# Understanding the Lexical Simplification Needs of Non-Native Speakers of English

**Gustavo Henrique Paetzold** and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{g.h.paetzold,l.specia}@sheffield.ac.uk

## Abstract

We report three user studies in which the Lexical Simplification needs of non-native English speakers are investigated. Our analyses feature valuable new insight on the relationship between the non-natives' notion of complexity and various morphological, semantic and lexical word properties. Some of our findings contradict long-standing misconceptions about word simplicity. The data produced in our studies consists of 211,564 annotations made by 1,100 volunteers, which we hope will guide forthcoming research on Text Simplification for non-native speakers of English.

## 1 Introduction

Text Simplification is a useful application both to improve other Natural Language Processing tasks and to assist language-impaired readers (Chandrasekar et al., 1996). When a simplifier aims to help people, understanding their needs becomes very important. In Lexical Simplification – the task of replacing complex words and expressions with simpler alternatives – this has been shown to be the case (Rello et al., 2013b; Rello et al., 2013a; Rello et al., 2013c). They describe several user studies conducted with readers suffering from Dyslexia and outline the most recurring challenges faced by them, as well as the most effective ways to overcome these challenges.

Given the widespread availability of content in English, non-native speakers of English become an important group to focus on. Reves and Medgyes (1994) elaborate on the differences between native and non-native English teachers from an educational and behavioural standpoint. However, we were not able to not find user studies that investigate the needs of non-native speakers from the perspective of Text Simplification. In this paper, we introduce three user studies that aim to do so.

Each user study pertains to one of three steps in the usual Lexical Simplification pipeline (Figure 1): Complex Word Identification, Substitution Selection and Substitution Ranking. In the sections that follow, we describe the findings of each user study.

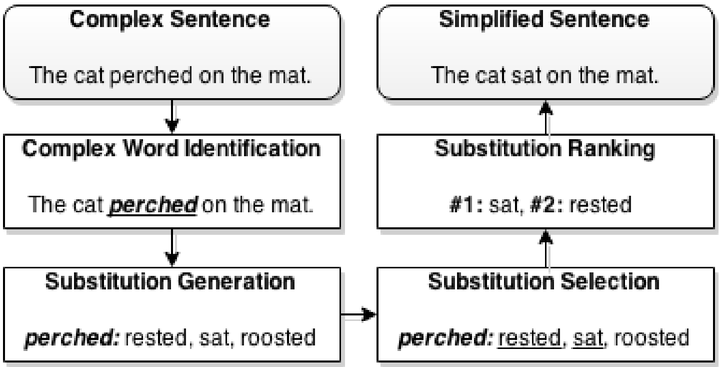


Figure 1: Common Lexical Simplification Pipeline

*Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*,
pages 717–727, Osaka, Japan, December 11-17 2016.

## 2 Complex Word Identification

Complex Word Identification (CWI) is the task of deciding which words should be simplified in a text. Effective CWI strategies identify words which should not be simplified, and hence prevent LS systems from making inappropriate replacements (Paetzold, 2015). As shown in (Paetzold and Specia, 2013; Shardlow, 2014), ignoring CWI can considerably decrease the quality of the output produced by a simplifier. The goals of our user study on CWI are to:

- Provide a better understanding on features of words that challenge non-native English speakers, and

- Create a dataset that allows us to conceive models that automatically identify complex words.

### 2.1 Data Sources

We selected 9,200 sentences with 20-40 words in length, at random, from three sources:

- **CW Corpus** (Shardlow, 2013b): Composed of 731 sentences from the Simple English Wikipedia in which exactly one word has been simplified by editors from the standard English Wikipedia. 231 sentences that conformed to our criteria were extracted.

- **LexMTurk Corpus** (Horn et al., 2014): Composed of 500 sentences from the Simple English Wikipedia containing one word simplified from the standard English Wikipedia. 269 sentences were extracted.

- **Simple Wikipedia** (Kauchak, 2013): Composed of 167,689 sentences from the Simple English Wikipedia, each aligned to an equivalent sentence in the standard English Wikipedia. We selected a set of 8,700 sentences from the Simple Wikipedia version that were aligned to an identical sentence in Wikipedia.

