Detecting Malapropisms Using Measures of Contextual Fitness

Torsten Zesch

Ubiquitous Knowledge Processing Lab
Department of Computer Science, Technische Universität Darmstadt, Germany

German Institute for International Educational Research
Frankfurt, Germany

ABSTRACT. *While detecting simple language errors (e.g. misspellings, number agreement, etc.) is nowadays standard functionality in all but the simplest text-editors, other more complicated language errors might go unnoticed. A difficult case are errors that come in the disguise of a valid word that fits syntactically into the sentence. We use the Wikipedia revision history to extract a dataset with such errors in their context. We show that the new dataset provides a more realistic picture of the performance of contextual fitness measures. The achieved error detection quality is generally sufficient for competent language users who are willing to accept a certain level of false alarms, but might be problematic for non-native writers who accept all suggestions made by the systems. We make the full experimental framework publicly available which will allow other scientists to reproduce our experiments and to conduct follow-up experiments.*

RÉSUMÉ. *Alors que la détection d'erreurs simples est aujourd'hui une fonctionnalité standard des traitements de texte un peu évolués, de nombreuses erreurs restent difficiles à repérér. C'est souvent le cas lorsque la forme correcte est remplacée par une autre forme valide et syntaxiquement plausible en contexte. Nous avons utilisé les révisions de Wikipédia pour extraire automatiquement une listes d'erreurs de ce type. Ces données permettent de se faire une meilleure idée de l'utilité réelle des indicateurs standard de conformité contextuelle, qu'ils soient linguistiques ou statistiques. Les taux de détection obtenus sont généralement suffisants pour des scripteurs compétents qui seraient prêts à accepter un certain niveau de fausses alarmes ; ils restent problématiques pour des scripteurs nonnatifs. L'ensemble du dispositif expérimental utilisé pour ce travail sera rendu public, ce qui permettra à d'autres chercheurs de reproduire nos expériences et d'approfondir nos résultats.*

KEYWORDS: *contextual fitness, malapropisms, Wikipedia revision history.*

MOTS-CLÉS : *conformité contextuelle, malapropisme, révisions de Wikipédia.*

1. Introduction

Measuring the contextual fitness of a term in its context is a key component in different NLP applications like speech recognition (Inkpen and Désilets, 2005), optical character recognition (Wick *et al.*, 2007), co-reference resolution (Bean and Riloff, 2004), or malapropism detection (Bolshakov and Gelbukh, 2003). The main idea is to test what fits better into the current context: the actual term or a possible replacement that is phonetically, structurally, or semantically similar.

In this article, we focus on the detection and correction of language errors. Whether or not context is needed for this purpose depends on the type of language error. English non-words like *nnie werds* can be easily detected without taking any context into account. As soon as we go beyond detection and also want to correct an error, context is needed as usually there is a set of possible corrections and we want to pick the one that fits the context best. For example, the error *nnie werds* could be corrected to either *nine wards*, *non-words*, or *nine words* depending on the context in which the error is observed.

A special class of errors are *real-word spelling errors* that come in the disguise of an existing word, like *My farther [father] was a baker.* or *What are [is] the price?* In both examples, the error can be detected on a syntactic level without looking at the meaning of the sentence, while in the sentence *People with lots of honey [money] live in big houses*, both words are syntactically valid, but one is much more likely in this context. The subset of real-word spelling errors that can only be detected on the semantic level is called *malapropisms* (Hirst and St-Onge, 1998). They can only be detected by looking at the semantic fitness of the word and its context.

If we already know the problematic words, supervised contextual fitness measures (Golding and Schabes, 1996; Jones and Martin, 1997; Carlson *et al.*, 2001) can be used. They are based on confusion sets, i.e. sets of words that are often confounded e.g. {peace, piece} or {weather, whether}, and recently the approach has also been applied to article and preposition errors (Han *et al.*, 2006; Tetreault *et al.*, 2010). Given a large corpus, the measure learns a model of which alternative from the confusion set is more likely in the current context. This approach yields very high precision, but only for the limited number of previously defined confusion sets. As a consequence, all the problematic cases have to be known in advance. For example, {honey, money} from the example above is not in any list of notorious confusables. We are probably facing a very long tail of infrequent malapropisms that needs to be tackled in an unsupervised way. This will be the main focus of this article.

The article is organized as follows: in section 2, we tackle the problem that only very small evaluation datasets for malapropism detection are available. We describe a method to create a dataset of naturally-occurring malapropisms by mining the Wikipedia revision history in section 2. In section 3, we give an overview of two unsupervised approaches: the statistical (Mays *et al.*, 1991; Wilcox-O'Hearn *et al.*, 2008) and the knowledge-based (Hirst and Budanitsky, 2005) one. In section 4, we present a comprehensive evaluation of contextual fitness measures showing that statistical and knowledge-based approaches can be combined in order to obtain better performance. We also describe the results obtained when participating in the pilot round of the Helping Our Own (HOO) Shared Task (Dale and Kilgarriff, 2011) which focuses on the detection and correction of errors in scientific documents (section 5). In order to allow the community to easily reproduce our results, we developed a comprehensive and flexible open-source framework for unsupervised and supervised error correction based on measures of contextual fitness which we describe in section 6.

2. Creating a Dataset of Naturally-Occurring Malapropisms

So far, evaluation of contextual fitness measures has relied on datasets with artificial errors (Mays *et al.*, 1991; Hirst and Budanitsky, 2005) which can be quite easily created from a corpus that is known to be free of spelling errors. For a certain amount of randomly chosen sentences from the corpus, a random word is selected and all strings with edit distance equal to a certain value are generated. If one of those generated strings is a known word from the vocabulary, it is picked as the artificial error. Creating artificial datasets in such a way has several disadvantages: (i) the artificial replacement might be a synonym of the original word and perfectly valid in the given context, (ii) the generated error might be very unlikely to be made by a human, and (iii) inserting artificial errors often leads to unnatural sentences that are quite easy to correct, e.g. if the word class has changed. However, even if the word class is unchanged, the original word and its replacement might still be variants of the same lemma, e.g. a noun in singular and plural, or a verb in present and past form. This usually leads to a sentence where the error can be easily detected using syntactical or statistical methods, but is almost impossible to detect for knowledge-based measures of contextual fitness, as the meaning of the word stays more or less unchanged. To estimate the impact of this issue, we analyzed 1,000 artificially created errors, and found 387 singular/plural pairs and 57 pairs which were in another direct relation (e.g. adjective/adverb). Such pairs are easy to detect using a statistical approach, as the resulting n-gram is often very infrequent, e.g. *see a houses* vs. *see a house*. Such a biased dataset is certainly not suited for a fair evaluation targeted at finding good measures of contextual fitness for malapropism detection.

