Substituting Phrasal Verbs in English as a Foreign Language for L1 Spanish speakers

Cornelius Fath, Maryam Geranmayeh, Ignacio Rubio Majano

Hauptseminar: Computational Approaches to Text Simplification
Advisors: Detmar Meurers, Sowmya Vajjala

Seminar für Sprachwissenschaft
Universität Tübingen
Summer Semester 2013

Abstract

In this work we present the tool PV-Sub for automatic text simplification. It tackles phrasal verbs, which are known to be challenging for Spanish L1 learners of English, and substitutes them with synonyms of Romance origin. We identify phrasal verbs by running a dependency parser on a text. Multiple dictionaries are employed for the retrieval of translations and synonyms. Using Levenshtein distance on the collected condidate verbs, we select substitutes which are likely to be of Romance origin. We apply two approaches to word sense disambiguation (Web-as-Corpus and the Lesk algorithm) in order to improve the adequacy of substitute verbs.

Contents

1 Introduction

In recent years, there has been an increasing interest in utilizing NLP tools to build assistive technologies, one of which is Automatic Text Simplification. While there are more technical applications of Automatic Text Simplification such as in the domain of machine translation or dependency parsers, the goal of this NLP task is to simplify text, in order to make it more accessible to a target audience such as children [De Belder and Moens, 2010], people with low literacy skills [Dell'Orletta et al., 2011], readers with aphasia [Carroll et al., 1998] or foreign language learners [Petersen and Ostendorf, 2007]. A Text Simplification system usually consists of a syntactic and a lexical simplification module, which are executed one after the other: the first module tries to simplify the grammatical complexity of a text by modifying its sentence structure, and the latter attempts to capture complex words in a sentence and substitute them with more suitable alternatives with respect to the target reader group. In either approach, the preservation of information is of prime importance. In this paper we go about a particular aspect of lexical text simplification, the motivation being to adapt English texts to a lower level of proficiency of foreign or second language learners.

An impulse for lexical simplification is the striking amount of Latin roots in the English vocabulary. Latin occurrences in English are partly traceable to early religious and scientific contexts. The larger portion, however, was delivered by the Normans from 1066. From the time of the invasion on, French became the language of formality and a huge amount of Latin-based words entered the lexicon of the upper class people, merchants and military officers [Levin and Novak, 1991]. The result is that about 50% of the words in the English language of today are Latin-based [Smith, 1995]. Finkenstaedt and Wolff [1973] mention figures about the origin of English vocabulary which can be seen in Figure 1. Latinate elements of the English lexicon represent a more formal and literate register than its native Germanic stock [Bar-Ilan and Berman, 2007], with the ratio being higher in expository than in narrative texts, and rising as a function

of age[1]. However, learners of English who have a Romance L1 are likely to find these Latinates easier to understand than phrasal verbs –which do not exist in their native languages–, and therefore will consider these adapted words as a practical simplification aid.

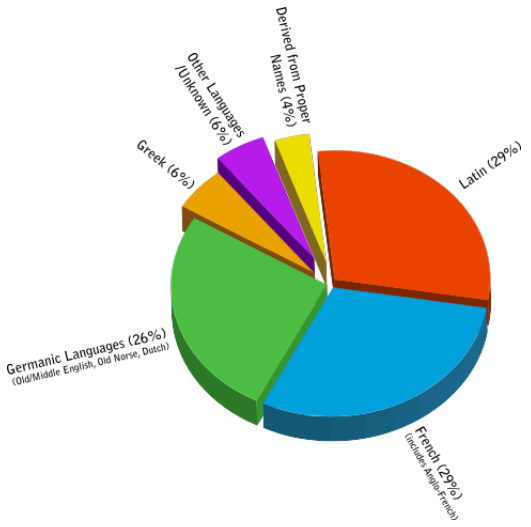

Figure 1: Origins of the English Language [Wikimedia Commons, 2007]

The subsequent thought is that learners of English who have a Romance L1 are likely to find these adopted words easy to understand, although they do not necessarily mean exactly the same. Why not make use of this common portion in both lexicons in order to help the learners?

Here, we describe **PV-Sub**: A simplification tool for speakers of Spanish[2], which substitutes phrasal verbs in English texts.

1.1 Phrasal Verbs

A phrasal verb is a phrase which consists of a verb in combination with a preposition or an adverb, and whose meaning is different from the meaning of its separate parts. Phrasal verbs exhibit a high frequency in written and especially in spoken English. For non native speakers, however, phrasal verbs are known to represent one of the most difficult aspects of learning English. Leacock et al. [2010] list a number of properties of phrasal verbs which contribute to the challenge of acquiring these items for second language learners of English. The

[1]This has often led to pairs of synonyms where the French version is more formal than the English one (compare *commence* to *start*), as well as triplets where the most formal version is a Latinate, followed by a French-derived term, the Anglo-Saxon term being the most informal (compare *to ask*, *to question* and *to interrogate*) [McWhorter, 2009]

[2]We assume that speakers of other Romance languages could also profit from the tool.

fact that the meaning of phrasal verbs cannot be derived from the meaning of its components, also referred to as non-compositionality, leads to the consequence that phrasal verbs must be learned by memorization. In addition, apart from the lexical form, learners must distinguish between phrasal verbs that allow their object to appear before their particle, and those which do not. Kurtyka [2001] suggests more features of phrasal verbs to be contributors to the complexity of phrasal verbs, such as restrictions on tense and collocational associations with other words, as well as the attachment of multiple meanings to a single phrasal verb.

