Word similarity, cognation, and translational equivalence

Grzegorz Kondrak

1 Introduction[1]

The focus of this presentation is the following observation: *words that are phonetically similar across different languages are more likely to be mutual translations*. This phenomenon has been exploited in the past to improve various tasks in Natural Language Processing (NLP). However, to the best of my knowledge, the proposition has never been explicitly stated or justified.

The term *mutual translations* should be understood here as words that can be used to express the same meaning. In particular, words that correspond to each other on two sides of a sentence in a bilingual corpus (*bitext*) are considered translations, as well as words that are used to define each other in a bilingual dictionary.

Even though the term *phonetic similarity* is used in the above formulation, phonetic similarity is usually reflected in orthographic similarity. If the languages use different scripts, orthographic similarity can be emulated by mapping one script to another, or converting both scripts to a more universal transcription, such as the International Phonetic Alphabet (IPA). Even if the mapping is imperfect, much of the similarity will be preserved. In this presentation, however, I focus on the written forms of the words.

The fact that similar words are more likely to be translations has been utilized in various tasks within Statistical Machine Translation (SMT), such as word alignment, sentence alignment, inducing translation models, and generating translation lexicons. Another application is automatic acquisition of transliterations (transliteration mining).

The publications that utilize the principle rarely if ever articulate it or explain it. In the first part of this presentation, I discuss the reasons behind the principle, which include the prevalence of cognates, loanwords, technical terms, and proper names. In the second part, I analyze the results of comparisons between French and English that provide insights into the issue of word similarity.

2 Definitions

In this section, I propose several definitions, accompanied by rationale and examples from French and English, respectively.

Cognates are word pairs in related languages that derive directly from the same word in the ancestor language. Because of gradual phonetic and semantic changes over long periods of time, cognates may no longer look similar and have quite different meanings. E.g., *pére* / *father*, *chef* / *head*.

Loanwords (also called *lexical borrowings)* are words that have been transferred form one language to another at some point of time, such as the word *reconnaissance* in English. The languages involved in the transfer need not be related.

Names are designations of persons, organizations, and places. They are normally not found in dictionaries. In English, and many other languages, names usually start with a capital letter. Names are rarely translated into other languages; instead, they are either copied verbatim, or transliterated on the basis of their pronunciation.

Unrelated words (as opposed to **related**) belong to neither of the previous three categories. Their forms cannot be traced to any common origin. However, they can be mutual translations.

False friends (*faux amis*) are pairs of words across languages that look or sound similar but have different meanings. In many cases, the similarity is purely accidental, e.g. *main* 'hand', but some false friends are cognates that have undergone semantic shifts, e.g. French *suave* 'sweet'.

True friends (*vrais amis*) are words that look or sound similar *and* are mutual translations. The words with identical spelling are called *homographs*. Aside from cognates and loanwords, true friends can be traced to nursery terms, onomatopoeia, and even accidental similarity.

Partial friends are similar words that have the same meaning in some, but not all, contexts. They are either true or false friends depending on the context. For example, *facteur* in French signifies not only 'factor' but also 'mailman'.

3 Theoretical view

We can classify cross-language word pairs according to the following three criteria: common origin, semantic similarity, and similarity of form (either phonetic or orthographic).

The three criteria vary in terms of the subjectivity and granularity. The first criterion is binary: words either have the same origin or not. The ones that do include cognates, loanwords, as well as names. In most cases, this can be established objectively. The other two criteria involve similarity, which is usually a subjective notion, and falls on a spectrum ranging from total synonymy to nothing in common. Here, I map semantic similarity onto a binary notion of *translatability*, which can be approximated by a bilingual dictionary look-up.

In order to convert form similarity into binary relation, I employ two imperfect projections: *identity* and *thresholding*. Identity is not subjective, but encompasses only a small subset of word pairs that are clearly similar. Thresholding, on the other hand, involves an arbitrary choice of a similarity measure and a threshold, which results in a relation that only partly correlates with human judgment. However, even human annotators would undoubtedly have difficulty with sharp demarcation of similar vs. dissimilar words.

The application of the above three criteria to the classification of cross-lingual word pairs produces eight categories listed in Table 1, with Spanish-English examples.