### 2.2 Annotation Process

400 non-native English speakers participated in this study, all students and staff from various universities around the world. Volunteers provided information about their native language, age, education level and English proficiency level according to CEFR (Common European Framework of Reference for Languages). They were given a set of sentences and asked to judge whether or not they could understand the meaning of each content word and to annotate all words that they could not understand individually, even if they could comprehend the meaning of the sentence as a whole. The exact instructions given to annotators are as follows:

> *For each sentence, mark all the words you do not understand, even if you understand the sentence as a whole. If you understand all of them, just select the "I understand all words!" option.*

In order to offer some form of financial compensation for their work, all annotators who successfully completed the task were automatically included in a monetary prize draw (£50). This compensation method was used in all user studies described in this paper.

For agreement analysis purposes, 200 sentences were annotated by 20 volunteers each, while the remaining 9,000 sentences were annotated by only one volunteer (i.e. each volunteer annotating an average of 22 sentences from the 9,000).

## 3 Dataset Analysis

The resulting dataset contains 158,624 annotations. 3,854 distinct words (6,388 in total) were deemed complex by at least one annotator. In the following sections, we discuss details of the data collected.

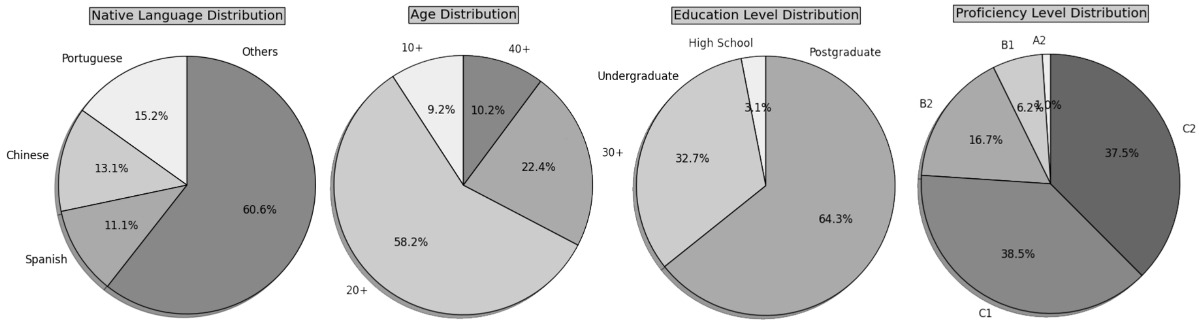
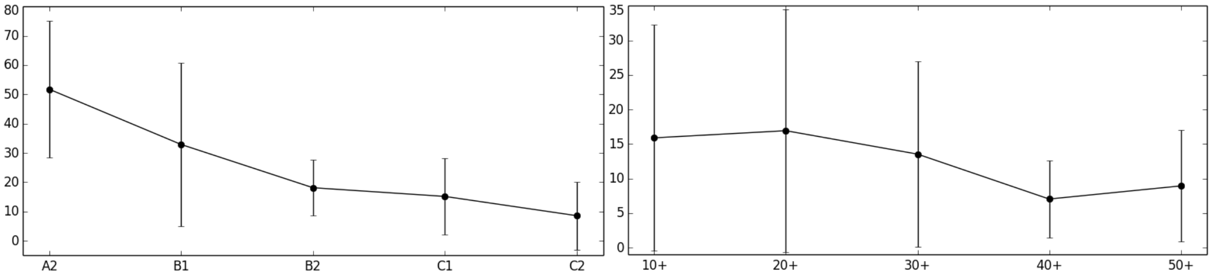


Figure 2: Annotators' backgrounds



(a) Proficiency level versus words deemed complex

(b) Age bands versus words deemed complex

Figure 3: Relationship between number of words deemed complex and the annotators' profiles

### 3.1 Profile of Annotators

Annotators spoke 45 different languages. The distributions with respect to native language, age, education and English proficiency levels are illustrated in Figure 2. As shown in Figures 3a and 3b, the data reveals interesting correlations between the number of complex words annotated and volunteers' age or English proficiency level.

Through F-tests, we found a significant difference ($p < 0.01$) between the band of 40+ years of age and the bands of 10+, 20+ and 30+ years of age. We also found significant differences between almost all English proficiency levels above A2, except between B2 and C1. Interestingly, B2 and C1 happen to have the same description in the London School Level Scale[1]: "*I speak and understand well but still make mistakes and fail to make myself understood occasionally*". We did not find significant differences among education levels.