Previous work on evaluating real-word spelling correction (Hirst and Budanitsky, 2005; Wilcox-O'Hearn *et al.*, 2008; Islam and Inkpen, 2009) used a dataset sampled from the *Wall Street Journal* corpus which is not freely available. Thus, we created a comparable English dataset of 1,000 artificial errors based on the freely available Brown corpus (Francis and Kuçera, 1964).[1] Additionally, we created a German dataset with 1,000 artificial errors based on the TIGER corpus.[2]

[1] `http://www.archive.org/details/BrownCorpus` (CC-by-na).

[2] `http://www.ims.uni-stuttgart.de/projekte/TIGER/`
The corpus contains 50,000 sentences of German newspaper text, and is freely available under a non-commercial license.

Unfortunately, there are very few sources of sentences with naturally-occurring malapropisms and their corrections. Recently, the revision history of Wikipedia has been introduced as a valuable knowledge source for NLP (Nelken and Yamangil, 2008; Yatskar *et al.*, 2010). We propose it as a possible source of naturally-occurring malapropisms, as it is likely that Wikipedia editors introduce such errors at some point, which are then corrected in subsequent revisions of the same article. The challenge lies in discriminating malapropisms from all sorts of other changes, including non-word spelling errors, reformulations, or the correction of wrong facts. For that purpose, we apply a set of precision-oriented heuristics narrowing down the number of possible error candidates. Such an approach is feasible, as the high number of revisions in Wikipedia allows us to be extremely selective.

2.1. *Mining the Wikipedia Revision History*

We access the Wikipedia revision data using the freely available Wikipedia Revision Toolkit (Ferschke *et al.*, 2011) together with the JWPL Wikipedia API (Zesch *et al.*, 2008a).[3] It outputs plain text converted from Wiki-Markup, but the text still contains a small portion of left-over markup and other artifacts. Thus, we perform additional cleaning steps removing (i) tokens with more than 30 characters (often URLs), (ii) sentences with less than 5 or more than 200 tokens, and (iii) sentences containing a high fraction of special characters like ":", which usually indicate Wikipedia-specific artifacts like lists of language links. The remaining sentences are part-of-speech tagged and lemmatized using TreeTagger (Schmid, 2004). Based on these cleaned and annotated articles, we form pairs of adjacent article revisions.

Sentence Alignment Fully aligning all sentences of the adjacent revisions is a quite costly operation, as sentences can be split, joined, replaced, or moved in the article. However, we are only looking for sentence pairs which are almost identical except for the malapropism. Thus, we form all sentence pairs and then apply an aggressive but cheap filter that rules out all sentences which (i) are equal, or (ii) whose lengths differ more than a small number of characters. For the resulting much smaller subset of sentence pairs, we compute the Jaro similarity (Jaro, 1995) between each pair. If the similarity is lower than a certain threshold t_{sim} (0.95 in this case), we do not further consider the pair. The small amount of remaining pairs is passed to the sentence pair filter for in-depth inspection.

Sentence Pair Filtering The sentence pair filter further reduces the number of remaining sentence pairs by applying a set of heuristics including *surface level* and *semantic level* filters. Surface level filters include:

– *Replaced Token* Sentences need to consist of identical tokens, except for one replaced token;

– *No Numbers* The replaced token may not be a number;

[3] `http://code.google.com/p/jwpl/`

– *UPPER CASE* The replaced token may not be in upper case;

– *Case Change* The change should not only involve case changes, e.g. changing *english* into *English*;

– *Edit Distance* The edit distance between the replaced token and its correction needs to be below a certain threshold.

After applying the surface level filters, the remaining sentence pairs are well-formed and contain exactly one changed token at the same position in the sentence. However, the change does not need to characterize a malapropism, but could also be a normal spelling error or a semantically motivated change. Thus, we apply a set of semantic filters:

– *Vocabulary* The replaced token needs to occur in the vocabulary. We found that even quite comprehensive word lists discarded too many valid errors, as Wikipedia contains articles from a very wide range of domains. Thus, we use a frequency filter based on the Google Web1T n-gram counts (Brants and Franz, 2006). We filter all sentences where the replaced token has a low unigram count. We experimented with different values and found 25,000 for English and 10,000 for German to yield good results;[4]

– *Same Lemma* The original token and the replaced token must not have the same lemma, e.g. *car* and *cars* would not pass this filter;

– *Stopwords* The replaced token must not be in a short list of stopwords (mostly function words);

– *Named Entity* The replaced token must not be part of a named entity. For this purpose, we applied the Stanford NER (Finkel *et al.*, 2005);

– *Normal Spelling Error* We apply the Jazzy spelling detector[5] and rule out all cases in which it is able to detect the error;

– *Semantic Relation* If the original token and the replaced token are in a close lexical-semantic relations, the change is likely to be semantically motivated, e.g. if *house* was replaced with *hut*. Thus, we do not consider cases where we detect a direct semantic relation between the original and the replaced term. For this purpose, we use WordNet (Fellbaum, 1998) for English and GermaNet (Lemnitzer and Kunze, 2002) for German.

2.2. *Resulting Datasets*

Using our framework for mining malapropisms in context, we extracted an English and a German dataset.[6] Although the output generally was of high quality, manual

[4] The absolute thresholds correspond to a relative frequency of about 10^{-8} for both languages.

