# Measuring the semantic distance between languages from a statistical analysis of bilingual dictionaries

Martin C. Cooper *

**Abstract**

A bilingual dictionary is a valuable linguistic resource which records, among other things, the differences in the segmentation of semantic space by the two languages and hence the difficulty in producing faithful translations between the two languages. Statistical analysis of nearly a hundred dictionaries has allowed us to determine how best to measure the semantic distance between languages from bilingual dictionaries. The distribution of the number of words in language A having n translations in language B, for n=1,2,3, etc., was found to have a specific shape depending on the semantic distance between the two languages. A sample of only a thousand words was sufficient to obtain an estimate of semantic distance.

We give a theoretical justification for this distance based on models of the historical evolution of monolingual and bilingual dictionaries.

Among our linguistic findings, we discovered, for example, that French is semantically closer to Basque than to German. We envisage an application of our semantic distance measure in the choice of an intermediate language when performing indirect translation, i.e. translating from language A to language B via a third language C.

**Keywords:** polysemy, historical semantics, stochastic model, bilingual dictionaries, comparative linguistics, lexicology.

**Abbreviated title:** Semantic distance between languages

*IRIT, Université de Toulouse III, 118 route de Narbonne, 31062 Toulouse, France (cooper@irit.fr)

# 1 Semantic differences between languages

Our aim is to compare languages on a purely semantic level, to answer questions such as to what extent do speakers of different languages have different segmentations of semantic space available in which to express their thoughts, or what is the relative difficulty of producing faithful translations between different pairs of languages.

A semantic distance measure between languages can also provide some insights into historical linguistics with a more objective foundation than the search for cognates in the two languages [121]. Our measure has the advantage of not being affected by borrowing of words between languages, provided only one meaning of the word is borrowed.

We identify three main differences which may exist between the lexical semantics of two languages: the stock of notions in semantic space having a single-word name, the segmentation of semantic space (including, among many other things, notions at the semantic-grammar border such as gender, plurality, tense, level of respect), and the network of polysemy/synonymy in the two languages.

It is clear that the set of denotata which are deemed to be worthy of naming differs in different cultures, and hence in different languages. Turkish has a word (*levhimahfuz*) for 'the tablet preserved in Heaven, containing God's decrees and the destiny of everyone'. Basque has a word (*bedakatü*) for 'to turn dry sausages so that they dry better'. Hungarian has a word (*csigatenyésztö*) for 'a breeder of edible snails'. Faroese has a word (*starilsi*) for 'the act of standing and staring out into the blue' and another (*eistnasúpan*) for 'soup made from sheep's testicles'. The English words *wicket* and *quango* are equally unlikely to have a one-word equivalent in other languages.

It is well known that, even in domains common to all languages, such as colours and kinship terms, there is not a one-to-one correspondence between the notions coded by single words in different languages [113, 115, 107]. The Japanese word *aoi* means both green and blue. English kinship terms such as cousin, aunt or grandmother do not specify whether we are related to these family members via our father or our mother, whereas in many other languages this information can be given by a single word. For example, *mormor* (Danish) = 'maternal grandmother' and *hala* (Turkish) = 'paternal aunt'. Similarly, the distinction *siostrzenica*/*bratanica* (Polish) = 'sister's daughter'/'brother's daughter' is lost in the English word *neice*. The English word *cousin* can refer to any of the eight possibilities among son/daugher of a paternal/maternal uncle/aunt.

The most common and apparently simple words of a language may not have a single corresponding notion in another language. For example, the definite article *the* has 13 translations in Icelandic but no equivalent in Russian. The personal pronoun *you* often has more than one translation depending on the level of respect the speaker wishes to show to the hearer. Japanese is an extreme

case, in which the following words (in decreasing order of respect from very respectful to downright rude) all mean 'you': *anatasama*, *o-taku*, *anata*, *kimi*, *omae*, *kisama* [114].

