Metaphors in Text Simplification: To change or not to change, that is the question

Yulia Clausen
Linguistics Department, SFB 1287
University of Potsdam
Potsdam, Germany
yulia.clausen@uni-potsdam.de

Vivi Nastase
Department of Computational Linguistics
University of Heidelberg
Heidelberg, Germany
nastase@cl.uni-heidelberg.de

Abstract

We present an analysis of metaphors in news text simplification. Using features that capture general and metaphor specific characteristics, we test whether we can automatically identify which metaphors will be changed or preserved, and whether there are features that have different predictive power for metaphors or literal words. The experiments show that the Age of Acquisition is the most distinctive feature for both metaphors and literal words. Features that capture Imageability and Concreteness are useful when used alone, but within the full set of features they lose their impact. Frequency of use seems to be the best feature to differentiate metaphors that should be changed and those to be preserved.

1 Introduction

Metaphor is ubiquitous in everyday language and central to human thought (Lakoff and Johnson, 1980). We find manifestations of it in colloquial and academic discourse, newspaper, school textbooks, political discourse and probably anywhere language is used. There are conflicting views though on whether metaphors are a useful communication device. Golden (2010) found that metaphors present in school textbooks can make the overall content comprehension more difficult. On the other hand, the essence of metaphor is to make abstract concepts, which are often hard to grasp, more easily understandable through concrete descriptions (e.g. Kövecses, 2017).

In this paper we investigate metaphors in the context of news texts simplification. On a corpus of parallel sentences from news articles and their simplified version (to grade 4 level, which corresponds to 9-10 years of age), we analyze metaphors that are kept, changed or added in the simplified version. Our aim is to verify whether we can characterize and automatically detect metaphors that help or do not help text understanding in the context of news articles.

The task of automatic text simplification has received a considerable amount of attention within NLP research. The proposed systems have, for the most part, addressed lexical and syntactic transformations, such as substitution of difficult words with simpler equivalents or altering the structure of sentences to make them more easily understandable (e.g. Barlacchi and Tonelli, 2013; De Belder and Moens, 2010; Drndarević and Saggion, 2012; Torunoğlu-Selamet et al., 2016; Vu et al., 2014).

The automatic handling of metaphorical language has also been researched extensively. However, the studies have mainly investigated the possibilities of automatic metaphor identification. Simplification of metaphorical language has not been explicitly addressed yet. This could be attributed to the fact that metaphor simplification is a challenging task for automatic implementation (cf. Drndarević and Saggion, 2012). Some approaches have considered the problem of automatic metaphor interpretation (e.g. Bollegala and Shutova, 2013; Shutova, 2013), which aims to find literal paraphrases for metaphorical expressions. It is not clear though whether the literal version is easier to understand than the original metaphor. Sometimes lexical simplifications for complex words can be too basic to convey the original meaning (cf. Vu et al., 2014).

We take a step towards filling the gap in metaphor simplification research. We combine information (in the form of features) from text simplification, and characteristics of metaphors to investigate whether there are specific features that can predict whether metaphors should be changed, and if these are different from features that are pre-

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 423–434
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

dictive of lexical simplification in general. We use parallel versions ("raw" and simplified) of news data from Newsela[1], a company that produces professionally simplified news texts, in which we annotate metaphors. In experiments that predict whether a target word will be changed or not, we analyze the performance of the features used w.r.t. the type of the target word – metaphoric or literal (the full set of features is described in Section 4.2). We find that the Age of Acquisition is the strongest feature overall, for both metaphoric and literal words. Imageability, Familiarity and Concreteness are useful when used alone, but within the context of the full set of features they lose their impact. Frequency of use is an important feature for distinguishing metaphors that should be changed from those to be preserved.

2 Related work

Text simplification has numerous facets, and can be approached from different angles. The general need for simplification can be predicted based on the readability of a text, from the point of view of sentence complexity (Štajner et al., 2017) or a combination of lexical, syntactic and semantic text characteristics (De Clercq and Hoste, 2016). Simplification can be targeted by identifying complex words (e.g. Paetzold and Specia, 2016; Yimam et al., 2018), and then performing lexical simplification (e.g. Glavaš and Štajner, 2015; Glavaš and Vulić, 2018; Horn et al., 2014; Kriz et al., 2018).

Lexical simplification systems often build on sentence-aligned simplification corpora and propose substitutes for complex words from a number of synonyms based on the words' frequency, length and suitability for the original context (De Belder and Moens, 2010; Drndarević and Saggion, 2012; Vu et al., 2014). Approaches influenced by machine translation have also been explored, as lexical simplification can be viewed as monolingual translation (e.g. Nisioi et al., 2017; Xu et al., 2016; Zhu et al., 2010). Other neural based models have also been developed, which exploit word embeddings and their closeness in the vector space as clues for substitution candidates. Glavaš and Štajner (2015) produce word simplifications in a large regular corpus using word embeddings to perform lexical substitution tasks. The simplification candidates are ranked based on features such as semantic and context similarity, and information load.