[5] `http://jazzy.sourceforge.net/`

[6] Using an English dump from April 5, 2011 and a German dump from August 13, 2010.

post-processing was necessary[7], as (i) for some pairs the available context did not provide enough information to decide which form was correct, and (ii) a problem that might be specific to Wikipedia – vandalism. The revisions are full of cases where a word is replaced with a similar sounding but greasy alternative. A relatively mild example is *In romantic comedies, there is a love story about a man and a woman who fall in love, along with silly or funny comedy farts [parts]*, where *parts* was replaced with *farts* only to be changed back shortly afterwards by a Wikipedia vandalism hunter. We removed all cases that resulted from obvious vandalism. For further experiments, a small list of offensive terms could be added to the stopword list to facilitate this process.

A related problem is correct words that get falsely corrected by Wikipedia editors (without the malicious intent from the previous examples, but with similar consequences). For example, the initially correct sentence *Dung beetles roll it into a ball, sometimes being up to 50 times their own weight* was "corrected" by exchanging *weight* with *wait*. We manually removed such obvious mistakes, but are still left with some borderline cases. In the sentence *By the 1780s the goals of England were so full that convicts were often chained up in rotting old ships*, the obvious error *goal* was changed by some Wikipedia editor to *jail*. However, actually it should have been the old English form for "jail" *gaol* which can be deduced when looking at the full context and later versions of the article. We decided to not remove these rare cases, because *jail* is a valid correction in this context.

After manual inspection, in which we had to remove about half of the cases returned by the heuristics described above, we are left with 466 English and 200 German malapropisms. Given that we restricted our experiment to 5 million English and German revisions, much larger datasets can be extracted if the whole revision history is taken into account. Our snapshot of the English Wikipedia contains 305 million revisions. Even if not all of them correspond to article revisions, it is safe to assume that more than 10,000 English malapropisms can be extracted from this version of Wikipedia using our methodology.

Using the same amount of source revisions, we found significantly more English than German malapropisms. This might be due to (i) English having more short nouns or verbs than German which are more likely to be confused with each other, and (ii) the English Wikipedia being known to attract a larger amount of non-native editors which might lead to higher rates of malapropisms. However, this issue needs to be further investigated, e.g. based on comparable corpora built on the basis of different language editions of Wikipedia. Further refining the identification of malapropisms in Wikipedia would allow evaluating how frequently such errors actually occur, and how long it takes the Wikipedia editors to detect them. A remaining problem is the

[7] The most efficient and precise way of finding malapropisms would of course be to apply measures of contextual fitness. However, the resulting dataset would then only contain errors that are detectable by the measures we want to evaluate – a clearly unacceptable bias. Thus, a certain amount of manual validation is inevitable.

boast / boost	foaming / forming	racial / radical
cape / cane	investing / inverting	remaining / renaming
cartridge / cartilage	irritation / irrigation	retried / retired
complied / compiled	laces / places	signer / singer
conference / confluence	layers / lawyers	tracks / tracts
confined / confided	mowing / moving	vane / vein
desserts / deserts	principal / principle	vehicles / vesicles

Table 1. *Examples of English confusables mined from the Wikipedia revision history*

activity of Wikipedia bots, i.e. autonomous agents that perform various tasks triggered by previous edits. Currently, these bots only target out-of-vocabulary spelling errors and some very common confusion sets, but as Hirst and Budanitsky (2005) note, a major source of malapropisms is the failed attempt of automatic spelling correctors to correct a misspelled word. Thus, the frequency of malapropisms in Wikipedia might be artificially high due to the activity of bots, but this remains to be investigated. Anyway, no matter whether the malapropisms are introduced by editors or by robots, they still need to be detected using measures of contextual fitness.

Note that once we have extracted a malapropism from the Wikipedia revision history which is not yet in the list of known confusables, we could train a supervised classifier based on the newly discovered confusion set in order to gain improved quality in error detection and correction. Table 1 lists some examples of extracted confusables. Only 36 out of 466 pairs occur more than once in the English dataset, i.e. we will always have a long tail of previously unseen examples which need to be detected at least once using unsupervised measures of contextual fitness.

Another interesting observation is that the average edit distance is around 1.4 for both datasets. This means that a substantial proportion of malapropisms involve more than one edit operation. Given that many measures of contextual fitness allow at most one edit, many naturally-occurring malapropisms will not be detected. However, allowing a larger edit distance enormously increases the search space resulting in increased run-time and possibly decreased detection precision due to more false positives. A solution might be not to rely on simple edit distance, but to use the knowledge about malapropisms gained from Wikipedia in order to learn a better model of what kind of words are being confused.

Related Work To our knowledge, we are the first to create a dataset of naturally-occurring malapropisms based on the revision history of Wikipedia. Max and Wisniewski (2010) used similar techniques to create a dataset of errors from the French Wikipedia. However, they target a wider class of errors including non-word spelling errors, and their class of real-word errors conflates malapropisms as well as other types of changes such as reformulations. Thus, their dataset cannot be easily used for our purposes and is only available in French, while our framework allows for creating datasets for all major languages with minimal manual effort. Bronner and Monz (2012) also use the Wikipedia revision history to classify edits into “factual” or "fluency", but it would be quite difficult to use the same techniques to classify malapropisms vs. non-malapropisms. Another possible source of malapropisms are learner corpora (Granger, 2002), e.g. the Cambridge Learner Corpus (Nicholls, 1999). However, learners are likely to make different mistakes than proficient language users, only a small fraction of observed errors will be malapropisms, and annotation of such errors is difficult and costly (Rozovskaya and Roth, 2010).

Now that we have created evaluation datasets containing artificial and naturally-occurring malapropisms, we use them for the evaluation of statistical and knowledge-based measures of contextual fitness.

3. Contextual Fitness Measures

Existing unsupervised measures of contextual fitness can be categorized into knowledge-based (Hirst and Budanitsky, 2005) and statistical methods (Mays *et al.*, 1991; Wilcox-O'Hearn *et al.*, 2008). Both test the lexical cohesion of a word with its context. For that purpose, knowledge-based approaches employ the structural knowledge encoded in lexical-semantic networks like WordNet (Fellbaum, 1998), while statistical approaches rely on n-gram counts collected from large corpora, e.g. the Google Web1T corpus (Brants and Franz, 2006).

