Multilingual and Cross-Lingual Complex Word Identification

Seid Muhie Yimam[†], **Sanja Štajner**[‡], **Martin Riedl**[†], **and Chris Biemann**[†]

[†]Language Technology Group, Department of Informatics, Universität Hamburg, Germany
[‡]Data and Web Science Group, University of Mannheim, Germany
{yimam, riedl, biemann}@informatik.uni-hamburg.de
sanja@informatik.uni-mannheim.de

Abstract

Complex Word Identification (CWI) is an important task in lexical simplification and text accessibility. Due to the lack of CWI datasets, previous works largely depend on Simple English Wikipedia and edit histories for obtaining 'gold standard' annotations, which are of mixed quality, and limited to English only. We collect complex words/phrases (CP) for English, German and Spanish, annotated by both native and non-native speakers, and propose language independent features that can be used to train multilingual and cross-lingual CWI models. We show that the performance of cross-lingual CWI systems (using a model trained on one language and applying it on the other languages) is comparable to the performance of monolingual CWI systems.

1 Introduction

The goal of lexical simplification (LS) is to replace words and phrases that are infrequent and difficult to understand with their simpler variants, which are easier to understand for various target readers, e.g. language learners (Petersen and Ostendorf, 2007; Aluísio et al., 2008), children (De Belder and Moens, 2010), and people with various cognitive or reading impairments (Feng et al., 2009; Rello et al., 2013; Saggion et al., 2015). Most LS systems have a Complex Word Identification (CWI) module at the beginning of their pipeline, which is then followed by the generation of possible substitution candidates, and the substitution candidates ranking (Paetzold and Specia, 2015, 2016a). Other systems do not have a separate CWI module but rather try to simplify any content word in the text, e.g. (Bott et al., 2012a; Glavaš and Štajner, 2015). They, however, still compare the complexity of the target word to all its substitution candidates, and in this way, perform the CWI task implicitly. The complexity comparison is usually performed taking into account the words frequency, length, ambiguity, or their combinations (Bott et al., 2012a; Glavaš and Štajner, 2015).

The 'gold standard' CWI datasets should ideally be compiled using human annotation of complex words and phrases in a controlled experiment (differentiating between target groups, e.g. native and non-native speakers). However, this is not always the case, e.g. (Shardlow, 2013; Horn et al., 2014). Currently the only existing 'gold standard' CWI corpus is the Semeval-2016 shared task CWI corpus for English (Paetzold and Specia, 2016b), annotated by non-native English speakers. In spite of the fact that such datasets are necessary for consistent automatic evaluation of LS systems and that CWI systems are known to improve the performance of automated LS systems (Paetzold and Specia, 2015), no similar datasets were built for any other language so far.

We address these needs by:

1) Collecting human annotations of complex words and phrases[1] by both native and non-native speakers in three languages (English, German, and Spanish), and for English, for three different text genres (Sections 3 and 4);

2) Proposing a language-independent set of features to build state-of-the-art automated CWI systems for all three languages (Section 5);

3) Showing that CWI systems using our language-independent feature set can be successfully trained on a dataset in one language and ap-

[1]In this paper, we interchangeably use **complex word**, **complex phrase**, or **hard word**, defined as a single word or a multi-word expression that causes difficulties in understanding the sentence or paragraph for a target reader.

plied on another language, thus reducing the need for compiling CWI datasets for various languages (Section 6).

2 Related Work

2.1 CWI Datasets

Currently the largest and most widely used CWI dataset, only available for English, is the SemEval-2016 shared task dataset (Paetzold and Specia, 2016b), which consists of 9,200 sentences collected from the older CW dataset created by Shardlow (2013), LexMTurk corpus (Horn et al., 2014), and Simple Wikipedia (Kauchak, 2013). Those previous datasets relied on Simple Wikipedia and edit histories as a 'gold standard' annotation of CWs, despite the fact that the use of Simple Wikipedia as a 'gold standard' for text simplification has been disputed (Štajner et al., 2012; Amancio and Specia, 2014; Xu et al., 2015). The SemEval-2016 CWI dataset, in contrast, is a collection of human annotations of CWs. Another improvement over the previous datasets is that all annotators were non-native English speakers, and therefore the two user groups (native and non-native English speakers) were not mixed as in the previous cases.