The challenge that phrasal verbs pose to learners is mostly apparent in learners avoidance behavior. Dagut and Laufer [1985] show in a study that Hebrew-speaking students tend to avoid using phrasal verbs when expressing themselves in English: in a multiple choice test, when presented with a phrasal verb form among the possible choices, one-word alternatives are preferred in more than half of the cases. Laufer and Dagut further propose that a possible reason for this behavior is the missing equivalent of phrasal verbs in Hebrew, and suggest that learners with an L1 that makes use of similar phrasal constructions should show less of a difficulty in phrasal verb usage. Hulstijn and Marchena [1989] address this matter in a similar study where they show that Dutch users do in fact use phrasal verbs when these are of the literal and completive category, but show difficulty in usage when phrasal verbs are of the figurative category, suggesting that the reason for the learner's avoidance of English phrasal verbs is mostly due to the idiomatic nature of many phrasal verbs rather than L1 interference.

González [2010] attempts to investigate whether native speakers of Spanish have a higher tendency to avoid using phrasal verbs than Swedish students (whose L1 makes use of phrasal verb constructions). Arguing that previous work on students' avoidance behavior has studied this phenomenon under experimental conditions, and that different methodological circumstances might exhibit different patterns of language use, González looks into natural corpora consisting of short essays from 3rd and 4th grade Spanish and Swedish learners of English, as well as a subcorpus of the British National Corpus consisting of short essays from native speakers of English. In this study, L2 speakers of English were considered to show avoidance behavior, since they used a significantly smaller number of certain particles (in this case the particle *out*) when compared with native speakers. From this study, González concludes that non-natives are less likely to use phrasal verbs than native speakers of English in essay writing. Furthermore, his findings suggest that Spanish students are more likely to underuse phrasal verbs when compared to the Swedish students. González, like Laufer and Eliasson [1993], proposes that L1 interference is a possible reason for this, since Swedish students are not unfamiliar with the phrasal verb construction.

While it is true that the missing equivalent to phrasal verb constructions may put Spanish learners of English at a disadvantage, this group of learners is con-

sidered to benefit from the large number of Latin-based words that are strongly represented in English, as has been discussed above. Cognates (vocabulary that is nearly identical or very similar between Spanish and English) can be found among all types of content words, and it is not an unimaginable idea that parallel to avoidance, preference for Latin-based single-word synonyms of phrasal verbs may play a role in the behavior of this group.

A study by Jiménez et al. [1996] on Spanish-English bilingual children who are expert readers in English found that the ability to identify cognates in decoding unfamiliar words was a key feature of bilingual reader's repertoire of skills when reading in English. Nagy et al. [1993] further address this matter in a study where 4th to 6th grade bilingual students were assessed for comprehension after being instructed to pay attention to cognates. Results of this study show that the students' performance on the items on the English multiple-choice for text comprehension significantly correlates with the students' ability to identify the cognates. This result, the authors suggest, reflects a transfer of Spanish lexical knowledge to reading in English.

The technique of cognate recognition for text comprehension is not restricted to bilinguals and is actually used by many language instructors to effectively facilitate the acquisition of English for Spanish speaking students through reading tasks. There exists data from language tests which addresses the fact that candidates of Spanish speaking groups are likely to make use of cognate recognition for their performance on English exams. Comparing TOEFL [Alderman and Holland, 1981] scores of examinees from different L1 backgrounds reports evidence that Hispanic examinees' performance on comprehension tasks among others is influenced by the existence of true cognates, words with a common English and Spanish root sharing the same meaning. Results of analyzing SAT scores of Mexican and Puerto Rican students [Schmitt, 1988] support these findings.

Phrasal verbs do not only pose a problem in production of the foreign language, but also in language comprehension. In order to determine to what extent they affect learners, we conducted a quick informal poll with 22 native Spanish speakers who had varying levels of proficiency in English. In a preliminary phase, we asked some of them to name features of English they considered difficult. Based on the results, we presented the whole group with a series of five linguistic phenomena[3] and asked them to choose the two that make English texts most difficult to understand for them. The conclusions were clear: the most difficult were phrasal verbs (19 votes out of 22 participants), followed by conjunctions (11). This information confirms our original intuition.

Different to most approaches on lexical text simplification, we do not detect difficult words by looking at word frequencies (e.g. [Carroll et al., 1998]), but lay the focus on a specific group of words which are known to be difficult.

[3]The phenomena presented were: irregular verbs, conjunctions, phrasal verbs, long sentences and technical terms

2 Tool Architecture

The purpose of our tool is to provide alternative English verbs for the phrasal verbs (henceforth PV) we find in a text and replace them.

The core idea for this replacement is that there are likely to be verbs in English which have a Romance origin and are therefore similar to their Spanish correspondent. An example of this is the PV "go on", where "continue" is in the set of synonyms and "continuar" in the set of translations. "Continue" would thus be a suitable substitute because it resembles a Spanish translation and might be easier for a Spanish L1 reader.