German has two different words meaning 'to eat', depending whether the subject of the verb is a human or an animal [112]. K'ichean (Mayan) languages have as many as seven different words meaning 'to eat', depending on the object of the verb [100]. Looking up the French verb *briller*, we find 11 possible translations [67]: 'to shine, sparkle, glitter, twinkle, glint, glow, blaze, flash, glisten, be shiney, stand out'. We consider all the above examples as being caused by differences in the segmentation of semantic space by the different languages.

Another important difference between languages lies in their different networks of synonymy/polysemy. The list of meanings of a word in a given language may be extended by metaphor or by metonymy, where metaphor refers to a semantic similarity and metonymy to any association in the real world. Other unrelated mechanisms, such as ellipsis (e.g. a daily < a daily paper) or folk etymology also exist but are much rarer [96, 98]. We use the umbrella-term *association* to cover all cases. Different associations occur in different languages. For example, in Hungarian *fűles* means 'donkey, with (long) ears, with a handle', and we can guess how these three meanings may have resulted from one another by association. The Basque word *nabar* means 'multicoloured, hypocrite' no doubt by association from the concrete to the abstract sense. The 'naked' and 'communist' meanings of the Basque word *gorri* may seem unrelated until we see the complete list of meanings including 'red, pink' from which these meanings were no doubt both derived. Sometimes two languages choose the same associations, as for example the common concrete and abstract meanings of *heart* (English) and *coeur* (French).

Sweetser [123] states a convincing case for the existence of universal patterns in the generation of new meanings by association, including, among others, the well-known tendancy to generate abstract meanings from concrete ones and the tendancy to generate cognitive meanings from perceptual ones (e.g. *to see* meaning 'to understand'). However, in the majority of cases, when a new concept has to be named in two languages, the two languages will choose two different associations or at least one of the languages will produce a neologism.

## 2 Experimental observations

In trying to measure the semantic distance between languages, one might naively think that simply counting the average number of translations per word in bilingual dictionaries would provide a useful measure. For example, in the case of two languages which have evolved from a common parent language (such as French and Italian from Latin), the number of translations of each individual word $w$ tends to incease as time passes and $w$ is used in new circumstances giving rise to new opportunities for different translations. However, neologisms with

a single translation are also continually entering the source language, implying that the average number of translations per word may not necessarily increase with time.

Another important point is that the number of translations listed depends heavily on the thoroughness with which the lexicographer records polysemy in the original language and synonymy in the target language. To investigate this we counted the number of translations of ten randomly chosen words (namely *bloom, contraband, enervate, gut, laziness, nothing, prolific, screw, stampede, tramp*) in ten English-Spanish dictionaries [2, 94, 13, 87, 39, 22, 64, 9, 23, 83] and in ten English-X dictionaries for languages X = Norwegian [75], Danish [88], Swedish [89], German [10], Dutch [4], Italian [14], Catalan [17], Russian [82], Polish [11], Turkish [55]. The standard deviation of the number of of translations was actually found to be slightly greater for the set of English-Spanish dictionaries than for the set of English-X dictionaries (3.86 compared to 3.12). The average number of translations in the English-Spanish dictionaries was found to be strongly correlated with the size of the dictionary (where size means the number of headwords). The correlation coefficient was $\rho = 0.967$. Note that $\rho \in [-1, 1]$ and $\rho = 1$ only if there is a linear dependence between the two variables [126].

Even for bilingual dictionaries of comparable size and between the same pair of languages, the average number of translations listed per word is extremely variable due to lexicographic choices. To illustrate this, Figure 1 gives the number of translations given in four French-English dictionaries of similar size for a random sample of 10 French words. The size of each dictionary is given in the list of references. In a different test, we took a random sample of about one thousand words in bilingual dictionaries of comparable size. The average number of translations listed per word was 1.9 in a Basque-English dictionary [3] and 2.1 for French-English dictionaries [80, 54, 67]. In other words, bilingual dictionaries for pairs of distant languages, such as Basque and English, do not necessarily give a greater average number of translations per word than dictionaries for pairs of closer languages, such as French and English.