### 3.2 Analysis of Data Sources

We found that the target words from the CW and LexMTurk datasets were deemed complex at least once by our annotators in only 51.9% and 40.8% of the instances, respectively. As for the remaining Simple Wikipedia instances, we discovered that at least one word in 27.3% of the instances was deemed complex by an annotator, which suggests that the simplified version of Wikipedia may still challenge non-native English speakers.

### 3.3 Features of Complex Words

We extracted and analysed 15 features that highlight the differences between simple words and those deemed complex by the annotators. We choose these features because they are the most widely used word complexity indicators in current work in Text Simplification. The features can be grouped in three types:

- **Morphological:** Word length and number of syllables, according to Morph Adorner (Burns, 2013).

[1] http://www.londonschool.com/level-scale

- **Semantic:** Number of senses, synonyms, hypernyms and hyponyms, according to WordNet (Fellbaum, 1998).

- **Lexical:** N-gram language model log-probabilities from the SubIMDB (Paetzold and Specia, 2016), Subtlex (Brysbaert and New, 2009) and Simple Wikipedia (Kauchak, 2013) corpora. We trained a specific language model using SRILM (Stolcke, 2002) from each of these corpora in order to estimate n-gram log-probabilities.

Table 1 shows the average feature values and standard deviations for all complex and simple words in the part of the dataset annotated by 20 volunteers. We define as complex any word which has been judged so by at least $n \in \{1, 5, 10\}$ annotators. The $[i, j]$ indicators present in the n-gram features of Table 1 refer to the number of tokens to the left ($i$) and right ($j$) of the words that was considered. Consequently, $[0, 0]$ refer to single-word frequencies. The column succeeding feature values indicate whether there was (•) or not (○) a statistically significant difference between complex and simple words ($p < 0.01$), given the results of an F-test.

| | $\mathbf{n=1}$ | | | $\mathbf{n=5}$ | | | $\mathbf{n=10}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **Feature** | **Complex** | **Simple** | $p$ | **Complex** | **Simple** | $p$ | **Complex** | **Simple** | $p$ |
| Length | $6.7 \pm 2$ | $6.6 \pm 2$ | ○ | $7.5 \pm 2$ | $5.9 \pm 2$ | ○ | $7.1 \pm 2$ | $6.1 \pm 2$ | ○ |
| Syllables | $2.1 \pm 1$ | $2.2 \pm 1$ | ○ | $2.3 \pm 1$ | $1.8 \pm 1$ | ○ | $2.2 \pm 1$ | $1.7 \pm 1$ | ○ |
| Senses | $6.6 \pm 8$ | $8.2 \pm 8$ | • | $2.1 \pm 2$ | $9.1 \pm 9$ | • | $1.1 \pm 1$ | $8.8 \pm 9$ | • |
| Synonyms | $16.7 \pm 21$ | $20.0 \pm 21$ | • | $5.3 \pm 6$ | $22.5 \pm 23$ | • | $2.3 \pm 3$ | $22.7 \pm 22$ | • |
| Hypernyms | $4.8 \pm 6$ | $5.5 \pm 6$ | • | $1.7 \pm 2$ | $6.1 \pm 8$ | • | $0.9 \pm 1$ | $5.9 \pm 7$ | • |
| Hyponyms | $24.7 \pm 48$ | $32.3 \pm 64$ | • | $4.0 \pm 13$ | $36.9 \pm 52$ | • | $0.8 \pm 2$ | $32.8 \pm 52$ | • |
| Subimdb[0,0] | $-5.3 \pm 1$ | $-4.8 \pm 1$ | • | $-6.5 \pm 1$ | $-4.6 \pm 1$ | • | $-6.6 \pm 1$ | $-4.5 \pm 1$ | • |
| Subtlex[0,0] | $-10.4 \pm 21$ | $-5.0 \pm 5$ | • | $-31.3 \pm 41$ | $-4.6 \pm 1$ | • | $-51.3 \pm 46$ | $-4.4 \pm 1$ | • |
| Simple[0,0] | $-5.9 \pm 10$ | $-4.3 \pm 1$ | • | $-11.5 \pm 22$ | $-4.2 \pm 1$ | • | $-8.4 \pm 14$ | $-4.2 \pm 1$ | ○ |
| Subimdb[1,1] | $-11.3 \pm 3$ | $-10.7 \pm 3$ | • | $-12.9 \pm 3$ | $-11.0 \pm 3$ | • | $-13.2 \pm 3$ | $-9.7 \pm 3$ | • |
| Subtlex[1,1] | $-19.3 \pm 28$ | $-13.4 \pm 18$ | • | $-40.0 \pm 45$ | $-16.4 \pm 23$ | • | $-59.7 \pm 52$ | $-13.8 \pm 21$ | • |
| Simple[1,1] | $-11.4 \pm 18$ | $-8.7 \pm 9$ | • | $-16.7 \pm 26$ | $-7.9 \pm 2$ | • | $-10.7 \pm 15$ | $-8.1 \pm 2$ | ○ |