2.1 Text Analysis

As a first step, PVs have to be detected in a given text. We use the *Stanford Parser* [Klein and Manning, 2003] to parse the text into its tree structure. In order to find PVs in the text we make use of *Stanford Dependencies* [De Marneffe et al., 2006]. Once we have the dependency structure of a sentence, we look for particles ("*prt*") and simultaneously for their governors, which combined form PVs. We also made a run with the MaltParser [Nivre et al., 2006], which approximately did the same job in detecting particles. In 271 sentences fully packed with PVs, the MaltParser differed in only two cases from the Stanford Parser. As the output format of the Stanford Parser was a little bit easier to handle, we picked it, but could have picked the MaltParser just as well. We also look for prepositions with their governors as they often really are particles misparsed as prepositions. An indicator for in fact being dealing with a PV is that the governor must always be a verb and appear before the particle. Additionally we have to keep track of the positions we find PVs at, in order to put substitutes in the right place. Moreover, the POS tag of each detected PV has to be stored, for ensuring that the substitute will have the same form. With the lemmatized PV at hand the quest for alternatives can begin.

2.2 Synonym and Translation Gathering

For finding alternatives, two sets of words are needed: Synonyms (English) and translations (Spanish).

We gather verbs from three different sources:

Synonyms:	WordNet [Miller, 1995]
Synonyms and translations:	WordReference.com
Translations:	SpanishDict.com

To gather synonyms, we search the PV in *WordNet* and *WordReference*. All the gathered candidates are stored in a set of **English synonyms**. Then we look

for Spanish translations of the PV on *WordReference* and *SpanishDict*. These translations are stored in a set of **Spanish translations**. Now every element of the synonym set is compared against every element of the translation set.

Technically, we access the pages via JSOUP[4] requests. JSOUP is a Java library which allows to call URLs and extract the HTML code of the retrieved pages. Also, it permits to traverse the DOM of the HTML and pick out individual objects. This is what we make use of for finding synonyms and translations of our search term in the HTML instead of using an API (2.2.2).

2.2.1 Individual contribution of each source

The following table gives an idea about what each source contributes to our two sets of synonyms and translations.

Search Term	**WordNet**	**WordRef.S**	**WordRef.T**	**SpanishDict**
carry on	conduct, deal, continue, uphold, preserve, continue, proceed	continue, preserve	seguir, proseguir, continuar, seguircon, conservar	continuar, seguir, dirigir, gestionar, mantener, llevar
get down	lower, unhorse, dismount, light, swallow, depress, deject, dismay, dispirit, demoralize, demoralise, begin, get, start, commence	horse, depress	bajar, apearse, desanimar	bajar, reducir, bajarse, descolgar, tragarse, tragar, escribir, deprimir, molestar, agacharse, levantarse de la mesa
get over	traverse, track, cover, cross, overcome, subdue, surmount, master	accept	recuperarse, mejorarse, sobreponerse, olvidarse, aceptar, creer	superar, hacer llegar, transmitir, cruzar, franquear, recuperarse, atravesar, reponerse, sobreponerse, olvidar, vencer, dominar, venir, ir

Table 1: Synonyms and translations provided by multiple sources

[4] http://jsoup.org/

2.2.2 APIs

The PV-Sub approach is a pipeline program based on an cognate recognition algorithm that attempts to substitute phrasal verbs with synonyms harvested from the field of candidates found in encyclopedic resources. The desired procedure to access these resources is through an interface, which directly interacts with them. Like this, the results (synonyms, translations) can be plugged in whenever available. In this way, the system stays flexible for new resources. However, APIs appear to be relatively sparse in the domain dictionaries, or at least are not easily accessed. Therefore, we opted for the following approach. WordNet is the resource which is employed the most by our program, as it is used for finding substitution candidates as well as disambiguating among them before replacement. In order to access this information we use the Java API for WordNet Searching (JAWS).

WordReference provides APIs and also encourages the usage of these for development. Unfortunately the APIs do not support libraries for the language the program is written in, namely Java. However, we have the kind permission of the author to send queries to the site via JSOUP.

SpanishDict does not provide APIs, however permits queries under certain guidelines. It is worth noting that the PV-Sub approach would benefit significantly from access to additional English-Spanish dictionaries, under the assumption that these dictionaries supply translations that are not captured by SpanishDict. Examples of such dictionaries are the Cambridge Dictionary and Longman Dictionary. API keys for these dictionaries are available upon request and permit users to send a definite number of queries per day. Unfortunately we received no response to our requests for API keys.

2.3 Word Comparison

The Levenshtein algorithm[Levenshtein, 1966], also called *edit distance*, is the most widely used to measure the difference between two strings. It is defined as the least number of edit operations (i.e. deletions, insertions or substitutions) required to transform a source string s into a target string t. This is the method we use to compare the synonyms with the Spanish translations. Schepens et al. [2012] show that, even without further adjustments, it is a reliable measure for finding cognate pairs from different languages. However, Schepens et al. also report false negatives due to the Levenshtein distance not taking into account grapheme-to-phoneme mappings. An example for this is that 'c' in English often corresponds to 'k' in German (e.g. *project* vs. *Projekt*). The Levenshtein distance treats this as a full substitution step. This is why we apply a few more dodges in order to pre-adapt the verbs to each other as much as possible and hereby make the comparison even more reliable. We employ evident self-defined rules based on observations about how Spanish words relate to English