Our focus in this paper is narrower. We aim to explore metaphors in text simplification, and check if there are specific features that predict whether a metaphor should be changed or not. To represent the instances in our data we use features that previous work on text simplification have shown to be beneficial, as well as features useful in metaphor identification tasks.

3 Data

The data for this study comes from Newsela, a company that provides professionally simplified news texts for school reading activities. The editors follow simplification guidelines and are assisted by a tool in detecting difficult words. There is no description of the criteria used by the tool to detect such words. Regarding metaphors, the instructions are brief and seem to draw attention to idioms rather than metaphors: "be literal in lower versions. No straight out metaphors, as in no 'paint into a corner' in 5th grade or below.".

Each Newsela article has five versions of different difficulty levels determined based on the Lexile[2] readability scores, which are used to measure the complexity of texts and assign them to appropriate grade levels. Using these parallel news texts allows for the quick identification of changed items to produce a dataset to which metaphor information is then added (Wolska and Clausen, 2017).

3.1 The dataset

We use a parallel corpus of 1,130 Newsela articles by Xu et al. (2015), where each original article has been aligned with its four simplified versions at the sentence level based on Jaccard similarity. For our study we look at the original (V0) and the most simplified (V4) versions, as between them we expect the most differences w.r.t. simplification strategies. From this corpus, we automatically sampled original sentences along with their equivalents from the chosen simplified version.

Each Newsela version covers several, unevenly distributed, grade levels. Because of the potential differences between the grade levels within versions, we sampled only articles at grade level 12 from the original version and grade level 4 from the most simplified version. The selected grade levels correspond to the largest subsets within the

[1] https://newsela.com/

[2] https://lexile.com/

respective versions. The sampling was randomized across documents to counter author and editor bias. The final dataset contains 582 documents, each consisting of one original sentence and one or more sentences in the corresponding simplified version.[3] All alignments were manually checked and corrected where necessary either by inserting missing sentences or by replacing wrong alignments with the correct ones. This resulted in 278 corrections, exemplified below ("m" indicates the manually inserted sentence, initially missing from the alignment):

V0 A year ago, [Shaw] Mychal suffered a concussion in a game that rendered him temporarily unable to walk or speak.
V4 Shaw suffered a concussion in a game last year.
V4-m Shaw could not walk or speak for a while.

3.2 Metaphor annotation

We focus on the two most common word classes – nouns and verbs. In the sampled documents, we annotated their occurrences in the original sentences as either metaphoric or not by following the guidelines of the metaphor identification procedure MIPVU (Steen et al., 2010).[4]

The annotation in this study builds on Wolska and Clausen (2017), where it was carried out as follows: one author initially identified metaphoric items in a smaller subset of the data. All unclear cases were then discussed with the second author and either resolved or left unannotated. The annotation was completed by the initial annotator. In this study, we use a version of the dataset with expanded annotations – every noun/verb left unannotated in the previous study was annotated for metaphoricity by the same annotator as in the initial study.

In MIPVU, metaphoricity is identified by examining a text on a word-for-word basis and determining the context and the basic senses of each word. "Words" are considered to be lexical units provided with separate part of speech tags.[5] A word is used metaphorically if its context sense can be sufficiently contrasted to and understood in comparison with its basic sense. The context sense of a word is "the meaning it has in the situation in which it is used", whereas the basic sense is taken to be "more concrete, specific, and human-oriented" (Steen et al., 2010, p. 33-35).

The senses are determined by means of a dictionary; we consult the Macmillan Dictionary[6], which is a standard reference used by the authors of the procedure. Different senses of a word correspond to separate, numbered descriptions within its grammatical category in a dictionary.

In an example from our dataset, given in (1), the verb *struggling* is used metaphorically, as there exists a more basic sense ("to use your strength to fight against someone or something"), which is contrasted to and compared with the contextual sense ("to try hard to do something that you find very difficult").[7]

(1) But now she's *struggling* to obtain documents required by the new law.

The quantitative information on the annotated dataset is summarized in Table 1.

Measure	**Count**
No. of sentences	566
No. of sentences containing metaphors	350
Mean No. of metaphors per sentence	1.7
No. of annotated metaphors	587
Verbs	354
Nouns	233
No. of annotated non-metaphors	2,952
Verbs	852
Nouns	2,100
No. of unique lexemes	2,261
Metaphoric	433
Non-metaphoric	1,828

Table 1: Statistics on the annotated dataset.

3.3 Simplification types

For each annotated word we marked its equivalent in the simplified version and determined the simplification type chosen by an editor.[8] There

[3] Documents where original and simplified versions were identical based on string comparison were excluded.