Table 1: Average and standard deviation of features of words deemed complex or simple by at least $n$ annotators. The $[i, j]$ indicators refer to the number of tokens to the left ($i$) and right ($j$) considered by n-grams. The $p$ columns' values indicate the presence (•) or not (○) of a statistically significant difference.

The results shed some light on word complexity for a non-native English speaker. They show that, unlike semantic and lexical features, length and number of syllables have little to do with complexity. Although it has been found that shorter words do promote understandability for readers suffering from Dyslexic (Rello et al., 2013b), our findings reveal that long words are not necessarily more difficult to understand for non-native English speakers.

When it comes to semantic properties, it can be noticed that, while both average and standard deviation values for complex words decrease as the number of complex judgements increase, the same does not happen for simple words. This phenomenon suggests a relationship between ambiguity and complexity, where complex words are more likely to be unambiguous. This finding is in line with those of (Shardlow, 2013a), who successfully modelled word complexity by exploring the hypothesis that complex words tend to be less ambiguous.

Interestingly, a somewhat similar relationship can be observed between lexical properties and word simplicity: while the n-gram log-probabilities of complex words are low on average and have high variance across all scenarios, the average log-probabilities of simple words are much higher, and they vary much less.

### 3.4 Agreement Analysis

Here we calculated Kappa's pairwise inter-annotator agreement coefficient (Carletta, 1996) for all pairs of annotators who were presented with sentences from the overlapping portion of the data. The values average to $0.616 \pm 0.05$, which is much higher than the agreement scores obtained in previous work in similar tasks. For example, the Lexical Substitution tasks of SemEval 2007 (McCarthy and Navigli,

2007) and 2010 (Mihalcea et al., 2010) obtain an agreement of 0.277 for their annotations, while the English Lexical Simplification task of 2012 obtains an agreement of 0.398 (Specia et al., 2012).

Perhaps more impressive is the agreement within certain classes of annotators. Although the agreement for annotators with the same proficiency level is lower (0.575 ± 0.07), the agreements within education levels and age bands are noticeably higher, reaching 0.638 ± 0.08 and 0.671 ± 0.08, respectively. The highest agreement is reached by annotators with the same native language: 0.718 ± 0.1. Inspecting the annotations, we found that the speakers of certain languages are sometimes challenged by words which, in most cases, are not considered complex by native speakers of any other languages. Table 2 illustrates the words with the highest percentage of variance (Brysbaert and New, 2009) between the number of times that they were deemed complex by the speakers of a specific native language, and the rest of the annotators.

| Language | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Arabic | fur | juvenile | apprenticed | city | serologic | link |
| Chinese | canton | inscribed | opium | referendum | thorax | contaminants |
| French | sewerage | subsequent | warships | escudo | dye | ridges |
| German | escape | strong | early | city | escudo | iconoclastic |
| Portuguese | rather | hurricane | undergo | southern | ruler | crude |
| Spanish | bailed | cryptanalysis | plaque | debris | demise | perm |

Table 2: Words with highest percentage of complexity variance per native language. Indexes in the first row indicate the words' percentage of variance rank, from highest to lowest.

## 4 Substitution Selection

Substitution Selection (SS) is the task of deciding which candidate substitutions can replace a complex word in a given context. Its goal is to prevent a simplifier from performing replacements that compromise the sentence's grammaticality and/or meaning. The goals of this user study were to:

- Understand what makes a good candidate substitute for a complex word, and
- Create a dataset to build more effective substitution selectors.

Notice that we skip the step of Substitution Generation in our user studies. We do so because we believe that the extensive experiments of Horn et al. (2014) and De Belder and Moens (2012) already provide sufficient insight with respect to the relationship between Substitution Generation and the needs of non-native English speakers.