In the SemEval-2016 CWI dataset, for each given sentence, annotators were asked to annotate all content words (nouns, verbs, adjectives, and adverbs as tagged by Freeling (Padró and Stanilovsky, 2012)) that they could not understand individually even if they could understand the meaning of the sentence as a whole. Annotators were presented only one target word at the time. In the training dataset (200 sentences), each target word was annotated by 20 people, while in the test set (9,000 sentences), each target word was annotated only by a single annotator. The goal of the shared task was to predict the complexity of a word for a single non-native speaker based on the annotations of a larger group of non-native speakers. This introduced a strong bias and inconsistencies in the test set (test sentences were annotated by only one annotator, but not all of them by the same one, involving a total of 400 different annotators), reflected in very low F-scores obtained across all systems (Paetzold and Specia, 2016b; Wróbel, 2016).

To the best of our knowledge, there are no CWI datasets for any language other than English, neither there are English CWI datasets covering different text genres and both native and non-native English speaker's needs.

2.2 State-of-the-Art CWI Systems

The systems of the SemEval-2016 shared task were ranked based on F-score (the standard F_1-measure) and G-score (a harmonic mean between accuracy and recall) on the *complex* class only.

The best system with respect to the G-score (77.40%), but at the cost of F-score being as low as 24.60%, uses a combination of threshold-based, lexicon-based and machine learning approaches with minimalistic voting techniques (Paetzold and Specia, 2016b). The second best system by the G-score (77.30%) also uses various lexical, morphological, semantic and syntactic features. The highest scoring system with respect to F-Score (35.30%), which obtained a G-score of 60.80%, uses threshold-based document frequencies on Simple Wikipedia (Wróbel, 2016).

The problem of those best performing systems is that their features cannot be obtained for other languages, as the lexicons used and Simple Wikipedia do not exist for other languages than English. Therefore, we propose a language-independent set of features and build fully-automated CWI systems using those features, which perform en par with the best SemEval-2016 shared task systems. Furthermore, we show that our systems, taking advantage of the language-independent set of features, can even be trained on one language and successfully applied on CWI task in a different language.

3 Collection of the New CWI Datasets

We collect the annotations of complex words and phrases (longer sequences of words, up to maximum 50 characters), using the MTurk crowdsourcing platform, from multiple native and non-native English speakers (collecting the information about whether they are native speakers or not) on three different text genres. Similarly, we collect complex phrases for German and Spanish, using the same UI and instructions given in the respective languages.[2]

[2]Data available under CC-BY at: https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/complex-word-identification-dataset.html.

3.1 Data Selection

The English dataset comprises texts from three different text genres: professionally written news, Wiki news (amateur written news), and Wikipedia articles (amateur written encyclopedic articles). For the NEWS dataset, we used 100 news stories from the EMM NewsBrief[3] compiled by Glavaš and Štajner (2013) for their event-centered simplification task. For the WIKINEWS, we collected 42 news articles from the Wikipedia news articles. To resemble the existing CW resources (Shardlow, 2013; Horn et al., 2014; Paetzold and Specia, 2016b), we also collected 500 sentences from Wikipedia, belonging to different categories (politics, economics, science, etc.) to ensure that we do not introduce a topic bias. For German and Spanish, a total of 978 and 1,387 sentences, respectively, were collected from German and Spanish Wikipedia articles; we take one HIT (Human Intelligence Task) from each article when there are enough sentences for a HIT.