Machine translation specialists, who aim at exploiting form similarity to find translations, are mainly interested in true friends, which correspond to categories (1) and (3). The difference between these two classes in immaterial for them. However, they need to avoid false friends, which are covered by categories (5) and (7).

Table 1. Classification of cross-lingual word pairs.

	Criteria			Example	
	Translations	Related	Similar	Spanish	English
1	+	+	+	*sal 'salt'*	*salt*
2	+	+	-	*pie 'foot'*	*foot*
3	+	-	+	*mucho 'much'*	*much*
4	+	-	-	*sangre 'blood'*	*blood*
5	-	+	+	*muerte 'death'*	*murder*
6	-	+	-	*carbon 'coal'*	*hearth*
7	-	-	+	*flor 'flower'*	*floor*
8	-	-	-	*fruto 'fruit'*	*door*

For historical linguists, on the other hand, who are interested primarily in identifying cognates, category (3) is the treacherous one. Fortunately, it contains few pairs, for it is unusual for unrelated words to converge in both form and meaning. However, distinguishing between cognates and loanwords is often difficult. Regular sound correspondences are helpful for this purpose.

Closely related languages, such as Spanish and Italian, contain numerous cognates, most of which fall into category (1). The more remote the relationship, the greater fraction of cognates falls into categories (2), (5), and (6).

Figure 1 shows the situation in a schematic way. The points on the graph represent pairs of words from two related languages. The black points denote cognates, while the white points denote unrelated words. The two axes correspond to the semantic and phonetic similarity. The pairs that are on the vertical axis have identical meanings (*synonyms*), while the pairs on the horizontal axis are identical in form (*homonyms*). The origin point of the graph is reserved for pairs that have exactly the same form *and* meaning.

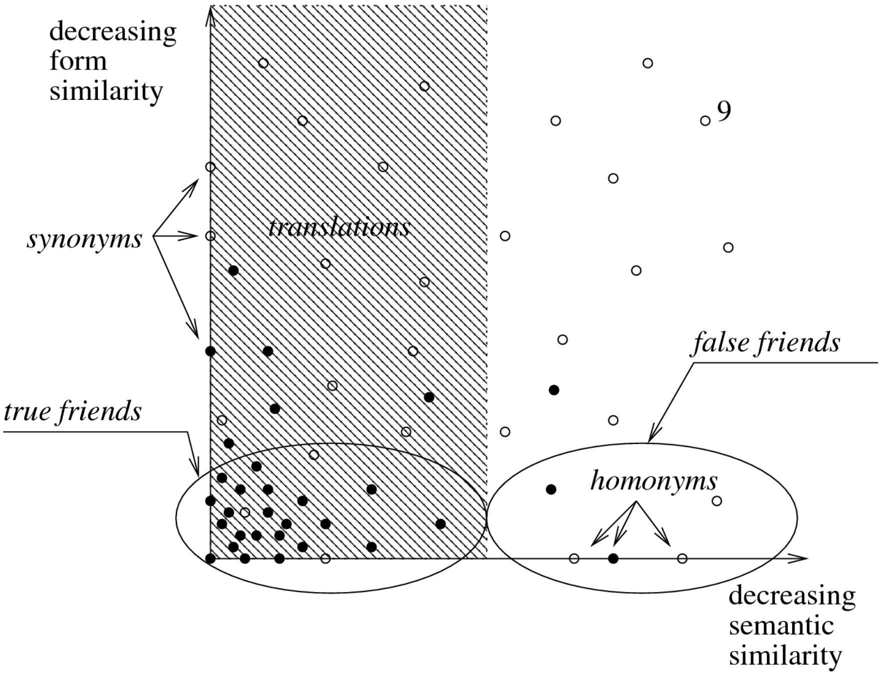

Figure 1. *A schematic depiction of the set of cross-lingual word pairs.*

The distribution of the unrelated words is shown as fairly uniform. Cognates, on the other hand, are clustered mostly in the vicinity of the plot origin. At the very moment of a language split, all word pairs are considered to be at the origin point. With time, cognates undergo phonetic and semantic changes that slowly cause them to disperse further and further away from the origin. Loanwords also start their existence in the vicinity of the origin point, but may move away from it. Names, on the other hand, tend to remain at the origin point.