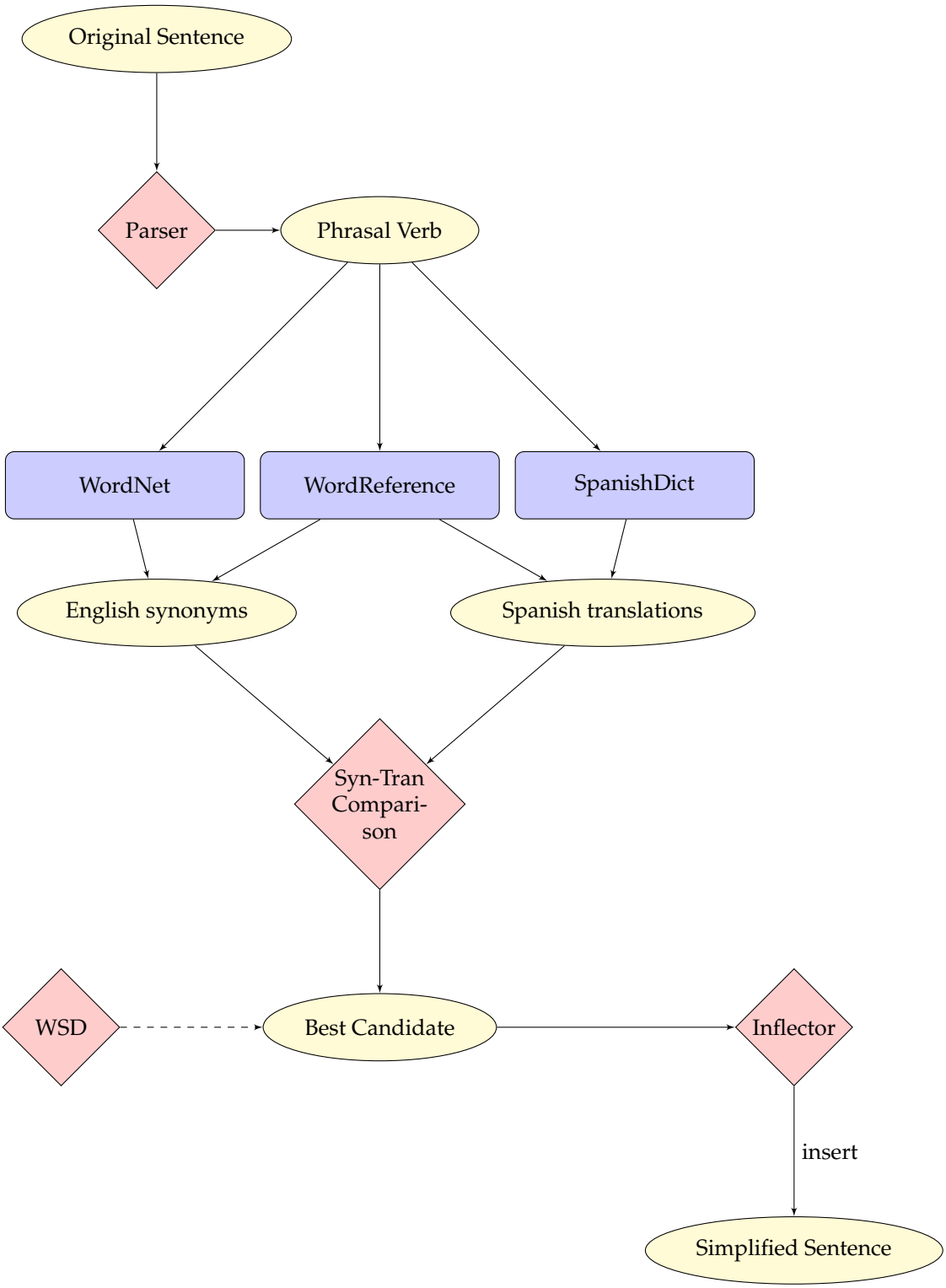

Figure 2: Dynamic view of the system components

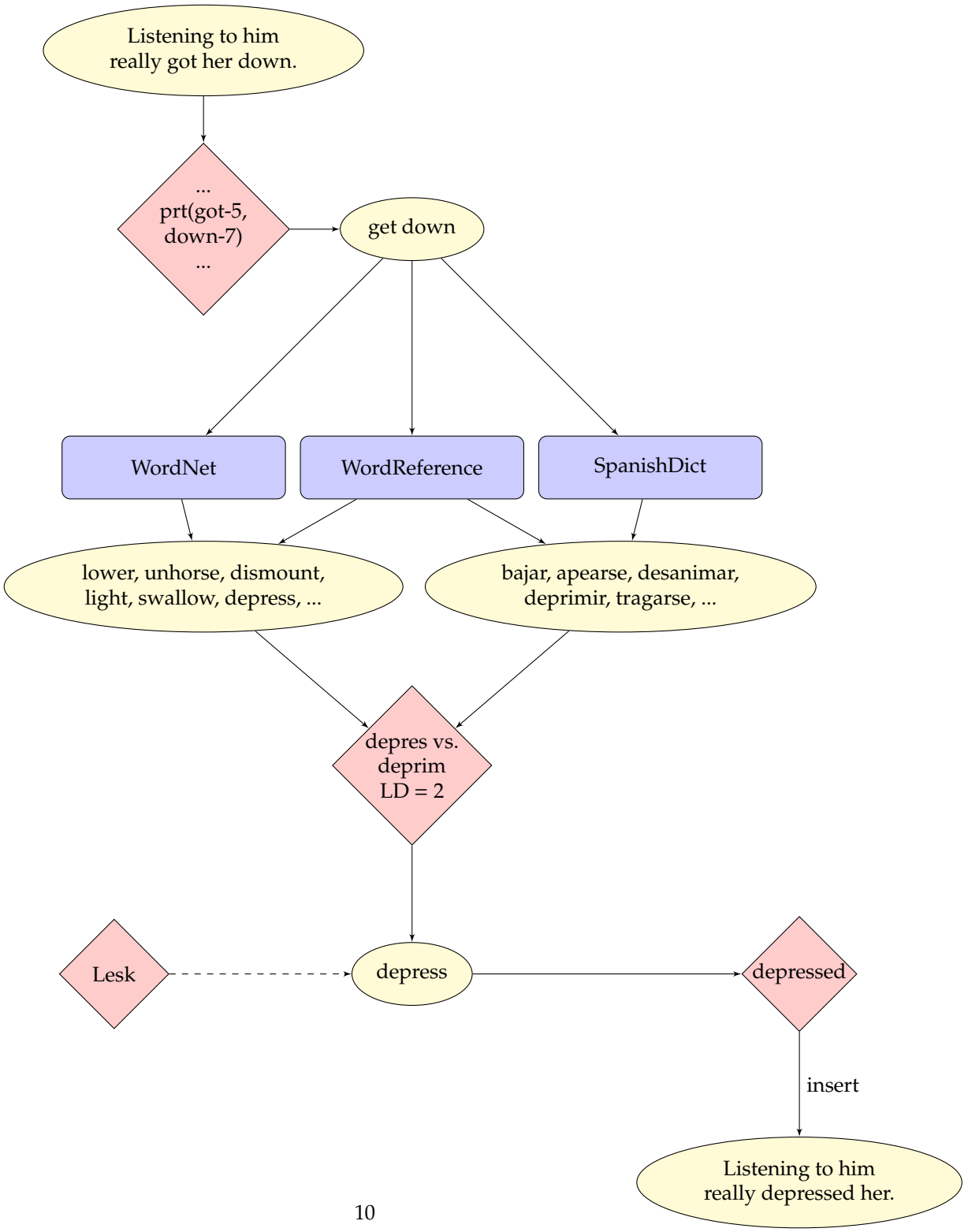

Figure 3: Example sentence run through the system

correspondents.

Adaptation of English synonyms

- Removal of double consonants
  English *a**ss**imilate*, Spanish *a**s**imilar*
- Removal of *h* in *th*
  English *empa**th**ize*, Spanish *empa**t**izar*
- Replacement of diphthongs by single vowels (multiple alternatives are generated)
  English *recl**ai**m*, Spanish *recl**a**mar*
- Removal of initial *e* if verb starts with *esc-, est-, esp-*
  English ***esc**ape*, Spanish ***esc**ape*

Adaptation of Spanish translations

- Removal of verb endings (*-ar, -er, -ir, -arse, -erse, -irse*)
  Spanish *retir**arse***, English *retir**e***
- Removal of initial *e* if verb starts with *esc-, est-, esp-*
  Spanish ***esp**ecular*, English ***sp**eculate*
- Transformation of *ñ* to *n*
  Spanish *acompa**ñ**ar*, English *accompa**ny***

Let us have a look at an example for how these adaptation rules are motivated. Consider the PV *to come across*. In the synonym set we find *encounter*; in the translation set we find *encontrar*. The Levenshtein distance of these two words is 3. That would mean that we would have to set our threshold at least to 3 if we want the two to be classified as being of the same origin. This is, according to the experience we gained in test runs, a relatively high threshold, considering the number of false positives. With the adaptation rules applied, we are left with the synonym *'enconter'* and the translation *'encontr'*, which results in a Levenshtein distance of 1. The rules allow us to ignore systematic spelling differences between English and Spanish and restrict the focus to