3.2 Procedure

For each language, we follow the same procedure except that the instructions and examples are provided in the same language as the dataset. Every single annotation task is cast into a HIT, which consists of 5–10 sentences forming a paragraph and is completed by 10 workers each. To select a complex phrase, workers can highlight single words or sequences of words using their mouse pointer. In order to control the annotation process, we do not allow users to select simple words such as determiners, numbers and stop words,[4] and very long phrases (more than 50 characters). We also have a compulsory question about whether the annotator is a native speaker or not, with a comment that the answer to this question does not influence the payment. To encourage annotators to carefully read the text and to only highlight complex words, we offer a bonus that doubles the original reward if at least half of their selections match selections from other workers. To discourage arbitrarily larger annotations, we limit the maximum number of selections that annotators can highlight to 10. If an annotator cannot find any complex word, we ask them to provide a comment. Examples 1, 2, and 3 show some of the CPs examples that were provided to the annotators for English, German and Spanish, respectively.

Example 1: *The Israeli official said the new ambassador to Cairo, Yaakov Amitai, was expected to travel to the Egyptian capital in December to present his credentials, but the embassy would not be staffed or resume normal activity until acceptable security arrangements were in place. Many Egyptians view Israel, which signed a peace treaty with Egypt in 1979 after four wars between the two countries, with hostility.*

Example 2: *Die Falschmeldung hatten die Yes Men (Kommunikationsguerilla) lanciert um an die Katastrophe in Bhopal vor 20 Jahren zu erinnern. Offiziellen Angaben zufolge starben 1.600 Menschen sofort und rund 6.000 weitere an den unmittelbaren Nachwirkungen. Bis heute summiert sich die Zahl der Opfer auf mindestens 20.000 Personen. Rund ein Fnftel der 500.000 Menschen die dem Gas ausgesetzt waren, leiden heute unter chronischen und unheilbaren Krankheiten , die sich offensichtlich zum Teil weiterverben knnen. Tausende erblindeten.*

Example 3: *Se ubica exactamente en la falda del cerro Uliachin y al pie de la laguna Patarcocha en la regin geogrfica de la puna donde est rodeada de montaas y lagunas. Se encuentra a pocos kilmetros del santuario nacional "Bosque de piedras de Huayllay" famoso por las misteriosas formas que le han dado el viento y el agua a los grandes macizos rocosos.*

Our data collection differs from previous works in several regards: 1) we allow annotators to select both single words and sequences of words. We think that such datasets are helpful in upstream tasks such as lexical simplification or paraphrasing. 2) We do not show a single sentence at a time, but rather multiple sentences (5-10), which allows annotators to select complex phrases based on larger contexts.

4 Analysis of Collected Annotations

A total of 181 workers (134 native and 47 non-native) participated in the annotation task and 25,617 complex phrase (CP) annotations have been collected, out of which 6,830 are unique CPs. The distribution of selected CPs across all annotators (*All*), native and non-native annotators separately, and the number of CPs selected by at least one native and one non-native annotator (*Both*) is presented in Table 1. The distribution of selected

[3]Freely available at: http://takelab.fer.hr/data/evsimplify/

[4]https://github.com/6/stopwords-json/

Dataset	All		Native		Non-native		Both
	Sing.	Mult.	Sing.	Mult.	Sing.	Mult.	
NewsBrief	2,373	10,358	2,032	5,981	1,824	2,923	1,860
WikiNews	1,565	5,687	1,253	4,052	1,091	756	896
Wikipedia	1,170	4,464	1,031	2,792	832	979	773
German	1,525	5,878	1,225	1,727	1,306	3,145	11,66
Spanish	3,983	10,297	3,952	10,080	236	12	172

(a) Annotation statistics (raw counts)

Dataset	All		Native		Non-native		Both
	Sing.	Mult.	Sing.	Mult.	Sing.	Mult.	
NewsBrief	18.64	81.36	25.36	74.64	38.42	61.58	14.61
WikiNews	21.58	78.42	23.62	76.38	59.07	40.93	12.36
Wikipedia	20.77	79.23	26.97	73.03	45.94	54.06	13.72
German	20.60	79.40	41.50	58.50	29.34	70.66	15.75
Spanish	27.89	72.11	28.16	71.84	95.16	4.84	1.21