### 4.1 Data Sources

We first created a list of 1,471 complex words by filtering any numbers, names, colours and stop words from the ones obtained in the CWI study. We then produced an average of 50 candidate substitutions for each word by combining the output of all Substitution Generation systems in the LEXenstein framework (Paetzold and Specia, 2015). These systems exploit complex-to-simple parallel corpora (Horn et al., 2014), word embedding models (Glavaš and Štajner, 2015; Paetzold and Specia, 2016), WordNet (Devlin and Tait, 1998; Biran et al., 2011) and the Merriam Dictionary[2] (Kajiwara et al., 2013). Using the strategy described in (Paetzold and Specia, 2015), we selected the 10 candidates with the highest semantic similarity to each complex word. Using the Text Adorning module of LEXenstein, we ensured that all candidates have the same conjugation form as the complex word itself.

Finally, we extracted, according to availability, up to three sentences from Wikipedia in which each of these complex words appear (2,554 in total), and created 25,540 annotation instances by replacing the complex word in each sentence with one of the 10 candidate substitutions selected.

[2] http://www.merriam-webster.com

### 4.2 Annotation Process

400 fluent speakers of English participated in this study, all students and staff from different universities around the world. We recruited native English speakers for this task because it requires that the annotator understands the meaning of the complex word in question in order to make the necessary judgements.

Each annotator was presented with 80 annotation instances accompanied by the original complex word for reference. The exact instructions given to annotators are as follows:

> *Judge the following candidate substitutions of complex words with respect to their grammaticality and meaning preservation. When judging, please ignore any grammatical errors that are not caused by the substitution.*

For each instance, annotators were presented with two options:

- The substitution preserves the sentence's *grammaticality*, and

- The substitution preserves the original sentence's *meaning*.

Volunteers could select both, either or none of them. The resulting dataset contains 25,540 annotated instances. For agreement analysis purposes, 1,600 instances were annotated by 5 volunteers each, while the remaining 23,940 instances were annotated by a single volunteer. In total, 31,940 annotations were gathered. Notice that word simplicity is not taken into account in this user study, given that we wish to study how here readers interpret word replaceability only.

### 4.3 Dataset Analysis

We calculated several features to compare the grammatical and/or meaning preserving substitutions against the remaining substitutions. Table 3 illustrate the average and standard deviation feature values of candidates annotated positively by at least three out of five annotators (Good), and not (Bad), with respect to their grammaticality, meaning preservation, and both of them jointly. We chose six features, which can be grouped as:

- Language model probabilities of the sentence with the candidate in place of the target, given four 3-gram language models trained over the SubIMDB (Paetzold and Specia, 2016), Subtlex (Brysbaert and New, 2009) and Simple Wikipedia (Kauchak, 2013) corpora. Language model sentence probabilities have been used in previous work to create very effective Lexical Simplification systems (Horn et al., 2014; Glavaš and Štajner, 2015; Paetzold and Specia, 2016). Language models were trained with SRILM.

- The cosine word vector similarity between the candidate and the target (Target Sim.), as well as the average cosine similarity between the vector of the candidate and the vectors of content words in the sentence (Context Sim.). These features have been used in the creation of unsupervised lexical simplifier (Glavaš and Štajner, 2015). The embeddings model was trained with word2vec (Mikolov et al., 2013) with the CBOW architecture and 500 dimensions over a corpus of 7 billion words extracted from various sources (Paetzold, 2015; Brysbaert and New, 2009; Kauchak, 2013).

- The probability of the candidate receiving the same POS tag attributed to the target (POS Prob.). This feature has been shown a strong indicator of grammaticality (Aluisio and Gasperin, 2010; Nunes et al., 2013). The POS tag conditional probability models were trained over POS tags produced by the Stanford Parser (Klein and Manning, 2003) over the NewsCrawl corpus[3].

The column following the average values for Good and Bad candidates contain • for features for which we found a statistically significant difference between the averages through a F-test ($p < 0.01$), and ○ for the remainder. The results suggest that, even though they are able to account for context, n-gram

[3] http://www.statmt.org/wmt11/translation-task.html

language model probabilities are much less effective in distinguishing good from bad candidates than the word vector distance between target and candidate words. Nonetheless, the same cannot be observed for the average similarity between candidate and context words.