[4] The annotation was done with the tool BRAT (Stenetorp et al., 2012): http://brat.nlplab.org/

[5] In MIPVU, phrasal verbs and compound nouns are regarded as single lexical units. Although we annotated them, in this paper we experiment only with the single words due to the non-availability of various features for multi-item words.

[6] https://www.macmillandictionary.com/

[7] The definitions of the basic and contextual senses: https://www.macmillandictionary.com/dictionary/american/struggle_1

[8] We encountered cases of clauses, such as coordinate and subordinate, not retained in the simplified version. These clauses were not annotated, as they might have been removed

	Simplification	Original sentence	Simplified sentence
metaphoric	same metaphor	*. . . like the magnetized nails, unable to resist a powerful magnetic **force** in the galactic bulge . . .*	*Like the magnetized nails, they would have been unable to resist a powerful magnetic **force** in the galactic bulge . . .*
	other metaphor	*Obama also has **grappled** publicly with reconciling King's teachings on nonviolence . . .*	*Obama has **wrestled** publicly with living up to King's teachings on nonviolence . . .*
	changed to non-metaphor	*In exchange for a 4 percent **piece** of their companies, entrepreneurs in the program will gain access . . .*	*. . . people in the program will give up a 4 percent **share** of their companies. In exchange they will get . . .*
	phrase with metaphor(s)	*But now she's **struggling** to obtain documents required by the new law.*	*But now she's **having a hard time** getting the papers that the new law requires.*
	phrase w/o metaphor(s)	*Utah officials say that since 2008, highway crashes have **dropped** annually on stretches of rural Interstate . . .*	*They say there have **been fewer** accidents where the speed limit was raised.*
	word removed	*Our goal is to provide Internet service to people in areas that can't afford to **throw down** fiber lines . . .*	*Our goal is to provide Internet service to people in areas that can't afford **Ø** usual Internet lines . . .*
non-metaphoric	same non-metaphor	*"In the past several hundred **years**, people have cultivated the habit of smoking wherever they want," she said.*	*"In the past several hundred **years**, people have "gotten used to" smoking wherever they want," she said.*
	other non-metaphor	*With nothing less at stake than the future of planet Earth, NASA has decided to crowdsource ideas to **detect** and track asteroids . . .*	*NASA wants to **find** and track asteroids, but it needs help. It is asking people around the world for ideas . . .*
	changed to metaphor	*That information could help the team's trainers **implement** practice plans that keep him spry the rest of the season.*	*That could help the team's trainers **make** plans that keep him healthy for the season.*
	phrase with metaphor(s)	*"Even after the Holocaust, our minority still encounters racism and discrimination," he said, noting that they are Europe's last **hired**, first fired.*	*His people still suffer unfair and insulting treatment, he said. They are the last in Europe **to get jobs**. They are also the first to be fired.*
	phrase w/o metaphor(s)	*On Thursday, the snowpack was a paltry 25 percent of **average** for this time of year.*	*The snowpack was just one-quarter of **what it usually is** for this time of year.*
	word removed	*SnapDragon is a cross of Honeycrisp with a Jonagold-like hybrid that's easier for **farmers** to manage.*	*SnapDragon is a cross of the tasty Honeycrisp apple and another kind that's easier **Ø** to grow.*

Table 2: Simplification types for metaphoric and non-metaphoric lexical items.

are six simplification options that were identified for metaphoric items in Wolska and Clausen (2017), which we now apply to non-metaphoric items as well. A word can be preserved (*same metaphor/same non-metaphor*)[9], replaced by another word of the same metaphorical status (*other metaphor* for metaphoric items and *other non-metaphor* for literal items), replaced by a word of opposite metaphorical status (*changed to non-metaphor* for metaphors and *changed to metaphor* for literal items), rephrased with metaphorical language (*phrase with metaphors*) or without (*phrase without metaphors*), or removed (*word removed*). See Table 2 for an overview with examples.

The annotation of the simplification types in Wolska and Clausen (2017) was done as follows: on a smaller subset of sentences annotated for metaphoricity, two authors identified and discussed the simplification choices. Once these were finalized, one author annotated the remainder of the dataset and the second author 99 instances.[10] Inter-annotator agreement on the common subset was $\kappa = .87$. In the present study, one author extended the annotations.

The quantitative information on the annotated simplification types is summarized in Table 3.[11] The statistics show that metaphors can be both useful and confusing for communication: 62% of the phrases that contained metaphors in the original article version contain a metaphor (the same or another one) in the simplified version. A small number of non-metaphors (2.3%) were replaced with metaphors in the simplified version.

With respect to the two word classes – nouns and verbs – we note considerable variation in the dataset (see Table 4). 93% of the verbs (186 metaphoric and 368 literal) appear less than five times; 67% (143 metaphoric and 256 literal) only once. The most frequent verbs annotated as metaphoric are *have* (22), *make* (18) and *take*

due to various reasons, e.g. complex syntactic structure. This is to be differentiated from the option *word removed*, where the changes are performed on the word level and which we annotate.