(b) Annotation statistics in percentages

Table 1: Distributions of selected CPs across all annotators (*All*), native and non-native annotators separately, and the number of CPs selected by at least one native and one non-native annotator (*Both*). The column *Sing.* shows the number/percentage of annotations selected by only one annotator while the column *Mult.* shows the number/percentage of annotations selected by at least two annotators.

dataset	uni-gram	bi-gram	tri-gram+	total
NewsBrief	10,631	1,592	508	12,731
WikiNews	6,242	727	289	7,258
Wikipedia	4,776	661	197	5,634
German	6,832	356	215	7,403
Spanish	11,000	1,975	1,305	14,280

(a) Distribution of collected CW (raw counts)

dataset	uni-gram	bi-gram	tri-gram+
NewsBrief	83.50	12.50	3.99
WikiNews	86.00	10.02	3.98
Wikipedia	84.77	11.73	3.50
German	92.29	4.81	2.90
Spanish	77.03	13.83	9.14

(b) Distribution of collected CW in percentages

Table 2: Distribution of collected CW annotations across different text genres and languages with CP lengths.

CPs according to their length is presented in Table 2, while the distributions of annotators (native and non-native) per each language and on average per HIT are presented in Table 3.

4.1 Analysis of English CPs

As we can see from Table 1, around 80% of English CPs have been selected by at least two annotators. However, when we separate the selections made by native and non-native speakers, we see that: (1) the percentage of multiply-selected CPs by native speakers stays stable across different genres, while this is not the case for the non-native speakers; (2) the percentage of multiply selected CPs by non-native speakers is always significantly lower (54%–62%) than the percentage of multiply selected CPs by native speakers (73%–75%), regardless of the text genre; and (3) the percentage of CPs selected by at least one native and one non-native annotator is very low (12%–15%).

	Number of Annotators		Avg. annotators per HIT	
dataset	Native	Non-native	Native	Non-native
NewsBrief	67	29	5.8	4.2
WikiNews	56	12	7.6	2.4
Wikipedia	31	13	6.9	3.1
German	12	11	3.9	6.1
Spanish	48	6	9.8	0.2

Table 3: Distribution of number of annotators (native and non-native) per each language and on average per HIT.

These results indicate a higher heterogeneity of complex phrases among non-native speakers, raising doubts in how well can we predict complex phrases for a non-native speaker based on the annotations of other non-native speakers, and thus offering a possible explanation for the very low F-scores obtained by the best systems on the SemEval-2016 shared task. The low inter-annotator agreement (IAA) between native and non-native speakers (column *Both*) further indicates that the lexical simplification needs are very different for those two target groups. The IAA is

calculated based on percentage of exact matches of annotations.

4.2 Analysis of German CPs

For German CWI task, we had fewer annotators (23 in total, 12 native and 11 non-native). They highlighted a total of 7,403 complex phrases (2,952 were selected by native and 4,451 by non-native speakers), out of which 2,711 are unique CPs. In this task, we had more non-native than native annotators per HIT (6.1 non-native and 3.9 native on average per HIT, see Table 3). In contrast to English and Spanish CP annotations, in the German task, more than 92% of the annotations are single words (Table 2). Unlike in the English CWI task, we found a higher IAA among non-native German annotators (70.66%) than native German annotators (58.5%). This might be due to the fact that we have more non-native than native annotators per HIT. The IAA between the native and non-native annotators was also higher for the German task (15.75%) than for the English task (Table 1).