4 Empirical analysis

In this section, I investigate the distribution of similar French-English word pairs in two resources: a dictionary and a bitext. The objective is to provide some empirical evidence for the observation stated in the introduction.

4.1 Preliminaries

In order to compute concrete statistics, we need to make a clear distinction between names and other words. This is not always straightforward (e.g., *Greek).* I adopt the following rule of thumb: a name must be capitalized and must not be listed in a standard dictionary.

The algorithm for deciding whether two morphologically complex words are related is slightly more complicated. The focus here is on the roots of the words, without paying much attention to affixes. If the roots are related, the words are considered related. In the case of compounds, one common root is sufficient.

A simple string similarity measure can be used to emulate the notion of orthographic similarity. The Longest Common Subsequence Ratio (LCSR) of two strings is defined as the length of their longest common subsequence normalized by the length of the longer string. It returns a value between 0 (no similarity) and 1 (identity).

4.2 Dictionary entries

The starting point of the analysis presented in this section is an automatically generated phrase translation table containing about eighty thousand French-English phrase pairs, such as *anneau de caoutchouc / rubber ring.* I limit my analysis to 8012 entries that map single French words to single English words. I manually annotated a randomly-selected sample of 1000 entries as either *related* or *unrelated*. Surprisingly, the related pairs are more frequent in the sample: 636 vs. 364. Of the related pairs, 140 are homographs, e.g. *horizontal*.

In the Cartesian product of the 1000 entries, which is composed of one million pair, there are 141 homographs, of which 140 are the translations mentioned above. All homograph translations are related words.[2] They consist mostly of words of Latin origin (e.g. *constellation*), but also quite a few direct loanwords (e.g. *folklore*, *chalet*, *cousin),* as well as words from other languages (e.g. *eldorado*). The one non-tranlational pair (*but* 'target') are unrelated false friends. Another pair that are almost identical, differing only by a single diacritic involves the French word to cur*é* 'parson', which evolved from the same Latin root as English *cure*. In this relatively small sample, all but one of the homographs across French and English are mutual translations.

Using the same Cartesian product approach, I identified 1097 homographs in a larger set of 8012 entries. 1016 (92.6%) of them correspond to two sides of the same dictionary entry. This means that homographs are over ten times more likely to be mutual translations than not, even under the assumption that all remaining pairs are false friends. Note that the set is likely to contain most of the common words in both languages.

I analyzed the remaining 81 homographs that occur in separate dictionary entries. The set contains 53 (65%) related and 28 unrelated pairs. The latter category is entirely composed of false friends. Many are function words from one language paired with a content word from another languages, e.g. *car* 'because', *or* 'gold'. Some pairs originate from different Latin words, e.g. *court* 'short', with the French and English words derived from *curtus* and *cohort*, respectively.

Most of the 53 related pairs are words of Latin origin. Some of them are partial friends, e.g. i*nexcusable*, which is translated as 'unpardonable'. A few pairs are clearly false friends, e.g. *concussion* 'embezzlement'. The majority of the pairs, however, are at various stages of semantic shifts, with partly overlapping meanings, such as *index*. A number of pairs contain words that belong to different parts of speech, such as *absorbent*, which is a verbal form in French, but a noun in English. In addition, there are some modern loanwords, such as *film* and *attaché*, which are partial friends.

The average length of the homographs measured in letters is 7.5 for related homographs, but only 4.1 for accidental homographs. This is because the latter are more likely to be short words. The longer the homographs, the more likely they are to be mutual translations.

The average similarity of translations in the 1000-entry sample according to the LCSR measure is 0.619. The corresponding values for the subset of related translations and unrelated translations are 0.815 and 0.276, respectively. Interestingly, the latter value is substantially larger than the average similarity of random pairs of French-English words, which is approximately 0.235 (ignoring diacritics). This shows that even disregarding directly related translations, translatability and similarity are not completely independent. The reasons behind the apparent correlation are discussed in Section 5.

4.3 Bitext alignment links

The resource that underlies the analysis in this section is the Blinker corpus (Melamed 1998), a word-aligned French-English bitext composed of 250 Bible verses (non-continuous). It contains 7510 English word tokens and 8191 French word tokens. The alignment links that associate words across two sides of the bitext are quite accurate. The total number of links is 10097.