[9] Morphological deviations are considered the same word.

[10] One erroneous instance had to be excluded.

[11] 30 of the annotated words are not included in the counts; they were excluded from the experiment part, as most of the features we use were not available for them.

Simplification type	Count
Metaphoric	584
same metaphor	299
other metaphor	43
changed to non-metaphor	101
phrase with metaphor(s)	20
phrase without metaphor(s)	14
word removed	107
Non-metaphoric	2,925
same non-metaphor	1,933
other non-metaphor	418
changed to metaphor	34
phrase with metaphor(s)	32
phrase without metaphor(s)	77
word removed	431

Table 3: Distribution of the simplification types.

(13). The verbs *say* (86) and *use* (16) are mostly used literally. Nouns behave similarly: 95% (167 metaphoric and 1,049 literal) appear less than five times; 68% (131 metaphoric and 746 literal) with frequency 1. The most frequent nouns are *drone* (9) (metaphoric), and *year* (47) and *school* (22) (literal). About 10% (65) verb types and 4% (56) noun types are used both metaphorically and literally, indicating that features that combine information about the word and its context are needed.

	Metaphoric		**Literal**	
	V	N	V	N
Min	1	1	1	1
Max	22	9	86	47
Mean	1.8	1.37	2.12	1.86
No. of types	196	169	401	1118

Table 4: Frequency distribution of metaphoric and literal verb/noun types.

4 Experiments

The purpose of these experiments is to test whether there are distinguishable characteristics that indicate whether a metaphoric/literal word should/should not be changed to make the text easier to understand, and also whether there are features that are particular to metaphoric or literal words with respect to simplification. We conducted two sets of experiments: on the full dataset (metaphoric and literal items), and on the metaphoric part of the data. Through the experiments on the full dataset we investigate whether there are different features indicative of metaphor and literal word simplification, respectively. In the second set of experiments we perform a more in depth exploration of the metaphoric part of the data and look at the changes within the fine-grained simplification types.

4.1 Experimental setup

For the first set of experiments, we group the simplification types in two classes: *preserved* and *changed*. Unchanged items (i.e. *same metaphor* and *same non-metaphor*) were assigned the *preserved* class. All other simplification types were combined as *changed*. The quantitative information on the items used in the experiments is provided in Table 5.

Simplification type	Count
Preserved	2,232
metaphoric	299
non-metaphoric	1,933
Changed	1,277
metaphoric	285
non-metaphoric	992

Table 5: Statistics on the coarse-grained simplification types.

The experiments were done with a Linear Support Vector Machine classifier using 10-fold cross-validation startegy.[12] The feature values were standardized prior to the experiments.[13] We report the results of the random baseline, and the distribution of the different phenomena in the data.

4.2 Features

Data analysis has shown that both metaphors and literal words can be changed to help comprehension, and either can be replaced with metaphoric or literal expressions. To determine whether there are identifiable characteristics that could make this distinction automatable, we compile a number of features that have been shown to be useful for text simplification and metaphor identification. The metaphor-sensitive features are *Imageability, Concreteness, WordNet senses* and *word's context*; the

[12] We use the SVM implementation in *scikit-learn* (Pedregosa et al., 2011): `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`

[13] Standardization was performed with the StandardScaler in *scikit-learn*: `http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

general features are *part of speech, vector space word representations, Age of Acquisition, word frequency* and *Familiarity*. The feature types used and their coverage in our dataset are described below.

Part of speech: The part of speech (POS) tagging was done using the NLTK toolkit[14] (Bird et al., 2009). The POS tags were then manually corrected where necessary. The two possible values are noun and verb.

Vector space word representations: We obtained vector space representations for each annotated word using Google's pre-trained *word2vec* model (Mikolov et al., 2013).[15]

Word embeddings have been successfully used in metaphor identification (e.g. Dinh and Gurevych 2016; Gutiérrez et al. 2016) as well as in lexical simplification tasks (e.g. Glavaš and Štajner 2015; Glavaš and Vulić 2018).

Age of Acquisition: Age of Acquisition (AoA) ratings were obtained from the AoA norms database of 51,715 English words (Kuperman et al., 2012). AoA denotes the approximate age at which a word is learned. The simplified news articles used in this study are intended for classroom use by 9-10 year old children. Words usually acquired after this age should be more readily changed/removed in the simplified version.

We extracted the AoA ratings by matching both word forms and lemmas (e.g. noun *testing/testing* vs. verb *testing/test*).

Imageability, Familiarity and Concreteness: Imageability stands for the ability of a word to evoke mental images; Familiarity refers to the frequency of exposure to a word; Concreteness describes the level of abstraction associated with the concept a word represents. The connection of these variables to metaphor comprehension has been shown in multiple studies (e.g. Marschark et al. 1983; Paivio et al. 1968; Ureña and Faber 2010). Concrete words are more easily learned, processed and remembered than the abstract ones (Paivio et al., 1968). It is quite likely then that abstract words will be discarded during simplification. Marschark et al. (1983) found a link between high imageability and easier processing for certain metaphor types. These features were successfully used in lexical simplification (e.g. Jauhar and Specia 2012; Vajjala and Meurers 2014).