3.1. *Statistical Approach*

Mays *et al.* (1991) introduced an approach based on the noisy-channel model. The model assumes that the correct sentence s is transmitted through a noisy channel adding "noise" which results in a word w being replaced by an error e leading to the wrong sentence s' which we observe. The probability of the correct word w given that we observe the error e can be computed as $P(w|e) = P(w) \cdot P(e|w)$. The channel model $P(e|w)$ describes how likely the typist is to make an error. This is modeled by the parameter α. The remaining probability mass $(1-\alpha)$ is distributed equally among all words in the vocabulary within an edit distance of 1. We refer to this set of words as $edits(w)$.

$$P(e|w) = \begin{cases} \alpha & \text{if } e = w \\ (1-\alpha)/|edits(w)| & \text{if } e \neq w \end{cases}$$

The source model $P(w)$ is estimated using a trigram language model, i.e. the probability of the intended word w_i is computed as the conditional probability $P(w_i|w_{i-1}w_{i-2})$. Hence, the probability of the correct sentence $s = w_1 \ldots w_n$ can be estimated as

$$P(s) = \prod_{i=1}^{n+2} P(w_i|w_{i-1}w_{i-2})$$

The set of candidate sentences S_c contains all versions of the observed sentence s' derived by replacing one word with a word from $edits(w)$, while all other words in the sentence remain unchanged. The correct sentence s is that sentence from S_c that maximizes $P(s|s') = \arg\max_{s \in S_c} P(s) \cdot P(s'|s)$.

3.2. *Knowledge-Based Approach*

Hirst and Budanitsky (2005) introduced a knowledge-based approach that detects real-word spelling errors by checking the semantic relations of a target word with its context. For this purpose, they apply WordNet as the source of lexical-semantic knowledge. The algorithm flags all words as error candidates and then applies filters to remove those words from further consideration that are unlikely to be errors. First, the algorithm removes all closed-class word candidates as well as candidates which cannot be found in the vocabulary. Candidates are then tested for having lexical cohesion with their context, by (i) checking whether the same surface form or lemma appears again in the context, or (ii) a semantically related concept is found in the context. In both cases, the candidate is removed from the list of candidates. For each remaining possible real-word spelling error, edits are generated by inserting, deleting, or replacing characters up to a certain edit distance (usually 1). Each edit is then tested for lexical cohesion with the context. If at least one of them fits into the context, the candidate is selected as a real-word error.

Hirst and Budanitsky (2005) use two additional filters: first, they remove candidates that are “common non-topical words”. However, it is unclear how the list of such words was compiled. Their list of examples contains words like *find* or *world* which we consider to be perfectly valid candidates. Second, they also applied a filter using a list of known multi-words, as the probability for words to accidentally form multi-words is low. It is unclear which list was used. We could use multi-words from WordNet, but coverage would be rather limited. We decided not to use these filters in order to better assess the influence of the underlying semantic relatedness measure on the overall performance. Such semantic relatedness measures are applied in order to determine the cohesion between a candidate and its context. In the experiments by Budanitsky and Hirst (2006), the measure by Jiang and Conrath (1997) yields the best results. However, a wide range of other measures have been proposed, cf. (Zesch and Gurevych, 2010). For example, some measures use a wider definition of semantic relatedness (Gabrilovich and Markovitch, 2007; Zesch *et al.*, 2008b) instead of only taxonomic relations in a knowledge source. As measures of semantic relatedness usually return a numeric value, we need to determine a threshold θ in order to come up with a binary related/unrelated decision. Budanitsky and Hirst (2006) used a characteristic gap in the standard evaluation dataset by Rubenstein and Goodenough (1965) that separates unrelated from related word pairs. We do not follow this approach, but optimize the threshold on a held-out development set of real-word spelling errors.

Dataset		P	R	F
English	Artificial	.77	.50	.60
	Natural	.54	.26	.35
German	Artificial	.90	.49	.63
	Natural	.77	.20	.32

Table 2. *Detection results of the statistical approach using a trigram model based on Google Web1T*

4. Evaluation

In this section, we report on the results obtained in our evaluation of contextual fitness measures using artificial and natural malapropisms in English and German. In our analysis, we focus on the *detection* of malapropisms, as it is more important than *correction* for two main reasons: first, in order to provide a correction, an error needs to be detected first. Second, once the user has been notified about a possible error, she usually knows the intended meaning and can correct the error without the need for further suggestions. [8]

4.1. *Statistical Approach*

Table 2 summarizes the results obtained by the statistical approach using a trigram model based on the Google Web1T data (Brants and Franz, 2006). On the English artificial errors, we observe a quite high F-measure of .60 that drops to .35 when switching to the naturally-occurring malapropisms which we extracted from Wikipedia. On the German dataset, we observe almost the same performance drop (from .63 to .32).

These observations correspond to our earlier analysis where we showed that the artificial data contains many cases that are quite easy to correct using a statistical model, e.g. where a plural form of a noun is replaced with its singular form (or vice versa) as in *I bought a cars [car].* The naturally-occurring malapropisms often contain much harder contexts, as shown in the following example: *Through the open window they heard sounds below in the street: cartwheels, a tired horse's plodding step, vices [voices].* While *voices* is clearly semantically related to other words in the context like *hear* or *sound*, the position at the end of the sentence is especially difficult for the statistical approach. The only trigram that connects the error to the context is "step , vices [voices]" which will yield a low frequency count even for very large trigram models. Higher order n-gram models could help to some extent, but suffer from the usual data-sparseness problems.

[8] An important exception are language learners which might take the erroneous suggestions of an automated system for granted.