I classified all 967 related word pairs in the corresponding sentences of the Blinker corpus as cognates (10%), loanwords (47%), and proper names (43%).[3] The cognates are words that go back all the way to Proto-Indo-European. Apart from numerals, these include words which have changed dramatically over the five thousand years, such as *coeur* / *heart* and *oeil* / *eye*.[4]

The number of pairs that are related and at the same time linked (aligned) in the bitext is 386 (40%). The remaining 60% of related pairs are not explicitly linked because they do not correspond to each other in the translated sentences, but the majority of them are mutual translations in the dictionary sense. For example, sentence #227 contains the word *division* three times on both sides of the bitext, inducing nine cross-lingual pairs, but only three of them are actually linked as corresponding translations.

If we disregard function words, all 174 instances of the homographs are related words. The majority of such pairs are names, but there are also many common words, such as *province* and *temple*. A few short function words, such as *a* and *on* are identical false friends, but these can be filtered by employing relatively small lists of function words. The closest an unrelated pair of content words comes to identity is French *cent* 'hundred' and English *sent*.[5]

We can try to binarize the subjective and continuous notion of form similarity by thresholding the LCSR measure st 0.66. In Blinker, 660 pairs exceed that threshold, out of which 631 (96%) are related. The set of 29 false friends includes accidental similarities across different parts of speech (*temple* / *remplie* 'full'), names (*Izharites* / *Amramites*), and shared affixes (*desolations* / *dévastations*).

In comparison, 92% of pairs in an independent list of 326 French-English false friends (Inkpen, Frunza, and Kondrak 2005) exceed the threshold of 0.66, which confirms that this thresholding approach is effective at identifying a great majority of words that are perceived as similar.

5 Discussion

The results in Section 4.2 suggest that unrelated translations exhibit greater similarity than random bilingual pairs, which seems to contradict the Saussurean principle of the arbitrariness of the linguistic sign. In order to confirm this observation, I performed another experiment using eight lists of 200 basic words in phonetic notation compiled by (Kessler 2001), which henceforth I refer to as *Kessler's set*. The lists represent Albanian, English, French, German, Latin, Hawaiian, Navajo, and Turkish. All cognates and borrowings are carefully annotated in the data. In Kessler's set, the average LCSR similarity value for 5029 unrelated translations is 0.142, whereas the corresponding value for over one million of pairs of words belonging to different languages is 0.129.

Another, independent confirmation of the phenomenon is provided by (Wichmann et al. 2010b). After analyzing the lists of 40 basic words across over ten thousand pairs of unrelated language families from different hemispheres, the authors found that the words for the same concept are slightly more similar to each other than are the words for different concepts. They attribute the difference to sound symbolism, which is further investigated in (Wichmann et al. 2010a).

Here, I propose a different explanation of the phenomenon. My intention is not to deny the influence of sound symbolism, which is clearly a factor, but to suggest another reason for the observed divergence. I posit the correlation between the following word characteristics: translatability, frequency, length, and similarity. Below, I consider these in order.

The key observation is that mutual translations are on average closer in terms of their length than random words. Let us define the *length ratio* of two words as the length of the shorter word divided by the length of the longer word. The length ratio is always a value between 0 and 1. In the set of 1000 French-English word pairs described in section 4.2, the average length ratio of unrelated translations is 0.758, as opposed to the average length ratio of 0.704 of random pairs of French-English words. Similarly, in Kessler's set, the corresponding values are 0.717 and 0.692.

In general, pairs of words with smaller average length difference also exhibit higher average LCSR similarity value. The mathematical explanation is that the length of the shorter word is the upper bound for the length of the longest common subsequence, which constitutes the numerator in the LCSR formula. Therefore, the greater the difference in length between the two words, the lower is the upper bound of the LCSR value. This agrees with the intuition that the similarity of length contributes to the overall similarity of words.