It is clear from this discussion that the average number of translations in bilingual dictionaries is not a reliable measure of the semantic distance between languages. However, a histogram of the number $NT(t)$ of words with $t$ translations, for $t = 1, 2, 3, \ldots$ was found to always have a distinctive shape for all bilingual dictionaries for the same pair of languages, and furthermore this shape is a good indicator of semantic distance between languages.

Figure 2 shows $NT(t)$ (plotted on a logarithmic scale) against $t$, for two English-French dictionaries. In both cases, the curve is almost a straight line. Figure 3 shows $NT(t)$ against $t$, for a French-Spanish dictionary. The curve is slightly concave.

Figure 4 shows $NT(t)$ against $t$, for X-English dictionaries for various languages X (Japanese, Basque, Turkish) all considered in comparative linguistics to be very distant from English. All three curves have a convex form. Note that

| | Larousse [63] | Le Robert [67] | Harrap's [54] | Hachette [80] |
|---|---|---|---|---|
| *apte* | 3 | 3 | 6 | 4 |
| *carabiné* | 5 | 3 | 0 | 4 |
| *croisière* | 1 | 1 | 2 | 1 |
| *emplumé* | 1 | 2 | 1 | 2 |
| *gazogène* | 1 | 1 | 2 | 1 |
| *leche-bottes* | 1 | 2 | 0 | 0 |
| *odieux* | 3 | 5 | 3 | 3 |
| *présidence* | 10 | 5 | 3 | 5 |
| *rugosité* | 4 | 6 | 2 | 2 |
| *tournis* | 3 | 3 | 4 | 0 |

Figure 1: The number of translations per word in different French-English dictionaries.

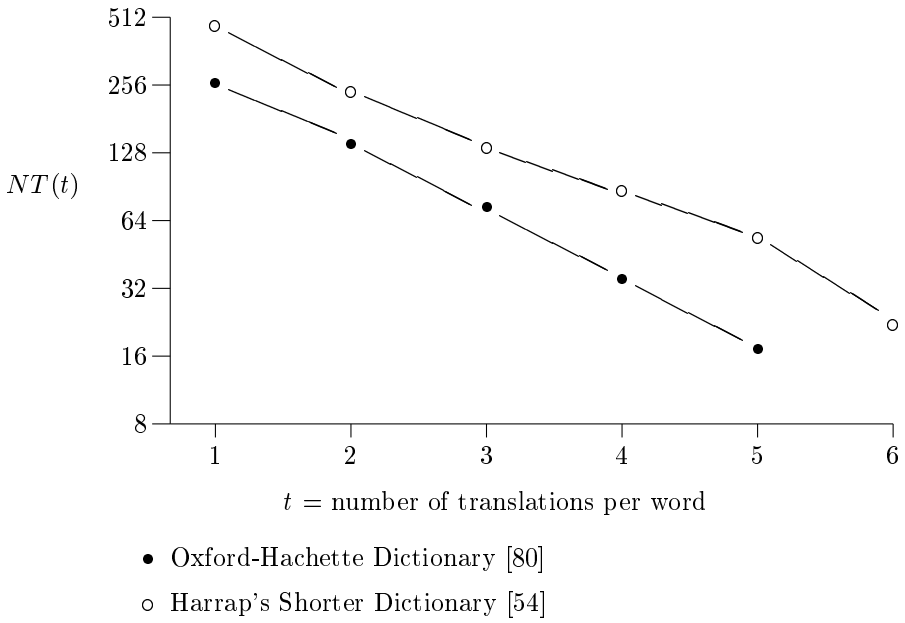


Figure 2: The number of words $NT(t)$ with $t$ translations in two French-English dictionaries.