Imageability and Familiarity ratings were obtained from the MRC Psycholinguistic Database (Wilson, 1988). This database contains up to 26 (psycho)linguistic attributes for 150,837 words. Concreteness ratings were extracted from a collection of English Abstractness/Concreteness ratings (Köper and Schulte im Walde, 2017).

We extracted the values for the word forms if present in the databases and for the respective lemmas otherwise. For a number of words, the values are missing (see Table 6). De Hertog and Tack (2018) use the third and first quartile values for Imageability and Concreteness, respectively, following an assumption that rarer words tend to have lower imageability and concreteness, while Gooding and Kochmar (2018) use the null value. We decided to assign instead a "neutral" value: the median value for each feature based on the ratings in the MRC.

	Available	**Missing**	
Imag	2,288	1,221	(35%)
Fam	2,293	1,216	(35%)
Concr	3,509	0	(0%)

Table 6: Counts for Imageability, Familiarity and Concreteness ratings.

Word frequency: In lexical simplification systems, it is common to substitute infrequent words with their more frequent synonyms (e.g. De Belder and Moens, 2010). As Kriz et al. (2018), we assume that highly frequent words are easier to understand, whereas infrequent words are more difficult and therefore will be removed/changed in the process of simplification.

We use word frequency counts from the SUBTLEX$_{\text{US}}$ database (Brysbaert and New, 2009), a corpus of subtitles for American English of 51M words. The frequencies are given per million words. We extracted the values based on the word forms in our data (3,503 words); 6 words (.2%) have frequency 0.

WordNet sense: The WordNet (Fellbaum, 1998) sense feature approximates a word's meaning in context. The values are the synset numbers representing the sense of a word in the original sentence. MIPVU uses sense information and comparison with a "basic" sense of a word

[14] `https://www.nltk.org/`

[15] The model can be downloaded from here: `https://code.google.com/archive/p/word2vec/`

to assign metaphoricity. The WordNet sense number could be an indication whether a word is metaphoric or not: the first sense is the more frequent, and could thus be considered basic, while the higher the sense number, the more likely it could be that the word is used metaphorically.

We use a Lesk-like (Lesk, 1986) method to disambiguate a target word relative to WordNet: a vector representation of the context of an annotated word (i.e. V0 sentence) is compared to a representation for each of the word's definitions in WordNet. The representations are generated using Google's pre-trained *word2vec* model. The context and each definition are compared using Word Mover's Distance (Kusner et al., 2015).[16] We chose the synset number whose definition is most similar to the word's context. The lookup in WordNet was done based on the word forms and matching POS tags. For 9 words (.3%) not found in WordNet the values are missing.

Word's context: This feature reflects the discrepancy between the level of abstractness of a metaphoric word and its context. It was operationalized with ratings of Concreteness (Köper and Schulte im Walde, 2017) and Imageability from the MRC database.

Turney et al. (2011) have shown that a word's degree of abstractness, relative to the context it appears in, can be successfully used to distinguish between literal and metaphoric meanings. Broadwell et al. (2013) used Imageability ratings to discover metaphors based on the assumption that they stand out of their context as being highly imageable.

We considered a symmetrical seven-word window centered on the target word. A word w's Concreteness context (CC) value is computed as:

$$CC_w = C\lceil n/2 \rceil - \left(\sum_{i=1}^{\lfloor n/2 \rfloor} C_i + \sum_{i=\lceil n/2 \rceil + 1}^{n} C_i \right)$$

where n is the size of the window. The Imageability context (IC) is calculated in the same way.

In the computation we used the context words with available Concreteness and Imageability scores in the database. If ratings for the target word itself or for all context words were not found, the value for the feature was set to missing. The overview of the value counts is given in Table 7.

	Available	Missing	
IC	2,258	1,251	(36%)
CC	3,503	6	(.2%)

Table 7: Counts for the words' context features: Imageability context (IC) and Concreteness context (CC).

4.3 Experiment 1: Metaphoric vs. literal words

To assess the impact of the different features on predicting whether a word should or should not be changed, we group the features based on the type of information they capture:

- *IFC* (Imageability + Familiarity + Concreteness) – informative for metaphoric words
- *WN+IC/CC* (WordNet sense + word's context) – different aspect of metaphor relevance
- *Freq+AoA* (word frequency + Age of Acquisition) – relevant for both metaphoric and literal items

The F-score results on the full dataset (1,277 changed, 2,232 preserved instances) for different feature combinations are presented in Figure 1.[17]

AoA has the highest Precision for the class *changed* in both metaphoric and literal cases. This shows that whereas this feature might be good in accurately detecting items that need simplification, it does not differentiate between metaphoric and literal usages in the current setting. Previous studies have shown that some correlation exists between the AoA and frequency of usage (e.g. Ghyselinck et al., 2004), but in this case the AoA feature and the Frequency feature have different effects when used alone (see Figure 1). In particular, the Frequency feature is not useful to determine whether a word should be changed or not, contrary to our expectations.