Another interesting finding from our results that agree with previous contributions (Aluisio and Gasperin, 2010; Nunes et al., 2013) refers to the probability of the candidate receiving the POS tag of the target (POS Prob.), which does indeed show a strong relationship with grammaticality.

| | **Grammaticality** | | | **Meaning** | | | **Joint (G/M)** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Feature** | **Good** | **Bad** | $p$ | **Good** | **Bad** | $p$ | **Good** | **Bad** | $p$ |
| Prob. Subimdb | $-0.9 \pm 0.3$ | $-1.0 \pm 0.3$ | ○ | $-1.0 \pm 0.3$ | $-0.9 \pm 0.3$ | ○ | $-0.9 \pm 0.2$ | $-1.0 \pm 0.3$ | ● |
| Prob. Subtlex | $-3.1 \pm 1.3$ | $-3.2 \pm 1.7$ | ● | $-3.2 \pm 1.4$ | $-3.2 \pm 1.7$ | ● | $-3.1 \pm 1.5$ | $-3.4 \pm 1.8$ | ○ |
| Prob. Simple | $-4.2 \pm 1.6$ | $-4.3 \pm 1.9$ | ● | $-4.2 \pm 1.8$ | $-4.3 \pm 1.9$ | ● | $-4.2 \pm 1.7$ | $-4.4 \pm 2.0$ | ○ |
| Target Sim. | $0.41 \pm 0.2$ | $0.29 \pm 0.2$ | ● | $0.39 \pm 0.2$ | $0.28 \pm 0.2$ | ● | $0.34 \pm 0.2$ | $0.27 \pm 0.2$ | ● |
| Context Sim. | $0.08 \pm 0.1$ | $0.06 \pm 0.1$ | ○ | $0.08 \pm 0.1$ | $0.06 \pm 0.1$ | ○ | $0.07 \pm 0.1$ | $0.06 \pm 0.1$ | ● |
| POS Prob. | $0.62 \pm 0.4$ | $0.44 \pm 0.4$ | ● | $0.53 \pm 0.4$ | $0.46 \pm 0.4$ | ○ | $0.58 \pm 0.4$ | $0.32 \pm 0.4$ | ● |

Table 3: Average and standard deviation of features of those words which were deemed grammatical, meaningful, or both by at least 3 annotators, and those that were not.

Out of the 356 candidates judged both grammatical and meaning preserving by at least three annotators, 171 (48%) are not listed in WordNet as either synonyms, hypernyms or hyponyms of the target word. This suggests that simplification strategies such as the ones of (Devlin and Tait, 1998) and (Biran et al., 2011), which extract candidate substitutions of complex words from WordNet, can suffer from low coverage.

### 4.4 Agreement Analysis

The average Kappa inter-annotator agreement scores for the data in this user study are $0.391 \pm 0.16$ for grammaticality, $0.424 \pm 0.16$ for meaning preservation, and $0.450 \pm 0.16$ for both of them jointly. Inspecting the data, we found that most disagreements resulted from situations in which the target word was part of a multi-word expression. Consider for example the target word *turn* in the sentence "*That in turn makes it difficult to affect policies to curb distracted driving*", which, in this case, is part of the multi-word expression *in turn*. Annotators were very much divided on whether or not candidate *reverse* preserved either grammaticality or meaning in this case: some judged it to be neither grammatical nor meaningful, while others claimed it to be grammatical or meaningful.

## 5 Substitution Ranking

In Substitution Ranking (SR), candidates are ranked according to their simplicity so that the complex word is replaced with the simplest candidate available. The goals of our user study are to:

- Discover which metrics best capture simplicity for non-native speakers, and
- Create a dataset for the evaluation and training of SR strategies.

### 5.1 Data Sources

We extracted 901 sentences from those collected in our Substitution Selection user study which had a minimum of two and maximum of four candidates annotated as both grammatical and meaning preserving by at least three annotators. In order to access the simplicity of these substitutes, we added the target complex word of each sentence to the set of candidates, and then replaced the target word in each sentence with a gap marker. Finally, we created an annotation instance for each pair of candidates, totalling 4,200 pairs (438 sentences * 3 pairs (3 candidates including complex word itself) + 436 * 6 pairs (4 candidates including complex word) + 27 * 10 pairs (5 candidates including complex word)).

### 5.2 Annotation Process

300 non-native English speakers participated in this study, all students and staff from different universities around the world. Volunteers provided anonymous information about their native language, age,

education level and English proficiency level. Each volunteer was presented with 70 annotation instances, each composed of a sentence with a gap and two candidates to fill it with.