Dataset	N-gram model	Size	P	R	F
Artificial-English	Google Web1T	$7 \cdot 10^{11}$	.77	.50	.60
		$7 \cdot 10^{10}$	.78	.48	.59
		$7 \cdot 10^{9}$	.76	.42	.54
	Wikipedia	$2 \cdot 10^{9}$	.72	.37	.49
Natural-English	Google Web1T	$7 \cdot 10^{11}$	.54	.26	.35
		$7 \cdot 10^{10}$	.51	.23	.31
		$7 \cdot 10^{9}$	.46	.19	.27
	Wikipedia	$2 \cdot 10^{9}$	.49	.19	.27
Artificial-German	Google Web1T	$8 \cdot 10^{10}$	.90	.49	.63
		$8 \cdot 10^{9}$	.90	.47	.61
		$8 \cdot 10^{8}$	.88	.36	.51
	Wikipedia	$7 \cdot 10^{8}$	.90	.37	.52
Natural-German	Google Web1T	$8 \cdot 10^{10}$	.77	.20	.32
		$8 \cdot 10^{9}$	.68	.14	.23
		$8 \cdot 10^{8}$	.65	.10	.17
	Wikipedia	$7 \cdot 10^{8}$	.70	.13	.22

Table 3. *Influence of the n-gram model on the detection quality of the statistical approach*

For building the trigram model, we used the Google Web1T data, which has some known quality issues and is not targeted towards the Wikipedia articles from which we sampled the natural errors. Thus, we also tested a trigram model based on Wikipedia. As it is much smaller than the Web model, we also created smaller Web models in order to evaluate the influence of the model size. Table 3 summarizes the results. We observe that "more data is better data" still holds, as the largest Web model always outperforms the Wikipedia model in terms of recall. If we reduce the size of the Web model to the same order of magnitude as the Wikipedia model, the performance of the two models is comparable. We would have expected to see better results for the Wikipedia model in this setting, but its higher quality does not lead to a significant difference.

Islam and Inkpen (2009) presented another statistical approach using the Google Web1T data (Brants and Franz, 2006) to create the n-gram model. It slightly outperformed the approach by Mays *et al.* (1991) when evaluated on a corpus of artificial errors based on the WSJ corpus. However, the results are not directly comparable, as Mays *et al.* (1991) used a much smaller n-gram model and our results show that the size of the n-gram model has a large influence on the results.

Even if statistical approaches quite reliably detect real-word spelling errors, the size of the required n-gram models remains a serious obstacle for use in real-world applications. The English Web1T trigram model is about 25GB, which currently is not

Dataset		P	R	F
English	Artificial	.35	.18	.24
	Natural	.32	.18	.23
German	Artificial	.34	.15	.21
	Natural	.38	.18	.25

Table 4. *Detection results of the knowledge-based approach using the JiangConrath semantic relatedness measure*

suited for being applied in settings with limited storage capacities, e.g. for intelligent input assistance in mobile devices. As we have seen above, using smaller models will decrease recall to a point where hardly any error will be detected anymore. Thus, we will now have a look on knowledge-based approaches which are less demanding in terms of the required resources.

4.2. *Knowledge-Based Approach*

Table 4 shows the results for the knowledge-based approach using the JiangConrath (Jiang and Conrath, 1997) relatedness measure. In contrast to the statistical approach, the results on the artificial errors are not higher than on the natural errors; another piece of evidence supporting our view that the properties of artificial datasets over-estimate the performance of statistical measures. In a re-evaluation of the statistical model, Wilcox-O'Hearn *et al.* (2008) found that it outperformed the knowledge-based method by Hirst and Budanitsky (2005) when evaluated on a corpus of artificial errors based on the WSJ corpus. [9] This is consistent with our findings.

As was pointed out before, Budanitsky and Hirst (2006) show that the measure by Jiang and Conrath (1997) yields the best results in their experiments on malapropism detection. In a similar fashion, we test another path-based measure by Lin (1998), the gloss-based measure by Lesk (1986), and the ESA measure (Gabrilovich and Markovitch, 2007) based on concept vectors from Wikipedia, Wiktionary, and WordNet. Table 5 summarizes the results. In contrast to the findings of Budanitsky and Hirst (2006), we could not find significant differences between the path-based measures. Even more importantly, other (non path-based) measures yield higher precision (at comparable recall levels) than any path-based measures. Especially ESA based on Wiktionary provides a comparatively good overall performance for English, while ESA based on Wikipedia provides good precision for both languages. The performance of ESA can be explained with its ability to incorporate semantic relationships beyond classical taxonomic relations (as used by path-based measures).

[9] They also tried to improve the model by permitting multiple corrections and using fixed-length context windows instead of sentences, but obtained discouraging results.

Dataset	Measure	θ	P	R	F
Artificial-English	JiangConrath	0.5	.35	.18	.24
	Lin	0.5	.27	.18	.21
	Lesk	0.5	.23	.18	.20
	ESA-Wikipedia	0.05	**.47**	.14	.21
	ESA-Wiktionary	0.05	.34	**.22**	**.27**
	ESA-Wordnet	0.05	.32	.17	.22
Natural-English	JiangConrath	0.5	.32	.18	.23
	Lin	0.5	.30	**.21**	.25
	Lesk	0.5	.22	.18	.20
	ESA-Wikipedia	0.05	**.49**	.15	.22
	ESA-Wiktionary	0.05	.35	**.21**	**.26**
	ESA-Wordnet	0.05	.32	.15	.21
Artificial-German	JiangConrath	0.01	.34	.15	.21
	Lin	0.01	.24	.13	.17
	Lesk	0.01	.38	.10	.16
	ESA-Wikipedia	0.05	**.53**	**.17**	**.25**
	ESA-Wiktionary	0.05	.49	.16	.24
	ESA-GermanNet	0.05	.33	.08	.13
Natural-German	JiangConrath	0.01	.38	**.18**	**.25**
	Lin	0.01	.33	.10	.15
	Lesk	0.01	.36	.08	.13
	ESA-Wikipedia	0.05	**.56**	.16	**.25**
	ESA-Wiktionary	0.01	.48	.13	.20
	ESA-GermaNet	0.01	.34	.15	.20

Table 5. *Detection results of knowledge-based approach using different relatedness measures*