We expected the "metaphor-specific" features (IFC) to have a higher impact on the metaphoric than on the literal words. When used alone they do lead to better prediction for changing metaphoric words compared to literal ones, but within the context of the full feature set, their impact is minimal (all, all-IFC/I/F/C). The imbalance in the data set could explain why, when using other features which can pick up on characteristics of literal

[16] We used the implementation in Gensim Python library (Řehůřek and Sojka, 2010).

[17] We report only the F-scores for this experiment due to space limitations.

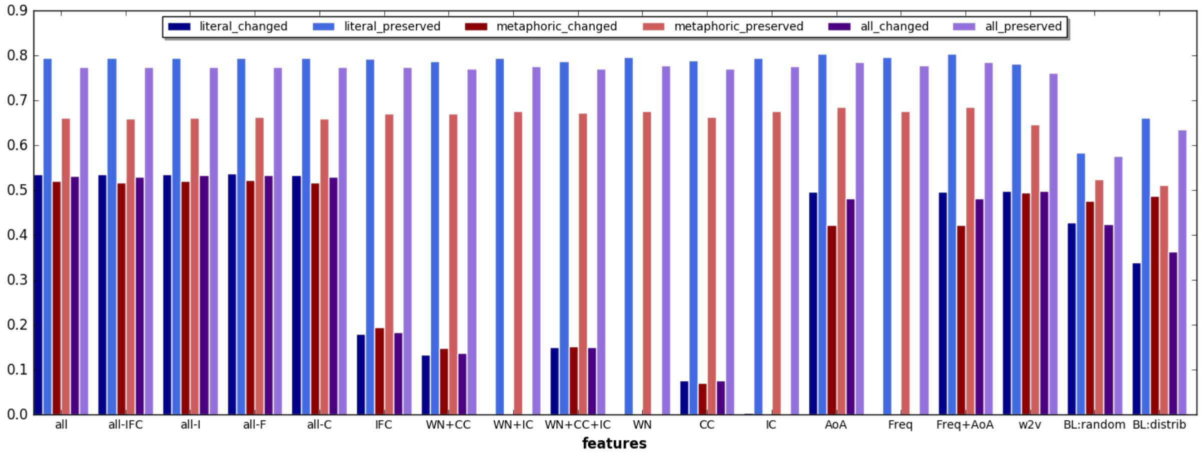

Figure 1: F-scores on learning changed vs. preserved on the full dataset, with results on all data and the metaphoric and literal subsets when using different feature combinations.

words (or both), the effect of Familiarity, Concreteness and Imageability is lost.

The WN and context features also behave in an interesting manner. Alone, neither of these has much impact on distinguishing words that should be changed or not (WN/CC/IC). But when combined, their predictive power grows, particularly considering the approximate 1:5 ratio between metaphoric and literal target words. We tested a binary representation of the WN feature: is the disambiguated sense the first one (the "basic" one) or not. This set-up led to worse results. This could mean that assuming that the first sense in WordNet is the "basic" sense is erroneous, even though it is the most frequent one.

	Changed		**Preserved**	
Feature(s)	Verbs	Nouns	Verbs	Nouns
all	.562	.515	.784	.770
IFC	.199	.175	.766	.779
WN+CC	.206	.093	.752	.780
WN+IC	.004	.000	.771	.779
WN+CC+IC	.220	.107	.750	.780
Freq+AoA	.480	.482	.803	.777
random	.424	.408	.523	.556
majority	.000	.000	.773	.780

Table 8: F-scores: changed vs. preserved on the full dataset, for nouns (833 changed, 1474 preserved) and verbs (444 changed, 758 preserved).

Looking at the results on the subsets corresponding to nouns and verbs (see Table 8), we note that there are differences in terms of the useful features. Predicting that nouns should be preserved is consistent w.r.t. the features used, and close to the majority baseline. Using all features leads to the best results overall, for both nouns and verbs, whether they should be changed or preserved. Metaphor-relevant features (IFC and contextual information) are not helpful in predicting verbs and nouns that need to be changed. However, they appear to be more relevant for verbs. The Frequency and Age of Acquisition combination seems to be more important for verbs than for nouns.

4.4 Experiment 2: Metaphoric words

We use the subset of 285 changed and 299 preserved metaphors to test the impact of different subsets of features for predicting change/preserve for metaphoric target words. The results are given in Table 9 for the complete metaphoric dataset.