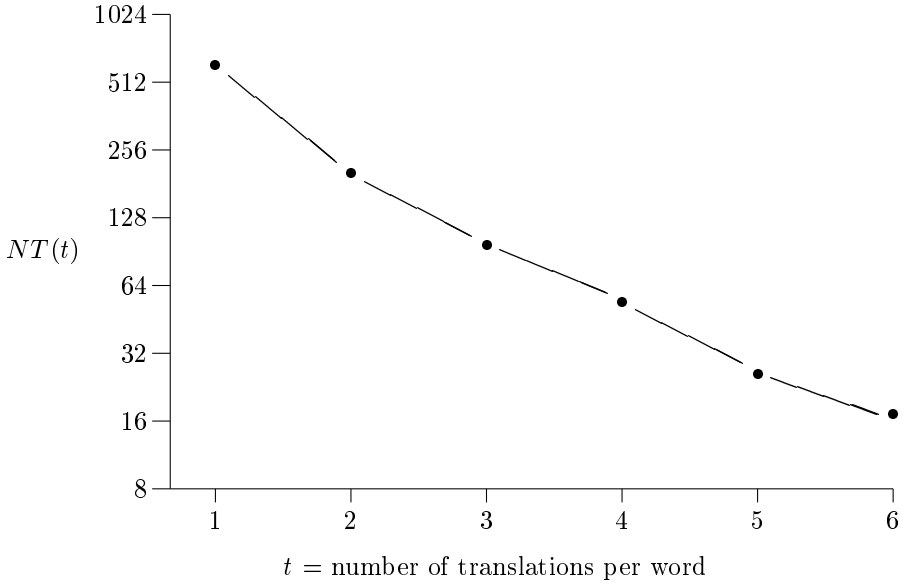


Figure 3: The number of French words $NT(t)$ with $t$ translations in Spanish (according to [61]).

it is always the shape of the left-hand side of the curve which is most important since the points on the left represent much larger samples and are hence less prone to random fluctuation than the points on the right.

In all the experiments reported in this paper, when analysing a bilingual language A - language B dictionary, we only considered entries whose headword in language A was a single word which was not a proper name, a foreign word or an abbreviation. Thus, for example, when language A was English, we did not count *Chinatown*, *haute cuisine*, *CD-ROM*, nor phrasal verbs such as *to take over*. No translations specific to idomatic expressions were included, since it is the whole expression which is translated rather than the words which compose it. For languages possessing regional variations, we restricted ourselves to just one dialect, such as British English, French as spoken in metropolitan France, Castilian Spanish and common Basque. For each dictionary, the figures quoted were obtained from a random sample of approximately 1000 words.

Given the shapes of the curves in Figures 2,3,4, we propose as a measure of semantic distance the following measure of convexity of the curve $\log NT(t)$ against $t$:

$$c = \frac{1}{N}\sum_{t=2}^{\infty}\{NT(t) - \sqrt{NT(t-1)NT(t+1)}\} \quad (1)$$

where $N = \sum_{i=1}^{\infty} NT(t)$ is the total number of words in the sample. The term $NT(t) - \sqrt{NT(t-1)NT(t+1)}$ is the distance from the point $(t, NT(t))$ to an exponential curve of the form $NT(t) = Ae^{-Bt}$ passing through the points