What could be the underlying reason of the fact that translations tend to differ less in length than non-translations? One possibility is that words that are mutual translations have similar frequency. Intuitively, translations refer to the same semantic concept, which tends to be expressed with similar frequency across languages. In order to confirm this intuition, I collected the frequencies of all words in the French-English set described in Section 4.2. The English word frequencies were taken from the CELEX database (Bayen et al. 1996), while the French frequencies were computed from *Le Monde Diplomatique* text corpus. The total number of word tokens in either resource are around 15 million. It turns out that the translation pairs in our dictionary data set exhibit a positive correlation of 0.573 with respect to the negative logarithm of their frequencies. On the other hand, the correlation for word pairs that are not mutual translations is close to zero. These numbers strongly support the observation that there is a connection between translatability and frequency.

Finally, it is wel known that there is a connection between word frequency and length (Zipf 1936). For example, (Piantadosi et al. 2011) calculate the correlation values between 0.1 and 0.4 for each of eleven European languages including both French and English. This completes the chain of reasoning that provides an explanation for the phenomenon which has been observed in the experiments; namely, that one reason of the greater similarity of translations is their similar frequency, which in turn is reflected in their similar length. This is a hypothesis that can be tested in the future on sets containing more languages and more concepts.

6 Conclusion

In this presentation, I have provided a theoretical justification and some empirical evidence for the observation that form similarity is positively correlated with translational equivalence. Data from a bilingual dictionary and an aligned bilingual corpus show that translation pairs tend to be similar, and that similar pairs tend to be translations. An interesting consequence of this bias is that electronic dictionaries and automatically generated bitext alignment links can be used for the purpose of evaluating word similarity measures (Kondrak 2005), and that cognate detection can improve machine translation quality (Kondrak 2003). In addition, I proposed a novel explanation of the word similarity divergence between mutual translations and random pairs of words. These are just a few examples of how computational linguists, lexicographers, and statistical language processing engineers can benefit from paying attention to each other's research.

Notes

[1] I am grateful to Lars Borin and Anju Saxena for their suggestion to investigate the observation that unrelated translations exhibit greater similarity than random word pairs. Comments made by John Nerbonne and Søren Wichmann during the 2011 workshop in Gothenburg were very helpful as well. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

[2] Unrelated homograph translations do exist, but those that can't be attributed to child language (e.g. *mama* 'mother') or onomatopoeia (e.g. *miau* 'meow') are extremely rare. One example is the word *bad*, which has the same meaning in English and Persian, but apparently no common origin.

[3] The lists of pairs are available on request.

[4] A list of over a hundred French-English cognate pairs that are still mutual translations is available at http://www.cs.ualberta.ca/~kondrak

[5] LeBlanc and Seguin (1987) identified 23,160 French-English cognate pairs, including 6,447 homographs, across two general-purpose dictionaries, each containing around 70,000 words.

References

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak
2005 Automatic Identification of Cognates and False Friends in French and English. P*roceedings of the International Conference on Recent Advances in Natural Language Processing.*

Harald Baayen and Richard Piepenbrock and Leon Gulikers
1996 *The CELEX2 lexical database*.

Brett Kessler
2001 *The significance of Word Lists.* Stanford: CSLI Publications.

Grzegorz Kondrak
2003 Cognates Can Improve Statistical Translation Models. *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.*

Grzegorz Kondrak
2005 Cognates and Word Alignment in Bitexts. *Proceedings of the Tenth Machine Translation Summit*.

Raymond LeBlanc and Robert Seguin
1987 Homographes et parographes dans l'enseignement de la langue seconde. 8th AILA World Congress.

Dan Melamed
1998 Manual Annotation of Translational Equivalence: The Blinker Project. IRCS #98-07. University of Pennsylvania.

Steven T. Piantadosi, Harry Tily and Edward Gibson
2011 Word lengths are optimized for efficient communication. *Proceeding of the National Academy of Sciences of the USA* 108(9).

Søren Wichmann, Eric W. Holman and Cecil H. Brown
2010a Sound Symbolism in Basic Vocabulary, *Entropy*, 12(4).

Søren Wichmann, Eric W. Holman, André Müller, Viveka Velupillai, Johann-Mattis List, Oleg Belyaev, Matthias Urban, and Dik Bakker.
2010b Glottochronology as a Heuristic for Genealogical Language Relationships. *Journal of Quantitat*ive *Linguistics* 17:4.

George Zipf
1936 *The Psychobiology of Language.* Routledge, London..