We further analyze the results of classifying originally metaphoric words as changed or preserved in the simplified texts. We look into the data subsets corresponding to the different metaphor simplification phenomena, and produce the recall results shown in Table 10. We cannot compute precision because all instances in each subset belong to one class (i.e. either *changed* or *preserved*).

The results for the metaphoric data preserve some of the tendencies seen on the complete dataset, and they also reveal some new insights. AoA leads to the highest Precision score for the class *changed* and has high Recall and F-score for the class *preserved*. Frequency of use appears to be the most useful in distinguishing between metaphors that should be changed or preserved. This is quite intuitive, as metaphors that are less common are more difficult to understand. Contrary to its impact in the first experiment – classifying whether a word should be changed or not,

Feature(s)	Changed P	Changed R	Changed F_1	Preserved P	Preserved R	Preserved F_1
all	.573	.561	.567	.590	.602	.596
all-IFC	.568	.558	.563	.586	.595	.590
all-I	.568	.554	.561	.585	.599	.592
all-F	.571	.561	.566	.589	.599	.594
all-C	.573	.561	.567	.590	.602	.596
IFC	.576	.519	.546	.581	.635	.607
WN+CC	.522	.505	.513	.542	.559	.550
WN+IC	.513	.540	.526	.539	.512	.525
WN+CC+IC	.540	.519	.530	.558	.579	.568
WN	.497	.554	.524	.523	.465	.492
CC	.495	.319	.388	.515	**.689**	.589
IC	.438	.186	.261	.499	**.773**	.606
AoA	**.602**	.530	.563	.598	**.666**	**.630**
Freq	.548	**.884**	**.677**	**.734**	.304	.430
Freq+AoA	.594	.586	.590	.611	.619	.615
w2v	.576	.572	.574	.595	.599	.597
random	.489	.463	.476	.513	.538	.525
majority	.000	.000	.000	.512	1.00	.677

Table 9: Results on learning changed vs. preserved on the subset of metaphoric items (285 changed / 299 preserved instances). Best results are given in bold.

regardless of whether it is metaphoric or literal – when analyzing metaphoric words and classifying them into changed/preserved, Frequency is the best feature. This effect is apparent also when looking at the subsets corresponding to the different simplification types (see Table 10).

Feature(s)	Changed to: other met	phr. with met	lit.	phr. no met	rem.	Pres.: same
all	.651	.600	.515	.500	.570	.602
all-IFC	.674	.600	.515	.500	.551	.595
all-I	.651	.600	.505	.500	.651	.599
all-F	.698	.600	.505	.500	.561	.599
all-C	.585	.560	.582	.500	.570	.602
IFC	.535	**.850**	.545	.286	.548	.635
WN+CC	.581	.500	.525	.500	.458	.559
WN+IC	.605	.550	.515	.571	.533	.512
WN+CC+IC	.488	.650	.545	.500	.486	.579
WN	.581	.450	.545	**.643**	.561	.465
CC	.326	.450	.327	.286	.290	**.689**
IC	.186	.200	.188	.143	.187	**.773**
w2v	.674	.600	.495	.571	.598	.599
w2v+IFC	.674	.600	.495	.571	.589	.605
AoA	.605	.600	.535	.571	.477	.666
Freq	**.884**	**.950**	**.931**	**.929**	**.822**	.304
Freq+AoA	.651	.650	.604	.643	.523	.619
random	.419	.450	.475	.357	.514	.518

Table 10: Recall: changed vs. preserved on the subset of metaphoric items (285 changed / 299 preserved instances) for each fine-grained simplification type. Best results are given in bold.

The word's context features (IC/CC) have the highest Recall scores for the preserved cases, but in combination with the WordNet senses feature they stop being useful for differentiating between the two classes. Just as in the first experiment, when used alone the IFC features are clearly useful, but within the full set of features they lose their predictive power. For the preserved items, the context features (IC/CC) show the best results. Those metaphors that were rephrased with metaphorical content are best described with the IFC features, whereas the WN senses feature is good when identifying paraphrases without metaphors.

5 Conclusion

The analysis of metaphor usage in original and simplified versions of the same news texts has shown that not all metaphors are alike, from the point of view of text comprehension. A large percentage of metaphors in our dataset were either preserved or replaced using metaphorical language, while a (much) smaller number of literally used words was replaced with a metaphoric expression.

The evaluation of the features most frequently used in literature for text simplification and metaphor identification has shown that for both metaphors and literal words, the most informative feature is the Age of Acquisition. Features that capture the imageability, familiarity and concreteness of a word have similar performance in predicting change/no change for both metaphorical and literal words when used alone. When used together with our other features, their predictive power diminishes. While not useful to separate changed and preserved words in the full dataset, for metaphoric words the frequency of usage is a telling feature, even at a fine-grained level.

One factor that could have influenced the results of these experiments is the incomplete coverage provided by the Imageability and Familiarity features. In future work we plan to improve the assignment of missing values by deriving a value using the scores assigned to the most similar words. We will further explore features that capture the interaction between a target word and its context, including contextual embeddings and the word's syntactic role.