4.3. *Combining the Approaches*

The statistical and the knowledge-based approach use quite different methods to assess the contextual fitness of a word in its context. This makes it worthwhile to combine both approaches. We ran the statistical method (using the full Wikipedia trigram model) and the knowledge-based method (using the ESA-Wiktionary relatedness measure) in parallel and then combined the resulting detections using two strategies: (i) we merge the detections of both approaches in order to obtain higher recall ("Union"), and (ii) we only count an error as detected if both methods agree on a detection ("Intersection"). When comparing the combined results in Table 6 with the best precision or recall obtained by a single approach ("Best-Single"), we observe that recall can be significantly improved using the "Union" strategy, while precision is only moderately improved using the "Intersection" strategy. This means that (i) a large subset of malapropisms is detected by both approaches which – due to their different sources

Dataset	Comb.-Strategy	P	R	F
Artificial-English	Best-Single	.77	.50	.60
	Union	.52	**.58**	.55
	Intersection	**.91**	.16	.27
Natural-English	Best-Single	.54	.26	.35
	Union	.38	**.36**	.37
	Intersection	**.81**	.11	.19
Artificial-German	Best-Single	.90	.49	.63
	Union	.77	**.51**	.61
	Intersection	**.93**	.06	.11
Natural-German	Best-Single	.77	.20	.32
	Union	.59	**.23**	.33
	Intersection	**.89**	.04	.08

Table 6. *Detection results obtained by a combination of the best statistical and knowledge-based configuration. "Best-Single" is the best precision or recall obtained by a single approach. "Union" merges the detections of both approaches. "Intersection" only detects an error if both methods agree on a detection*

of knowledge – mutually reinforce the detection leading to increased precision, and (ii) a small but otherwise undetectable subset of malapropisms requires considering detections made by one approach only.

5. HOO Shared Task 2011

The Helping Our Own (HOO) Shared Task aims to promote the development of automated tools and techniques that can assist authors of scientific papers. For that purpose, a set of scientific papers written by non-native speakers of English was provided, in which errors had been annotated using the tagset from the Cambridge Learner Corpus (Nicholls, 1999). The task was to detect and also to correct the errors because non-native speakers might not know the correct word even if pointed to an error. For a more detailed description of the task setup, see Dale and Kilgarriff (2011).

The development data contained 1,264 errors and the test data 1,057 errors. However, relatively few of them are malapropisms. One of the rare examples is file "0046" from the development data that contains ... *untagged copra are often used to do emotion classification research*, where the writer mistakenly replaced *corpora* with *copra*. As *copra* (dried coconut meat) is a valid word, the error cannot be detected using a lexicon-based spell checker. In this case, the correction would rather be ... *untagged copra is often used* ... because of the number agreement error. Such errors can only be detected using methods that analyze the contextual fitness of each term in a sentence. However, the unsupervised contextual fitness measures evaluated in this article can be applied to a wider range of error classes.

		Detection			Correction		
	Method	P	R	F	P	R	F
Single	Jazzy	.08	**.19**	.10	.03	.08	.05
	Knowledge (JC)	.23	.10	.14	.17	.08	.10
	Knowledge (ESA-WN)	.29	.09	.14	.25	.08	.12
	Statistical (Web1T)	.37	.10	.15	.32	.08	.14
	Statistical (ACL)	**.68**	.08	.14	**.67**	.08	.14
Combined	Union (all)	.07	**.21**	.10	.03	.08	.04
	Union (w/o Jazzy)	.23	.10	.14	.18	.08	.11
	Intersection (all)	.80	.08	.14	.78	.08	.14
	Intersection (w/o Jazzy)	**.94**	.08	.14	**.94**	.08	.14

Table 7. *Overview of evaluation results. Best values are in bold*

5.1. *Experimental Setup*

We use the statistical and knowledge-based approach as described above. For the statistical approach, we apply n-gram models based on (i) the Google Web1T n-gram corpus (Brants and Franz, 2006), and (ii) all the papers in the ACL Anthology Reference Corpus (Bird *et al.*, 2008). For the knowledge-based approach, we test the path-based relatedness measure by Jiang and Conrath (1997) [JC] as well as the concept vector-based ESA measure (Gabrilovich and Markovitch, 2007) with vectors created from WordNet (WN).[10] As a baseline, we use the open-source spell checker Jazzy[11] as provided by the DKPro Core framework. As our framework allows to easily combine spell checkers, we try different combinations of Jazzy, the knowledge-based, and the statistical approach.

– *Union (all)* All approaches are run in parallel and detections are merged. In the case that two approaches detect the same error but suggest a different correction, we select the correction with the higher confidence score.

– *Union (w/o Jazzy)* All approaches but Jazzy are merged.

– *Intersection (all)* All approaches are run in parallel, but only errors that are detected by each of the spell checkers are retained.

– *Intersection (w/o Jazzy)* All approaches but Jazzy are intersected.

5.2. *Results & Discussion*

Table 7 summarizes our results. The traditional spell checker Jazzy provides the best recall of the single approaches, but at the price of the lowest precision. For the other approaches, the balance between precision and recall is controlled by a threshold

[10] We also tested other measures but found the differences to be negligible.

[11] http://jazzy.sourceforge.net/

parameter. This threshold needs to be exceeded in order to flag a word as an error. For the knowledge-based approach, we threshold the semantic relatedness between a replacement candidate and its context. For the statistical approach, it is the parameter α that controls the prior probability of each word to be an error. We used a parameter setting that provided higher precision with acceptable recall levels, and found that the F-measure is relatively stable for non-extreme settings of the parameters. The detection and correction precision of the statistical approach gets a significant boost using the ACL corpus n-gram model, but at the price of an even lower recall.

Regarding the combination experiments, we find that merging all approaches but Jazzy did not significantly increase recall indicating that the statistical and the knowledge-based approaches more or less detect the same errors. In contrast, recall increases when merging all approaches which shows that the errors detected by Jazzy are somewhat complementary to those detected by the other methods. The "Union" combination strategy focuses on recall, but – in the setting of this challenge – high precision is more important than high recall, as writers might be tempted to take the detected errors and suggested corrections for granted. This could result in a document with more errors than before. Thus, we also used the "Intersection" strategy which should yield better precision. When intersecting the detection of different approaches, we obtain very high precision, but low recall.