4.3 Analysis of Spanish CPs

For the Spanish CWI task, we had 54 annotators, 48 native speakers and 6 non-native speakers. A total of 14,280 annotations are collected (14,032 from the native and 248 from the non-native speakers) with 6,061 CPs being unique. Given a low number of participating non-native speakers, we excluded the non-native Spanish annotations from further experiments. We found a lower IAA among Spanish native speakers than among English native speakers. This lower IAA for Spanish is mainly due to the fact that annotators highlighted mostly multiple phrases (23% of the annotations, see Table 2).

5 Classification Experiments

We developed a binary classification system for the CWI task with a performance comparable to the state-of-the-art systems of the SemEval-2016 shared task. We base our discussions on the F-scores, but also report on the G-score (both calculated on the *complex* class only, as in the shared task) to compare our systems with the SemEval-2016 best systems. We have normalized and transformed all features to a common and language-independent feature space in order to build a multilingual CWI system. This multilingual CWI system design help us to conduct cross-lingual experiments.

5.1 Language-independent Feature Space

We use four different, language-independent sets of features.

Length and frequency features: Lexical substitution systems (Bott et al., 2012b; Glavaš and Štajner, 2015), and most of the CWI systems in the SemEval-2016 shared task use length- and frequency-related features. We use three length features: the number of vowels, the number of syllables, and the number of characters in the word. The number of syllables in the word are computed using the *texhyphj* tool,[5], which is a Java implementation of the Liang (1983) hyphenation algorithm available in multiple languages. We also use three sets of frequency features: frequency of the word in Wikipedia, frequency of the word in the Google Web 1T 5-Grams, and frequency of the word in the HIT/paragraph. In order to build a language independent feature representation, we normalized all the length and frequency features. For the length of vowels and syllables features, we normalize the count by dividing it with the token length. The length of the word (number of characters) was normalized by dividing the observed length with the average length of all words in the specific language of the datasets used to collect CPs. We have found that, for the English dataset, the average length of a word was 5.3 while for German and Spanish, it was 6.5 and 6.2 characters, respectively. Similarly, the frequency of the word in Wikipedia and Web1T corpus was normalized by dividing the frequency of the word by the maximum frequency of the word in the Wikipedia and Web1T corpus of the respective language.

Syntactic features: Based on the work of Davoodi and Kosseim (2016), the part of speech (POS) tag influences the complexity of the word. We used POS tags predicted by the Stanford POS tagger (Toutanova et al., 2003). However, the pre-trained models for the Stanford POS tagger are trained based on various POS tagged data: Penn Treebank[6] for English, the Stuttgart-Tübingen tag set (STTS)[7] for German, and the DEFT Spanish

[5] github.com/dtolpin/texhyphj

[6] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

[7] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

Treebank tag set[8] for Spanish. We have transformed the tag sets into universal POS tags based on the work of Petrov et al. (2012)[9].

Word embeddings features: The work of Ammar et al. (2016) introduced a single shared embedding space for more than fifty languages. For estimating multilingual embeddings, two methods called *multiCluster* and *multiCCA*, are designed with dictionaries and monolingual data. For our task, we have used the pre-trained embeddings model for the 3 languages.[10] We use the word2vec representations of content words (both complex and simple) as a feature, and also compute cosine similarities between the vector representations of the word and its context paragraph or sentence. The paragraph and sentence representations are computed by averaging the vector representations of the content words.

Topic Features: We use topic-relatedness feature that is extracted based on an LDA (Blei et al., 2003) model, which was trained on English, German and Spanish Wikipedia using 100 topics. We compute the cosine similarity between the word-topic vector and the document (the HIT in this case) vector as a feature. While this requires training a topic model for each language, the feature is still language-independent since we merely use the similarity between complex word candidate and context to gauge its in-topic-ness.

5.2 Classification Algorithms

We have used different machine learning algorithms from the scikit-learn machine leaning framework:[11] KNeighborsClassifier (KNN), NearestCentroid (NC), ExtraTreesClassifier (EXT), RandomForestClassifier (RF), and GradientBoostingClassifier (GB), and Support Vector Machines (SVM), and report only the results of the best classifiers based on NearestCentroid (NC).