Acknowledgements

We thank the reviewers for the interesting comments. Yulia Clausen's research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 317633480 – SFB 1287, project A03.

References

G. Barlacchi and S. Tonelli. 2013. ERNESTA: A sentence simplification tool for children's stories in italian. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, pages 476–487, Berlin, Heidelberg. Springer-Verlag.

S. Bird, E. Klein, and E. Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

D. Bollegala and E. Shutova. 2013. Metaphor interpretation using paraphrases extracted from the web. *PLoS ONE*, 8(9):e74304.

G. A. Broadwell, U. Boz, I. Cases, T. Strzalkowski, L. Feldman, S. Taylor, S. Shaikh, T. Liu, K. Cho, and N. Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'13, pages 102–110, Berlin, Heidelberg. Springer-Verlag.

M. Brysbaert and B. 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, Instruments & Computers*, 41(4):977–990.

J. De Belder and M.-F. Moens. 2010. Text simplification for children. In *SIGIR: Workshop on Accessible Search Systems*, pages 19–26.

O. De Clercq and V. Hoste. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.

D. De Hertog and A. Tack. 2018. Deep learning architecture for complex word identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 328–334.

E.-L. Do Dinh and I. Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33.

B. Drndarević and H. Saggion. 2012. Towards automatic lexical simplification in Spanish: An empirical study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 8–16.

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

M. Ghyselinck, M. B. Lewis, and M. Brysbaert. 2004. Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multitask investigation. *Acta psychologica*, 115 (1):43–67.

G. Glavaš and S. 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)*, pages 63–68.

G. Glavaš and I. Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–45.

A. Golden. 2010. Grasping the point: A study of 15-year-old students' comprehension of metaphorical expressions in schoolbooks. In Graham Low, Zazie Todd, Alice Deignan, and Lynne Cameron, editors, *Researching and Applying Metaphor in the Real World*, pages 35–62. John Benjamins B. V.

S. Gooding and E. Kochmar. 2018. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.

E. D. Gutiérrez, E. Shutova, T. Marghetis, and B. Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193.

C. Horn, C. Manduca, and D. Kauchak. 2014. Learning a lexical simplifier using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463.

S. K. Jauhar and L. Specia. 2012. UOW-SHEF: SimpLex - lexical simplicity ranking based on contextual and psycholinguistic features. In **SEM 2012: 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 2012)*, pages 477–481.

M. Köper and S. Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30.

Z. Kövecses. 2017. *The Routledge Handbook of Metaphor and Language*, chapter Conceptual Metaphor Theory. Routledge.

R. Kriz, E. Miltsakaki, M. Apidianaki, and Ch. Callison-Burch. 2018. Simplification using paraphrases and context-based lexical substitution. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 207–217.

V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. 2015. From word embeddings to document distance. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, volume 37, pages 957–966.

G. Lakoff and M. Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26.

M. Marschark, A. N. Katz, and A. Paivio. 1983. Dimensions of metaphor. *Journal of Psycholinguistic Research*, 12(1):17–40.

T. Mikolov, G. Corrado, K. Chen, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, pages 1–12.

S. Nisioi, S. Štajner, S. Paolo Ponzetto, and L. P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91.

G. Paetzold and L. Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

A. Paivio, J. C. Yuille, and S. A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, Monograph Supplement*, 76(3, Pt. 2):1–25.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

R. Řehůřek and P. Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

E. Shutova. 2013. Metaphor identification as interpretation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 276–285.

G. Steen, L. Dorst, J. B. Herrmann, A. Kaal, T. Krennmayr, and T. Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins B.V.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

D. Torunoğlu-Selamet, T. Pamay, and G. Eryiğit. 2016. Simplification of Turkish sentences. In *Proceedings of The First International Conference on Turkic Computational Linguistics*, pages 55–59.

P. D. Turney, Y. Neuman, D. Assaf, and Y. Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690.

J. M. Ureña and P. Faber. 2010. Reviewing imagery in resemblance and non-resemblance metaphors. *Cognitive Linguistics*, 21:123–149.

S. Vajjala and D. Meurers. 2014. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Recent Advances in Automatic Readability Assessment and Text Simplification*, 165(2):194–222.

S. Štajner, S. Paolo Ponzetto, and H. Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 4096–4102.

Tu Thanh Vu, Giang Binh Tran, and Son Bao Pham. 2014. Learning to simplify children stories with limited data. In *9th Asian Conference on Intelligent Information and Database Systems*, pages 31–41.

M. Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

M. Wolska and Y. Clausen. 2017. Simplifying metaphorical language for young readers: A corpus study on news text. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 313–318.

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

W. Xu, C. Napoles, E. Pavlick, Q. Chen, and Ch. Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

S. M. Yimam, Ch. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78.

Z. Zhu, D. Bernhard, and I. Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.