For each instance, volunteers were asked to judge which candidate made the sentence easier to understand. They could also indicate that both candidates made the sentence equivalently complex/simple. The exact instructions given to annotators are as follows:

> *For each of the following instances, select which candidate makes the sentence easier to understand. If the words are equally complex/simple, select the "The words are equally simple" option. Please overlook any grammatical or spelling errors.*

All instances were annotated by 5 volunteers. A total of 21,000 annotations were produced.

Once instances were annotated, we used the algorithm introduced by (Wauthier et al., 2013) to infer rankings from binary comparisons, and hence produce 901 instances composed of a sentence, a target complex word, and a set of candidate substitutions ranked according to their simplicity.

### 5.3 Dataset Analysis

For validation purposes, we computed the correlation between the simplicity rankings and the same 15 features used in the dataset analysis for our Complex Word Identification user study, described in Section 3.2. Notice that in our work we measure simplicity as the ease with which one can understand a given portion of text, which is the opposite of the definition of complexity used in our study on CWI. We use three evaluation metrics: Spearman correlation ($r$), Pearson correlation ($\rho$) and TRank. The TRank metric was introduced by (Specia et al., 2012) and measures the proportion of times in which the word ranked simplest by a given feature is also ranked simplest by the annotators.

The values in Table 4 reveal that word length and number of syllables correlate poorly with word complexity, while simpler words tend to be more ambiguous and occur more frequently in corpora. These findings reinforce the ones from our CWI user study. More importantly, our results show that correlation and TRank scores of [1,1] (one token to the left and right) n-grams consistently outperform the scores of single-word frequencies ([0,0]) according to all metrics used. These findings contradict a long-standing assumption that context is not an important factor in word simplicity estimation (Devlin and Tait, 1998; Carroll et al., 1999; Biran et al., 2011; Rello et al., 2013b; Shardlow, 2013a).

| **Feature** | $r$ | $\rho$ | **TRank** |
|---|---|---|---|
| Length | 0.172 | 0.179 | 0.386 |
| Syllables | 0.097 | 0.095 | 0.340 |
| Senses | −0.345 | −0.349 | 0.505 |
| Synonyms | −0.288 | −0.297 | 0.454 |
| Hypernyms | −0.289 | −0.297 | 0.472 |
| Hyponyms | −0.309 | −0.300 | 0.453 |
| Subimdb[0,0] | −0.419 | −0.436 | 0.539 |
| Subtlex[0,0] | −0.465 | −0.467 | 0.556 |
| Simple[0,0] | −0.490 | −0.468 | 0.578 |
| Subimdb[1,1] | −0.463 | −0.473 | 0.579 |
| Subtlex[1,1] | −0.496 | −0.496 | 0.590 |
| Simple[1,1] | **−0.501** | **−0.475** | **0.593** |

Table 4: Simplicity correlation analysis between features and annotations

### 5.4 Agreement Analysis

The average Kappa inter-annotator agreement scores for this user study resemble the ones reported in Section 3.4: although the agreement between all annotators is encouraging ($0.454 \pm 0.05$), the scores are even higher for annotators with similar backgrounds. Annotators within the same education level, age band and proficiency level reach agreement scores of $0.468 \pm 0.01$, $0.482 \pm 0.02$ and $0.486 \pm 0.01$, respectively. Like what was observed in our user study on Complex Word Identification, the highest agreement comes from annotators that speak the same native language ($0.601 \pm 0.15$). This serves as further evidence that one's native language plays an important role on vocabulary acquisition.

## 6 Conclusions

We have described three user studies conducted with the goal of understanding the simplification needs of non-native speakers of English.

In our Complex Word Identification study we learned that words which are simpler to non-native English speakers have much higher probabilities according to language models, both alone and in context, while those which are more complex to them tend to have a smaller number of senses in WordNet. In contrast with what was reported by (Rello et al., 2013b) in experiments with readers who suffer from Dyslexia, we found no evidence of a relationship between the non-natives' perception of complexity and neither word length or number of syllables. Our experiments also showed that while a reader's English proficiency level indicates how many words will pose a challenge to them, their native language indicates which words these will be.