On the SemEval-2016 shared task dataset, our system obtains an F-score of 35.44% and a G-score of 75.51%. The best system of the shared task by G-score obtained a 77.40% G-score, but with much lower F-score (24.60%) than ours, and the best shared task system by F-score obtained a 35.50% F-score, but with much lower G-score (60.80%) than ours. Therefore, our best system can be seen as comparable to the state-of-the-art CWI systems, but with the crucial difference of using a language-independent feature set.

5.3 Experimental Setups

We first build nine new datasets (three different genres times two different groups of annotators for English, native and non-native datasets for German and the native dataset for Spanish), by marking a word as *complex* if at least one annotator selected it as complex.

We further perform three sets of experiments:
Set I: Monolingual experiments on nine datasets (for all three languages).
Set II: Cross-language experiments.
Set III: Cross-group experiments.

The first set of experiments can be seen as benchmarking of CWI task on different languages and text genres. The second set of experiments explores the possibility of training a CWI system on one language and applying it on another language, which if possible, would imply that we do not need to collect CWI datasets for all languages. The third set of experiments explores whether the simplification needs of native and non-native speakers can be generalized.

In all three sets of experiments, we use the NC classifier and the same set of features (cf. Section 5.1), and we always use training sets of 200 sentences (to have the same size training dataset as in the SemEval-2016 shared task) and the rest of each dataset for testing (controlling for not having the same sentences in training and test sets in any experiment).

The distributions of the *complex* class in our nine new datasets and the SemEval-2016 shared task dataset are presented in Table 4. As can be noted, the percentages of *complex* instances are similar for both training and test sets in all our datasets, while this is not the case for the SemEval-2016 shared task. The unbalanced percentage of *complex* instances in training and test sets of the SemEval-2016 shared task is the consequence of the training dataset being annotated by 20 annotators and the test set being annotated by only one annotator, which is probably the cause for the very

[8] https://web.archive.org/web/20160325024315/http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

[9] https://github.com/slavpetrov/universal-pos-tags

[10] [http://128.2.220.95/multilingual/data/

[11] http://scikit-learn.org/stable/supervised_learning.html

Dataset	Native				Non-Native			
	Train		Test		Train		Test	
	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex
NewsBrief	970	459	768	360	1,068	361	860	270
Wiki news	898	531	436	250	1,119	310	516	170
Wikipedia	856	573	268	225	985	444	355	133
German	1,117	393	586	187	1,014	497	536	238
Spanish	1,529	647	1,189	435	–	–	–	–
Shared	–	–	–	–	1,531	706	84,090	4,131

(a) Raw counts of *complex* and *simple* instances in our training and test sets

Dataset	Native				Non-Native			
	Train		Test		Train		Test	
	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex
NewsBrief	67.88	32.12	68.09	31.91	74.74	25.26	76.11	23.89
Wiki news	62.84	37.16	63.56	36.44	78.31	21.69	75.22	24.78
Wikipedia	59.90	40.10	54.36	45.64	68.93	31.07	72.75	27.25
German	73.97	26.03	75.81	24.19	67.11	32.89	63.73	28.06
Spanish	70.27	29.73	73.21	26.79	–	–	–	–
Shared	–	–	–	–	68.44	31.56	95.32	4.68

(b) Percentages of *complex* and *simple* instances in our training and test sets

Table 4: Distribution of *complex* and *simple* instances in our nine new datasets and the SemEval-2016 shared task dataset.

low F-scores achieved by all systems on the shared task (Section 2). In order to avoid this problem, we used exactly the same annotation procedure for both training and test sets. For Spanish, we only report results for native annotators since we did not collect enough non-native annotations (cf. Section 4.3).

6 Results and Discussion

We present and discuss the results of each set of experiments in a separate subsection. In all experiments, as a baseline system, we use threshold-based document frequency using the English Simple Wikipedia, German Wikipedia and Spanish Wikipedia articles. We present results of all experiments based on the F_1-measure.