real deviations. One step remains before the comparison takes place. Having taken away the schematic Spanish verb suffixes, we are still facing suffixes of the English verb, which are in general less schematic. Take e.g. the verbs *accompany/acompañar*. With the rules we receive *'acompany/acompan'*. By adapting the length of the two, i.e. stripping the English to the length of the Spanish verb, we can even get rid of this non-significant deviation. The above example will then end up in *'enconte/encontr'*, which also means Levenshtein distance 1.

However, this stripping should only be applied if the English verb stays more than a stub and the part thrown away stands in a reasonable relation to the length of the word. Otherwise, it would be possible to derive that two words are similar, even though they are not. If we compare e.g. *surface/subir*, we get *'surface/sub'* after the rules. After stripping we would compare *'sur'* to *sub* and

get a Levenshtein distance of 1. That is, of course, not desirable, so we limit the stripped suffix to length 2.

Having processed and adapted the two words as much as possible, a low Levenshtein distance tells us that they are likely to be of the same origin. If so, the synonym can be seen as a valid alternative for the PV. We refer to this alternative as "best alternative" hereafter.

2.4 Inflection and Replacement

Once a suitable alternative for the PV is found, it has to be set in the right form in order to fit in the text. The POS tag of the PV has been stored before and can now be used to inflect the alternative. We use an external list of 189 irregular verbs and a fistful of rules for inflecting regular verbs. The inflected alternatives can now be inserted in the previously stored positions.