In our Substitution Selection study we found that, despite disregarding contextual information altogether, word vector distances between a complex word and a candidate substitution are more reliable than language model probabilities in capturing both grammaticality and meaning preservation. We also found that the conditional probability of a candidate with respect to the grammatical role of the complex word is a reliable indicator of grammaticality. From our agreement analysis, we found evidence that single-word replacements tend to compromise the understanding of multi-word expressions.

In our Substitution Ranking study, we found further evidence that, unlike ambiguity indicators and language model probabilities, length and number of syllables have little to do with word simplicity for non-native speakers of English. N-gram probabilities proved the most reliable simplicity indicators among the features evaluated, which contradicts the assumption often made in earlier work that context offers no important clues on a word's simplicity.

Table 5 summarises the data produced from each of these studies. Some examples of models and applications that could be built from these datasets are readability assessment tools, semantic analysers, text profilers and full lexical simplifiers. All of the datasets described herein can be downloaded from `http://ghpaetzold.github.io/data/User_Studies_NNS.zip`.

| User Study | Sentences | Words | Word Pairs | Annotators | Annotations |
|---:|:-:|:-:|:-:|:-:|:-:|
| Complex Word Identification | 9,200 | 87,244 | - | 400 | 158,624 |
| Substitution Selection | 2,554 | 25,540 | - | 400 | 31,940 |
| Substitution Ranking | 901 | 3,193 | 4,200 | 300 | 21,000 |
| Total | **12,655** | **115,977** | **4,200** | **1,100** | **211,564** |

Table 5: Summary of annotated data produced: number of unique sentences, words and word pairs annotated, as well as the number of annotators who participated and annotations produced.

## Acknowledgements

This work has been partially supported by the European Commission project SIMPATICO (H2020-EURO-6-2015, grant number 692819).

## References

Sandra Aluisio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: The porsimples project for simplification of portuguese texts. In *Proceedings of the 2010 NAACL Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th ACL*, pages 496–501.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41:977–990.

Philip R Burns. 2013. Morphadorner v2: A java library for the morphological adornment of english language texts. *Northwestern University, Evanston, IL.*

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of the 9th EACL*, pages 269–270.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics*.

Jan De Belder and Marie-Francine Moens. 2012. A dataset for the evaluation of lexical simplification. In *Computational Linguistics and Intelligent Text Processing*, pages 426–437. Springer.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd ACL*, pages 63–69.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd ACL*, pages 458–463.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Rocling*, pages 59–73.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st ACL*, pages 1537–1546.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st ACL*, pages 423–430.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th SemEval*, pages 48–53.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 9–14.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Bernardo Pereira Nunes, Ricardo Kawase, Patrick Siehndel, Marco a. Casanova, and Stefan Dietze. 2013. As simple as it gets - a sentence simplifier for different learning levels and contexts. In *Proceedings of the 13th ICALT*, pages 128–132.

Gustavo H. Paetzold and Lucia Specia. 2013. Text simplification as tree transduction. In *Proceedings of the 9th STIL*, pages 116–125.

Gustavo Henrique Paetzold and Lucia Specia. 2015. Lexenstein: A framework for lexical simplification. In *Proceedings of The 53rd ACL*, pages 85–90.

Gustavo Henrique Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of The 30th AAAI*.

Gustavo Henrique Paetzold. 2015. Reliable lexical simplification for non-native speakers. In *Proceedings of the 2015 NAACL Student Research Workshop*, pages 9–16.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013a. Simplify or help?: text simplification strategies for people with dyslexia. In *Proceedings of the 10th W4A*, pages 1–10.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013b. Frequent words improve readability and short words improve understandability for people with dyslexia. *Human-Computer Interaction*, pages 203–219.

Luz Rello, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. 2013c. One half or 50%? an eye-tracking study of number representation readability. *Human-Computer Interaction*.

Thea Reves and Peter Medgyes. 1994. The non-native english speaking efl/esl teacher's self-image: An international survey. *System*, 22(3):353–367.

Matthew Shardlow. 2013a. A comparison of techniques to automatically identify complex words. In *Proceedings of the 51st ACL Student Research Workshop*, pages 103–109.

Matthew Shardlow. 2013b. The cw corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *Proceedings of the 9th LREC*, pages 1583–1590.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 1st SemEval*, pages 347–355.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of the 2002 ICSLP*, pages 257–286.

Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. 2013. Efficient ranking from pairwise comparisons. In *Proceedings of the 30th International Conference on Machine Learning*, pages 109–117.