6.1 Monolingual Results (Setup I)

Table 5 presents the baseline as well as the results of the CWI systems for the nine datasets using the multilingual features. All of the CWI systems perform better than the baseline system. We can also see that for English, the CWI systems based on the datasets collected from native speakers perform better than CWI systems based on the datasets collected from non-native annotators.

6.2 Cross-Language Results (Setup II)

In the cross-language CWI systems, we train the source model in one language and test on the

Dataset	Native		Non-native	
	Our (NC)	Baseline	Our (NC)	Baseline
NEWS	**69.97**	66.01	**62.35**	60.28
WIKINEWS	**69.25**	66.56	**57.89**	51.50
WIKIPEDIA	**70.79**	67.20	**58.31**	53.53
GERMAN	**54.92**	51.37	**58.50**	56.57
SPANISH	**45.83**	44.04	–	–

Table 5: Results of our CWI system (NC) and the baseline system on our nine datasets using the multilingual features. The baseline is based on document frequency thresholds of Wikipedia corpora in the respective languages, with better system marked in bold. (Setup I)

datasets for other languages (both native and non-native datasets separately). As we can see from Table 6, when we use a CWI model trained on one of the English datasets and test it on the German datasets annotated by native or non-native speakers, we obtain similar results to (and, in some cases, even better than) those of the CWI models trained on German datasets. The same holds when we test the English CWI models on the native Spanish dataset.

When we train the CWI system on the Spanish native dataset and test it on the German datasets, we observe a slight decrease in performance in comparison to monolingual German CWI systems, but still very close.

The CWI systems trained on German datasets and applied on English datasets, however, show a

Training		Testing								
		NEWS		WIKI NEWS		WIKIPEDIA		GERMAN		SPANISH
		Native	Non-Native	Native	Non-native	Native	Non-native	Native	Non-native	Native
NEWS	Native	**69.97**	60.13	**71.45**	**59.76**	67.24	**57.14**	53.89	58.32	45.19
	Non-Native	67.49	**62.35**	69.24	58.53	67.95	55.28	53.02	58.92	44.79
WIKI NEWS	Native	69.25	57.69	70.91	58.84	64.98	54.34	54.54	58.42	44.48
	Non-Native	68.56	57.89	69.49	58.37	66.36	51.67	56.03	58.31	43.26
WIKIPEDIA	Native	69.75	58.54	71.02	58.64	**70.79**	55.61	52.93	58.64	45.29
	Non-Native	68.80	59.95	68.63	57.36	67.12	58.31	51.53	**59.14**	44.39
GERMAN	Native	67.42	57.55	64.12	51.01	61.99	50.48	**54.92**	57.69	42.76
	Non-Native	66.99	58.51	68.33	55.53	67.27	54.09	53.83	58.50	41.52
SPANISH	Native	66.07	58.17	68.69	55.43	62.67	51.89	53.53	56.82	**45.83**

Table 6: Results of the cross-group and cross-language experiments using for the nine datasets, with better system marked in bold.) (Setups II and III)

drop in the performance in comparison to monolingual English CWI systems. The same holds for the CWI systems trained on the Spanish native dataset and applied on the English test sets.

Therefore, we see that the CWI systems trained on one language can be used to identify complex words in another language.

6.3 Cross-Group Results (Setup III)

For the English datasets, training the CWI systems on native datasets and using them to identify complex words for non-native speakers seems to lead to worse performances than training the CWI systems on the non-native English datasets (Table 6). The opposite (training the CWI systems on non-native English datasets and using them to identify complex words for native speakers), however, seems to lead to better results than training the systems on the native English datasets.

For the cross-group German experiments, the results are exactly the opposite from those for English. One possible explanation could be the higher IAA between English native annotators and German non-native annotators (cf. Table 1) and the number of annotators per HIT being higher for English native and German non-native annotators (cf. Table 3).