2.5 Word Sense Disambiguation

In a broad sense, PV-Sub is a lexical simplification approach that is based on choosing isolated target words in a text and substituting them with a synonym. Given the polysemous nature of most words in natural languages, any substitution approach of this sort must deal with word sense disambiguation at some point. Up until now, there have been no semantic restrictions on our set of candidate synonym verbs, and, as a result, the final output would contain a verb that is synonymous with the original phrasal verb, but whose meaning could substantially differ from the intended one, therefore rendering the sentence unintelligible, which is the opposite of what a text simplification tool is supposed to do. As mentioned in the introduction, cognates do not necessarily share the same meaning, but can deviate slightly or even have completely different meanings. A word sense disambiguation unit could aid to filter synonyms of this sort.

As an example where word sense disambiguation could be of benefit, consider the sentence *"The tourists were taken in by conmen"*. WordNet provides 17 synsets for the phrasal verb *'take in'*, ranging from one to ten synonyms per synset. Without using any semantic filtering criteria, our algorithm encounters, among others, the synonym *'to adopt'* and puts it into the bag of candidate synonyms. Given the similarity between this verb and the Spanish verb *'adoptar'*, our program is likely to consider it as an ideal pair and consequently select it as the best substitution for the target phrasal verb, and thus return the sentence *"The tourists were adopted by conmen"*. This sentence has a clearly different meaning than the original.

Intuitively, a human reader/simplifier would work around this semantic problem by using clues from other predicates within the sentence. In this case,

'conmen' (provided that the reader knows its meaning) calls into mind words like *'to deceive'*, *'to fool'* or *'hoax'*, rather than *'to adopt'*, and this remains the main strategy to choose between different meanings of *'to take in'*.

We employ two different disambiguation methods in our study: one approach is based on n-gram patterns in a corpus (*Web-as-Corpus*) and applied after PV-Sub, and the other on a dictionary- and knowledge-based method (Lesk) which restricts the set of synonyms for PV-Sub.

2.5.1 Web-as-Corpus

Inspired by approaches by Elghafari et al. [2010] who predict prepositions and Boyd et al. [2012] who detect determiner and preposition errors with the web as corpus, we perform web searches for getting an impression of how likely the newly created sentence is to occur and therefore how likely the best alternative is to be an appropriate verb alternative.

The window we pick begins with one token as left context, the best alternative, and ends with two tokens as right context. If the sentence's boundaries are narrower than this, the window is smaller. The term in this window is searched with Yahoo or Google and the number of hits is retrieved. Another three searches with the same context are done with verbs from the senses list that has been created for the PV. Note that these verbs do not have to do with similarity to Spanish translations anymore. They are just the next ones on the WordNet list. This choice is motivated by the ordering of senses of the word by WordNet. WordNet lists the most common above others. However, as they state in their guidelines, the semantically tagged text they use as resource is somewhat too sparse to base research upon. Still WordNet seems the most accurate source for our purposes. In the few cases when WordNet does not provide enough synonyms for the comparison, we use WordReference as a secondary resource for synonyms. We also worked with a "second best alternative", but it turned out that its Levenshtein distance is often quite high so the choice of the second best alternative is too random.

The setting is so that if one of the three additional verbs occurs at least ten times as often in the given context as the best alternative and at least six times, the supposedly best alternative gets overridden by the alternative verb with most hits in the particular context.

Unfortunately, neither Google nor Yahoo offer free API usage anymore. Therefore we launch the searches from our own JSOUP request and parse the retrieved HTML pages for the number of hits. Both Google and Yahoo limit the number of searches. Nevertheless, we can obtain enough results with our method to evaluate its benefit.

2.5.2 The Lesk Approach

The Lesk algorithm, introduced by Lesk [1986], is based on the intuition that words in a sentence are topically related in meaning with other words in the same sentence. Disambiguation of a target word lies in identifying which sense of the word shares a wider range of semantic properties (meaning) with other content words present in the sentence. The basic implementation of this algorithm assigns a score to the candidate senses of the word in question by counting word overlaps between the dictionary definitions of the target word and other content words in the sentence. Since this algorithm is very dependent on the wording of the dictionaries in use, its performance is very sensitive to cases in which the definitions are kept concise, which is often the case.

Various extensions of the Lesk algorithm attempt to tackle this problem. Here we employ an adaptation of the Lesk algorithm put forward by Banerjee and Pedersen [2002]. This extended version of the Lesk algorithm seeks to find the best combination of senses in a short window of words around the target rather than disambiguating one word at a time. Furthermore, overlaps are not only sought in dictionary glosses of the words, but instead a database is constructed for each sense of a word, containing extended definitions which are gathered from words that stand in semantic relations to the word in consideration (such as holonyms and meronyms in the case of nouns, or hypernyms and troponyms in the case of verbs). The scoring of sense combinations is performed by counting definition overlaps in a heterogeneous scheme of selecting synset pairs for gloss–comparison from this database, assigning a higher weight to phrase overlaps. The mentioned semantic relations can be captured by using WordNet.

In our approach, the extended Lesk algorithm is employed as follows. Once a phrasal verb has been identified, instead of collecting all synonyms that are found in WordNet and WordReference, only those synonyms are put in the bag of synonyms which belong to the sense (synset) that the extended Lesk algorithm finds to be the most semantically relevant. In this way, we hope to discard spurious synonyms of the phrasal verbs before they enter the stage of word comparison. Since WordNet is our only source which provides the synset architecture as well as semantic relations between them, it is the only source used for PV replacement in this approach.