A comparison with the results by other participants of the shared task (Dale and Kilgarriff, 2011) showed that our approach yields the best performance in many error classes, but is not well suited for article and preposition errors that together constitute about 36% of all errors in the dataset. Thus, the possible recall of our approach is already limited. A solution would be to combine the unsupervised detection (targeted towards all kinds of errors) with the supervised classification (targeted towards frequent errors like article or preposition errors).

6. Open Source Framework

Reproducing results from previous research is often hindered by the high costs of re-implementing everything from scratch. We thus provide open source implementations of all software components required to reproduce the experiments described in this article. The framework is called *DKPro Spelling*[12] and it is available under the Apache Software License (ASL), Version 2.[13]

We re-implemented the statistical approach by Mays *et al.* (1991) and the knowledge-based approach by Hirst and Budanitsky (2005). Besides the algorithm itself, both approaches rely on external resources which we also need to support. The statistical approach relies on a large database of n-gram counts like the Web1T corpus (Brants and Franz, 2006). Access to an n-gram database is encapsulated using the

[12] `http://code.google.com/p/dkpro-spelling-asl/`

[13] `http://www.apache.org/licenses/LICENSE-2.0`

generic provider resources for n-gram counts from the DKPro Core Framework.[14] It currently supports jWeb1T[15] and BerkeleyLM (Pauls and Klein, 2011). Implementations for other n-gram databases can be easily added. The knowledge-based approach relies on the ability to measure the semantic relatedness between terms. Instead of re-implementing the semantic relatedness measures, we use the DKPro Similarity package[16] that provides a wide range of measures. As all measures in the package implement the same interface, they can be easily exchanged in an experiment. We made use of this property when testing the influence of the semantic relatedness measure on knowledge-based detection of real-word spelling errors in section 4.2.

6.1. *Provided Experimental Setups*

The effort needed to set up an NLP experiment depends on many factors. A crucial first step is that researchers make the implementations of their algorithms available, but an often underestimated part of a scientific experiment is that it depends on other factors like data import, preprocessing, and evaluation of the results. All those steps need to be done by every researcher working on a certain task which results in wasting large amounts of time that could be spent more productively. Additionally, all this "scaffolding" erected in a hurry is likely to contain errors that can potentially distort the results to the extent that they are rendered meaningless. Thus, we publicly provide experimental frameworks for all the experiments described in this article:[17]

- HOO 2011 (Dale and Kilgarriff, 2011) that targets a wide range of error classes,
- detecting and correcting real-word spelling errors, and
- creating datasets via error mining from Wikipedia revision history.

Each experimental module contains the necessary code for reading the data format, preprocessing the data (e.g. sentence splitting, tagging, parsing, etc.), and evaluating the results.

7. Summary

We show that the Wikipedia revision history is a rich source for mining naturally-occurring errors and that it might provide a basis for creating other evaluation datasets whenever no suitable error annotated corpus is available. We extract a dataset with naturally-occurring malapropisms and their contexts, and show that using this dataset for evaluating statistical and knowledge-based measures of contextual fitness provides a more realistic picture of the quality of malapropism detection. In particular, using

[14] `http://code.google.com/p/dkpro-core-asl/`

[15] `http://code.google.com/p/jweb1t/`

[16] `http://code.google.com/p/dkpro-similarity-asl/`

[17] We also support the HOO 2012 Shared Task (Dale *et al.*, 2012) that targets preposition and article errors, which is beyond the scope of this article.

artificial datasets over-estimates the performance of the statistical approach, while it under-estimates the performance of the knowledge-based approach.

We show that n-gram models targeted towards the domain from which the errors are sampled do not improve the performance of the statistical approach if larger n-gram models are available. We further show that the performance of the knowledge-based approach can be improved by using semantic relatedness measures that incorporate knowledge beyond the taxonomic relations in a classical lexical-semantic resource like WordNet. Finally, by combining both approaches, significant increases in precision or recall can be achieved.

While being far from perfect, the achieved quality in detecting malapropisms is generally sufficient for competent language users who are willing to accept a certain level of false alarms in order to minimize the number of embarrassing mistakes in a document that would have gone unnoticed otherwise. The results on the Helping Our Own 2011 dataset show that contextual fitness measures are certainly not yet good enough for automatically correcting text. Especially non-native writers, as targeted by the HOO initiative, are likely to accept wrong suggestions made by the system. Thus, increased research effort is necessary in order to further improve results.

In future work, both – the statistical as well as the knowledge-based approach – will benefit from a better model of selecting likely confusables for a given word instead of the brute-force strategy of testing all candidates within a certain edit distance. Although a deeper analysis is necessary, the nature of the confusables mined from the Wikipedia revision history suggests that malapropisms are the product of a cognitive process where similar sounding words are confused rather than the result of typing errors. On the side of extracting errors from the Wikipedia edit history, we are going to further improve the extraction process by incorporating more knowledge about the revisions. For example, vandalism is often reverted very quickly, which can be detected when looking at the full set of revisions of an article.

We make the full experimental framework publicly available which will allow the scientific community to reproduce our experiments as well as conducting follow-up experiments. The framework contains (i) methods to extract natural errors from Wikipedia, (ii) reference implementations of the knowledge-based and the statistical methods, and (iii) the evaluation datasets used in our experiments.

Acknowledgements

We would like to thank the anonymous reviewers for their very detailed comments, also regarding numerous language errors that neither the system described in this article nor intensive proofreading could detect. We would also like to thank Robert Dale for organizing the “Helping Our Own” Shared Task that triggered most of our work described here.

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806.

8. References

Bean D., Riloff E., “Unsupervised Learning of Contextual Role Knowledge for Coreference Resolution”, *Proceedings of HLT/NAACL*, Boston, Massachusetts, USA, p. 297-304, 2004.

Bird S., Dale R., Dorr B., Gibson B., Joseph M., Kan M.-Y., Lee D., Powley B., Radev D., Tan Y. F., “The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics”, *In Proceedings of Language Resources and Evaluation Conference (LREC 08)*, Marrakesh, Morocco, 2008.