7 Conclusions

Complex word identification (CWI) task is an important task in text accessibility and text simplification. So far, however, this task has only been addressed on the Wikipedia sentences and taking into account mostly the needs of non-native English speakers. Moreover, languages other than English did not receive any attention with regard to building either the CWI datasets or automated CWI systems.

We have collected a total of nine 'gold-standard' CWI datasets: six datasets for English (three genres times two groups of annotators), two datasets for German (for native and non-native speakers), and one dataset for Spanish native speakers.

Furthermore, we have developed a state-of-the-art automated CWI system with language-independent feature representations, and showed that it performs well regardless of text genre and language.

Most importantly, we demonstrated that it is possible to train CWI systems in one language and use them to identify complex words in a different language, by demonstrating that CWI systems trained with English datasets annotated by native and non-native speakers can be used to reliably identify complex words in German and Spanish with a drop of only 1-2% in performance, whereas CWI systems trained with German training sets annotated by non-native speakers can be used to identify complex words in English with maximal drop of only 2-4% in performance.

These results imply that state-of-the-art CWI systems can be built for many languages without a need for collecting new CWI datasets in those languages: it is safe to use existing CWI datasets for other languages.

The full dataset is available for download via the first author's homepage.

References

Sandra M. Aluísio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*. New York, NY, USA, DocEng '08, pages 240–248.

Marcelo Adriano Amancio and Lucia Specia. 2014. An Analysis of Crowdsourced Text Simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Gothenburg, Sweden, pages 123–130.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR* abs/1602.01925.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)* 3:993–1022.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012a. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proceedings of COLING 2012*. Mumbai, India, pages 357–374.

Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. 2012b. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proceedings of COLING 2012*. Mumbai, India, pages 357–374.

Elnaz Davoodi and Leila Kosseim. 2016. CLaC at SemEval-2016 Task 11: Exploring linguistic and psycho-linguistic Features for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California, USA, pages 982–985.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*. Geneva, Switzerland, pages 19–26.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Athens, Greece, EACL '09, pages 229–237.

Goran Glavaš and Sanja Štajner. 2013. Event-centered simplification of news stories. In *Proceedings of the Student Research Workshop at the International Conference on Recent Advances in Natural Language Processing*. Hissar, Bulgaria,, pages 71–78.

Goran Glavaš and Sanja Štajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China, pages 63–68.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. A Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland, USA, pages 458–463.

David Kauchak. 2013. Improving Text Simplification Language Modeling Using Unsimplified Text Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria, pages 1537–1546.

Franklin M. Liang. 1983. *Word hy-phen-a-tion by comput-er*. Ph.D. thesis, Stanford University, Department of Linguistics, Stanford, CA., USA.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey, pages 2473–2479.

Gustavo Paetzold and Lucia Specia. 2015. LEXenstein: A Framework for Lexical Simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. Beijing, China, pages 85–90.

Gustavo Paetzold and Lucia Specia. 2016a. Benchmarking Lexical Simplification Systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, pages 3074–3080.

Gustavo Paetzold and Lucia Specia. 2016b. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California, USA, pages 560–569.

Sarah E. Petersen and Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of Workshop on Speech and Language Technology for Education*. Farmington, Pennsylvania, USA, pages 69–72.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey, pages 2089–2096.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Proceedings of the INTERACT 2013: 14th IFIP TC13 Conference on Human-Computer Interaction., 2013*. Cape Town, South Africa, pages 203–219.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarević. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing* 6(4):14:1–14:36.

Matthew Shardlow. 2013. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Sofia, Bulgaria, pages 69–77.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2003)*. Edmonton, Canada, pages 982–985.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What Can Readability Measures Really Tell Us About Text Complexity? In *Proceedings of the LREC'12 Workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*. Istanbul, Turkey.

Krzysztof Wróbel. 2016. PLUJAGH at SemEval-2016 Task 11: Simple System for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California, USA, pages 953–957.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics* 3:283–297.