The next step is to determine which words in the sentence are relevant to use in the task of disambiguating the phrasal verb. Let us look closer at this problem by considering different cases. In the sentence *"Recently she had taken up archaeology"*, there is no word that could help disambiguate the phrasal verb among its many meanings; in this case, a human reader would understand that *archaeology* is implied as a hobby, but the word does not help to better understand the meaning of the phrasal verb per se. Consider also the sentences *"She didn't get away with it"* and *"The next morning Bill turned up in a new jeep"*, where no relevant words can be found for this purpose. In other cases, such as

"I woke up this morning, my baby was gone", the temporal modifier does provide useful information to the disambiguation of *woke up*, but most often it is not the case.

In order to tackle this problem, we use information provided by the dependency parser. We consider to be *content words* those tokens which are nouns and which are found in a dependency relation governed by the verb of the phrasal verb[5]. We then run each of these words through Lesk in conjunction with the phrasal verb, and select the synset of the phrasal verb which achieves the highest score, thus restricting the set of meanings. Finally, among the candidate synonyms found in the final synset, we search for a Latinate.

3 Test Run

For evaluating the performance of our tool, we use *"The Adventures of Lara Cruft. A phrasal verb story in 6 parts."* and eight other phrasal verb-rich texts with kind permission of the *EFL Theatre Club* [Streames, 2011]. For an excerpt see Appendix A.

3.1 Parser Results

The orientation point for correctly identified PVs is the concordance with those highlighted as PVs –or verb plus preposition– by the author of the texts, and constitute 151 items. The Stanford Dependency parser identifies 78 of these as phrasal verbs or verb plus preposition. 26 items are found in addition. Those are not necessarily false positives. One of the reasons for this is that the texts follow a certain syllabus, and the highlighting provided by the author sometimes misses out on reoccurring PVs. All in all, PV-Sub looks up 104 verbs.

	PV-Sub		
	nude	with WSD	
		WaC	Lesk
Identified PVs (Stanford)	104		
Results yielded	93		40
Suitable alternatives	25	26	19
- better than nude PV-Sub		11	9
Non-suitable alternatives	68	67	21
- worse than nude PV-Sub		9	2

Table 2: Overview of the results

[5]We expect these words to be the most relevant, but this choice is based on the analysis of a set of sample sentences, and could be further improved by adding other categories of words, such as verbs and adjectives, or elements from surrounding sentences.

3.2 Alternatives

For 11 verbs PV-Sub does not find any alternatives. We evaluate the quality of the found alternatives manually, i.e. we look at each context and decide whether the inserted verb is suitable. First, we look at the best alternative PV-Sub suggests. Out of 93 alternatives, we consider 25 to be suitable, 68 not suitable.

go on - continue
Original phrase: What should they do, go on or turn back?
PV-Sub: What should they do , continue or turn back ?

Figure 4: Example for PV-Sub + Web-as-Corpus

3.2.1 Web-as-Corpus

With the Web-as-Corpus approach as described in section 2.5.1 we find 26 suitable alternatives and 67 non-suitable alternatives. Out of the 26 suitable cases, 11 are suitable while the PV-Sub best alternatives are not. In 9 cases out of the 67 non-suitable alternatives, the nude PV-Sub had suggested a suitable one.

turn down - reject
Original phrase: all the companies turned me down
nude PV-Sub: all the companies spurned me
PV-Sub + WAC: all the companies rejected me

Figure 5: Example for PV-Sub + Web-as-Corpus

In 30 cases, the best alternative suggested by PV-Sub gets overridden by an alternative suggested by the Web-as-Corpus approach due to the high number of hits on the web, of which 10 were valid replacements. Overriding is just based on hits, not on suitability. Although the Web-as-Corpus approach on its own does not perform better than PV-Sub, it helps finding a good alternative as it only overrides the PV-Sub if it is relatively "sure", i.e. reports a very high number of hits.

3.2.2 Lesk

The Lesk algorithm yields results, i.e. a preferred synset, in 40 out of 104 phrasal verbs identified by the Stanford Dependency Parser. In 19 cases, we considered the suggested verb to be an adequate synonym for the target phrasal verb. Out of these, in 9 cases the Lesk suggestion actually was an improvement to the nude PV-Sub suggestion.

take back - return
Original phrase: so she had to take it back to the shop
nude PV-Sub: so she had to reclaim it to the shop
Lesk + PV-Sub: so she had to return it to the shop

Figure 6: Example for Lesk + PV-Sub

In 2 of the 21 non-adequate Lesk suggestions, nude PV-Sub provided a better alternative. Summing up, there are 11 cases where Lesk's suggestion deviates from the nude PV-Sub suggestion: 9 constituted an improvement, and 2 yielded worse results. These results prove that word sense disambiguation is beneficial to our task.

4 Future Work

The tool described in this paper focuses on a very small set of vocabulary in the English language. It is a set known to be difficult for a particular target group, and it is replaced with alternatives of a set known to be easy for the same group. It can thus be seen as an example application of lexical text simplification with a very narrow field of application.