Bolshakov I. A., Gelbukh A., “On Detection of Malapropisms by Multistage Collocation Testing”, *Proceedings of the 8th International Workshop on Applications of Natural Language to Information Systems at NLDB*, 2003.

Brants T., Franz A., “Web 1T 5-Gram Version 1”, 2006.

Bronner A., Monz C., “User Edits Classification Using Document Revision Histories”, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, p. 356-366, 2012.

Budanitsky A., Hirst G., “Evaluating Wordnet-Based Measures of Lexical Semantic Relatedness”, *Computational Linguistics*, vol. 32, n° 1, p. 13-47, 2006.

Carlson A. J., Rosen J., Roth D., “Scaling Up Context-Sensitive Text Correction”, *Proceedings of the 13th Conference on Innovative Applications of Artificial Intelligence Conference (IAAI)*, p. 45-50, 2001.

Dale R., Anisimoff I., Narroway G., “HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task”, *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, Montreal, Canada, p. 54-62, 2012.

Dale R., Kilgarriff A., “Helping Our Own: The HOO 2011 Pilot Shared Task”, *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France, p. 242-249, September, 2011.

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

Ferschke O., Zesch T., Gurevych I., “Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia’s Edit History”, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, Portland, OR, USA, p. 97-102, 2011.

Finkel J. R., Grenager T., Manning C., “Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling”, *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, p. 363-370, 2005.

Francis W. N., Kuçera H., “Manual of Information to Accompany a Standard Corpus of Present-day Edited American English, for Use with Digital Computers”, 1964.

Gabrilovich E., Markovitch S., “Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis”, *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, p. 1606-1611, 2007.

Golding A. R., Schabes Y., “Combining Trigram-Based and Feature-Based Methods for Context-Sensitive Spelling Correction”, *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (ACL)*, p. 71-78, 1996.

Granger S., *A Bird’s-Eye View of Learner Corpus Research*, John Benjamins Publishing Company, p. 3-33, 2002.

Han N.-R., Chodorow M., Leacock C., “Detecting Errors in English Article Usage by Non-Native Speakers”, *Natural Language Engineering*, vol. 12, n° 2, p. 115, May, 2006.

Hirst G., Budanitsky A., “Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion”, *Natural Language Engineering*, vol. 11, n° 1, p. 87-111, March, 2005.

Hirst G., St-Onge D., *WordNet: An Electronic Lexical Database*, Cambridge: MIT Press, chapter Lexical Ch, p. 305-332, 1998.

Inkpen D., Désilets A., “Semantic Similarity for Detecting Recognition Errors in Automatic Speech Transcripts”, *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT ’05*, p. 49-56, 2005.

Islam A., Inkpen D., “Real-Word Spelling Correction Using Google Web1T 3-Grams”, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1241-1249, 2009.

Jaro M. A., “Probabilistic Linkage of Large Public Health Data File”, *Statistics in Medicine*, vol. 14, p. 491-498, 1995.

Jiang J. J., Conrath D. W., “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”, *Proceedings of the 10th International Conference on Research in Computational Linguistics*, Taipei, Taiwan, 1997.

Jones M. P., Martin J. H., “Contextual Spelling Correction Using Latent Semantic Analysis”, *Proceedings of the 5th Conference on Applied Natural Language Processing*, p. 166-173, 1997.

Lemnitzer L., Kunze C., “GermaNet – Representation, Visualization, Application”, *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, p. 1485-1491, 2002.

Lesk M., “Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone”, *SIGDOC ’86 Proceedings of the 5th Annual International Conference on Systems Documentation*, p. 24-26, 1986.

Lin D., “An Information-Theoretic Definition of Similarity”, *ICML ’98: Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 296-304, 1998.

Max A., Wisniewski G., “Mining Naturally-Occurring Corrections and Paraphrases from Wikipediaâs Revision History”, *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC’10)*, p. 3143-3148, 2010.

Mays E., Damerau F. J., Mercer R. L., “Context Based Spelling Correction”, *Information Processing & Management*, vol. 27, n° 5, p. 517-522, 1991.

Nelken R., Yamangil E., “Mining Wikipedia’s Article Revision History for Training Computational Linguistics Algorithms”, *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (WikiAI)*, WikiAI08, 2008.

Nicholls D., “The Cambridge Learner Corpus – Error Coding and Analysis for Lexicography and ELT”, *Summer Workshop on Learner Corpora*, Tokyo, Japan, 1999.

Pauls A., Klein D., “Faster and Smaller N-Gram Language Models”, *Proceedings of ACL*, Association for Computational Linguistics, Portland, Oregon, June, 2011.

Rozovskaya A., Roth D., “Annotating ESL Errors: Challenges and Rewards”, *The 5th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL-HLT)*, 2010.

Rubenstein H., Goodenough J. B., “Contextual Correlates of Synonymy”, *Communications of the ACM*, vol. 8, n° 10, p. 627-633, 1965.

Schmid H., “Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors”, *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 2004.

Tetreault J., Foster J., Chodorow M., “Using Parse Features for Preposition Selection and Error Detection”, *Proceedings of the ACL 2010 Conference Short Papers*, p. 353-358, 2010.

Wick M., Ross M., Learned-Miller E., “Context-Sensitive Error Correction: Using Topic Models to Improve OCR”, *9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, p. 1168-1172, 2007.

Wilcox-O’Hearn A., Hirst G., Budanitsky A., “Real-Word Spelling Correction with Trigrams: A Reconsideration of the Mays, Damerau, and Mercer Model”, *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, 2008.

Yatskar M., Pang B., Danescu-Niculescu-Mizil C., Lee L., “For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia”, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT ’10, p. 365-368, 2010.

Zesch T., Gurevych I., “Wisdom of Crowds Versus Wisdom of Linguists – Measuring the Semantic Relatedness of Words”, *Journal of Natural Language Engineering*, vol. 16, n° 1, p. 25-59, 2010.

Zesch T., Müller C., Gurevych I., “Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary”, *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, 2008a.

Zesch T., Müller C., Gurevych I., “Using Wiktionary for Computing Semantic Relatedness”, *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, Chicago, IL, USA, p. 861-867, 2008b.