which are estimated on L2 readers and textbook texts targeting a specific level on the CEFR scale. The novelty of the NT2Lex resource with respect to the common methodology of generating graded lexicons is concerned with the fact that the lexical entries are disambiguated for word senses and are also linked to WordNet synsets. We argued that this linkage gives us a better insight into word sense complexity in a foreign language.

We found that the estimated frequency and word sense distributions are in line with what one expects to observe in the target language. Moreover, the distributions of lexical entries per level in NT2Lex also appeared to be consistent with previous findings in terms of lexical complexity. As regards the features of lexical ambiguity, age of acquisition and concreteness, we observed that the lexical entries in the most basic levels of the resource (i.e. A1/A2) are more polysemous and acquired the earliest by non-native speakers, whereas the lexical entries in the more advanced levels (i.e. B2/C1) portray a significantly higher degree of abstractness and are acquired at a later

developmental stage.

We could thus conclude that the resource enables us to get a better grasp on what kind of vocabulary should be understood *a priori* when reading Dutch as a foreign language at a particular proficiency level. Of course, we should highlight that the assumptions that can be drawn are still limited in the sense that they are mainly based on expert knowledge drawn from pedagogical texts. Indeed, we lack extensive experimental data on what vocabulary is effectively understood when reading Dutch at a particular proficiency level and by a specific learner depending on his/her characteristics (e.g. native language, age, experience, etc.). As a future perspective, we therefore aim to contrast the knowledge we gained through the resource with this kind of receptive learner data.

## Acknowledgments

This research was funded by an F.R.S.-FNRS research grant. We would like to thank Anne-Sophie Desmet and Dorian Ricci for their helping hand and valuable input in the preparation of the resources and tools.

## References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1 edition. O'Reilly, Beijing ; Cambridge Mass.

Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114.

Marc Brysbaert, Michaël Stevens, Simon De Deyne, Wouter Voorspoels, and Gert Storms. 2014. Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, 150:80–84.

Annette Capel. 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(01).

Annette Capel. 2012. Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3.

John B. Carroll, Peter Davies, and Barry Richman. 1971. *The American Heritage Word Frequency Book*. Houghton Mifflin, Boston.

Council of Europe. 2001. *Common European Framework of Reference for Languages*. Cambridge University Press, Cambridge, UK.

Scott Crossley, Tom Salsbury, and Danielle McNamara. 2009. Measuring L2 Lexical Growth Using Hypernymic Relationships. *Language Learning*, 59(2):307–334.

Luise Dürlich and Thomas François. 2018. EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. To appear.

Nick C. Ellis. 2002. Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Thomas François, Nuria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3766–3773, Reykjavik, Iceland. European Language Resources Association (ELRA).

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: A CEFR-Graded Lexical Resource for Swedish Foreign and Second Language Learners. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 213–219, Portorož, Slovenia. European Language Resources Association (ELRA).

Nuria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *Electronic lexicography in the 21st century: thinking outside the paper : proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia, 2013, págs. 132-151*, pages 132–151.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Stefan Th. Gries. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4):403–437.

Sujay Kumar Jauhar and Lucia Specia. 2012. UOW-SHEF: SimpLex – Lexical Simplicity Ranking Based on Contextual and Psycholinguistic Features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 477–481, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emmanuel Keuleers, Marc Brysbaert, and Boris New. 2010. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42(3):643–650.

P. de Kleijn and E. Nieuwberg. 1993. *Basiswoordenboek Nederlands*. Wolters Leuven, Leuven.

Stephen Krashen. 1989. We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis. *The Modern Language Journal*, 73(4):440–464.

Batia Laufer and Geke C. Ravenhorst-Kalovski. 2010. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language*, 22(1):15–30.

Xiaofei Lu. 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, 96(2):190–208.

Bernard Lété, Liliane Sprenger-Charolles, and Pascale Colé. 2004. MANULEX: a grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, 36(1):156–166.

Carla Marello. 2012. Word lists in Reference Level Descriptions of CEFR (Common European Framework of Reference for Languages). In *Proceedings of the XV Euralex International Congress*, pages 328–335.

Michelle L. Millis and Scoti B. Bution. 1989. The effect of polysemy on lexical decision time: Now you see it, now you don't. *Memory & Cognition*, 17(2):141–147.

James Milton. 2010. The development of vocabulary breadth across the CEFR levels. A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe. In Inge Bartning, Maisa Martin, and Ineke Vedder, editors, *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, number 1 in Eurosla Monographs Series, pages 211–232. European Second Language Association.

Paul Nation and Robert Waring. 1997. Vocabulary size, text coverage, and word lists. In Norbert Schmitt and Michael McCarthy, editors, *Vocabulary: description, acquisition and pedagogy*, Cambridge language teaching library. Cambridge Univ. Press, Cambridge.

Ildikó Pilán, Elena Volodina, and Lars Borin. 2016a. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *TAL*, 57(3):67–91.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016b. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, pages 2101–2111, Osaka, Japan. Association for Computational Linguistics.

M. C. Postma, E. Miltenburg, R. Segers, A. Schoen, and P. T. J. M. Vossen. 2016. Open Dutch WordNet. In *Proceedings of the Eigth Global Wordnet Conference*.

David Qian. 1999. Assessing the Roles of Depth and Breadth of Vocabulary Knowledge in Reading Comprehension. *Canadian Modern Language Review*, 56(2):282–308.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016a. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 230–236, Portorož, Slovenia.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016b. Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. In *Actes de la 23ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'16)*, pages 221–234, Paris, France.

James B. Tharp. 1939. The Measurement of Vocabulary Difficulty. *The Modern Language Journal*, 24(3):169–178.

Frank Van Eynde. 2004. Part of speech tagging en lemmatisering van het Corpus Gesproken Nederlands. Technical report, Centrum voor Computerlinguïstiek, KU Leuven, Belgium.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016. SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint 5th NLP4CALL and 1st NLP4LA workshops (SLTC 2016)*, pages 76–84, Umeå, Sweden. Linköping University Electronic Press.

Piek Vossen, Attila Görög, Rubén Izquierdo, and Antal Van den Bosch. 2012. DutchSemCor: Targeting the ideal sense-tagged corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 584–589, Istanbul, Turkey. European Language Resources Association (ELRA).

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, Massachusetts.