Extensions of this tool could append at various points. Currently, PV-Sub considers every phrasal verb found in a text for replacement. However, by applying complexity filters to the target words, phrasal verbs could be replaced only if they meet certain complexity criteria, for example low frequency counts. Since the target users of PV-Sub are foreign language learners of English, it seems valid to attempt to assist learning through comprehension while simplifying a text. This could be achieved by replacing the actual substitution step of the program with a feature that suggests Latinate synonyms for phrasal verbs only when the user selects these words to be looked up, for example by providing a pop-up window when the phrasal verb is highlighted or hovered over, instead of changing the original form of the text.

The tool could be modified to target word classes other that phrasal verbs –such as nouns and adverbs–, especially if these are to be selected by a complexity filter or by the user. On a more abstract level, the algorithm used in the design of this tool could be applied to different language pairs. For example, with the significant Germanic or French influence present in the English Vocabulary, it seems plausible to look into relevant transformation rules between cognates of these language pairs and to use them in a similar manner as PV-Sub does for English and Spanish.

References

Donald L Alderman and Paul W Holland. Item performance across native language groups on the test of english as a foreign language. *ETS Research Report Series*, 1981(1):i–106, 1981.

Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pages 136–145. Springer, 2002.

Laly Bar-Ilan and Ruth A Berman. Developing register differentiation: the latinate-germanic divide in english. *Linguistics*, 45(1):1–35, 2007.

Adriane Boyd, Marion Zepf, and Detmar Meurers. Informing determiner and preposition error correction with word clusters. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 208–215. Association for Computational Linguistics, 2012.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, 1998.

Menachem Dagut and Batia Laufer. Avoidance of phrasal verbs—a case for contrastive analysis. *Studies in second language acquisition*, 7(01):73–79, 1985.

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

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83. Association for Computational Linguistics, 2011.

Anas Elghafari, Detmar Meurers, and Holger Wunsch. Exploring the data-driven prediction of prepositions in english. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 267–275. Association for Computational Linguistics, 2010.

Thomas Finkenstaedt and Dieter Wolff. *Ordered profusion; studies in dictionaries and the English lexicon*, volume 13. C. Winter, 1973.

Rafael Alejo González. L2 spanish acquisition of english phrasal verbs: A cognitive linguistic. *Corpus-based approaches to English language teaching*, page 149, 2010.

Jan H Hulstijn and Elaine Marchena. Avoidance. *Studies in second language acquisition*, 11(03):241–255, 1989.

Robert T Jiménez, Georgia Earnest García, and P David Pearson. The reading strategies of bilingual latina/o students who are successful english readers: Opportunities and obstacles. *Reading Research Quarterly*, 31(1):90–112, 1996.

Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

Andrzej Kurtyka. Teaching english phrasal verbs: A cognitive approach. 2001.

Batia Laufer and Stig Eliasson. What causes avoidance in l2 learning. *Studies in second language acquisition*, 15(01):35–48, 1993.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies*, 3(1):1–134, 2010.

Michael Lesk. 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*, pages 24–26. ACM, 1986.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

Harry Levin and Margaretta Novak. Frequencies of latinate and germanic words in english as determinants of formality. *Discourse Processes*, 14(3): 389–398, 1991.

John McWhorter. *Our Magnificent Bastard Tongue: The Untold History of English*. Gotham, 2009.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

William E Nagy, Georgia Earnest García, Aydin Y Durgunoğlu, and Barbara Hancin-Bhatt. Spanish-english bilingual students' use of cognates in english reading. *Journal of Literacy Research*, 25(3):241–259, 1993.

Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219, 2006.

Sarah E Petersen and Mari Ostendorf. Text simplification for language learners: a corpus analysis. In *SLaTE*, pages 69–72. Citeseer, 2007.

Job Schepens, Ton Dijkstra, and Franc Grootjen. Distributions of cognates in europe as based on levenshtein distance. *Bilingualism: Language and Cognition*, 15(01):157–166, 2012.

Alicia P. Schmitt. Language and cultural characteristics that explain differential item functioning for hispanic examinees on the scholastic aptitude test. *Journal of Educational Measurement*, 25(1):1–13, 1988.

E.L. Smith. *Contemporary Vocabulary*. Bedford/St. Martin's, 1995. ISBN 9780312101282. URL https://books.google.de/books?id=AlN1HAAACAAJ.

Dominic Streames. EFL theatre club, 2011. URL http://eftheatreclub.co.uk/.

Wikimedia Commons. Latin influence in english, 2007. URL http://en.wikipedia.org/wiki/Latin_influence_in_English.

A
The invitation

Uncle Bob has a neighbour called Boris. Most people don't take to Boris when they meet him. He usually comes across as a bit rude. Worst of all he tends to go on and on about the local council with whom is wages a constant battle. He refuses to recycle and he still uses a fridge full of CFC gases. The Mayor himself even came round to try to talk Boris into changing the fridge but Boris wouldn't give in. Anyway, one day Uncle Bob's girlfriend Eliza decided to throw a party and Uncle Bob decided to invite Boris. Eliza is a hippy which didn't go down well with Boris. He thinks all hippies are drug addicts. To make matters worse, the party was in aid of a local environmental group. At the party Boris managed to live up to his reputation. He explained that we should all be spaying more CFC's into the atmosphere, not less. Winter is already too cold, so a little Global Warming would do use some good. But it turned out that Boris was a big hit. He really livened up the party and in fact he went down so well that the local resident have elected him their representative. It seems that if there is ever any disagreement with the local council, as soon as they see Boris coming, the council will immediately back down.