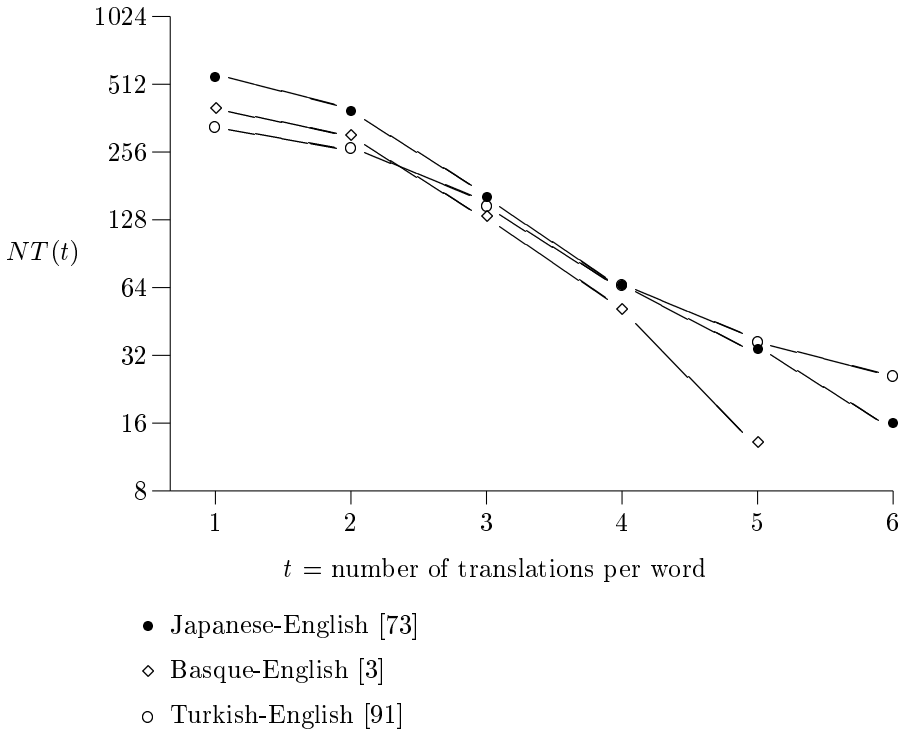


Figure 4: Number of words $NT(t)$ in language X with $t$ translations in English, for various languages X considered to be distant from English.

| language X | convexity $c$ |
|---|---|
| German [8] | -0.0177 |
| French [54, 67, 80] | -0.0114 |
| Portuguese [85] | -0.0113 |
| Spanish [49] | -0.0045 |
| Dutch [4] | 0.0189 |
| Norwegian [75] | 0.0211 |
| Faroese [46] | 0.0313 |
| Tibetan [90] | 0.0480 |
| Russian [82] | 0.0483 |
| Greek [84] | 0.0492 |
| Polish [11] | 0.0509 |
| Mongolian [70] | 0.0509 |
| Hungarian [15] | 0.0591 |
| Turkish [91] | 0.0597 |
| Estonian [40, 41] | 0.0641 |
| Japanese [73] | 0.0686 |
| Basque [3] | 0.0975 |

Figure 5: Values of convexity $c$ for X-English dictionaries for various different languages X.

$(t-1, NT(t-1))$, $(t+1, NT(t+1))$. This exponential curve corresponds to a straight line in the plots of $\log NT(t)$ against $t$. The factor $N$ in the formula for the convexity $c$ provides independence of sample size; doubling all values of $NT(t)$ $(t = 1, 2, \ldots)$, for example, does not change the value of the convexity. We can note that the convexity $c$ lies in the range -1 to +1, and usually falls close to 0. To avoid inaccuracies due to the small values of $NT(t)$ for large values of $t$, we only calculated the sum in equation (1) for $t = 2$ to 9.

Figure 5 presents the values of $c$, in increasing order as one reads down the table, calculated from random sampling of X-English dictionaries for various languages X. (The value for French-English is a weighted average of three dictionaries). There is a strong positive correlation between the order of languages in the table in Figure 5 and our expectations based on historical linguistics

[97]. Indeed, according to our measure $c$, Basque, Japanese and Estonian are semantically distant from English, whereas German, French and Portuguese are relatively close to English. We give a possible theoretical explanation of this phenomenon in Section 4.

It is interesting to note that, despite the fact that English contains twice as many words derived from Latin as from the original Germanic stock [114], our measure $c$ indicates that English still remains semantically close to German.

From the values of $c$ given in Figure 5, Basque would appear to live up to its reputation as being a truly isolated language [125], since it is the furthest, among those languages studied, from English. However, when we analyse Basque-French and Basque-Spanish dictionaries, we get a different picture. Indeed, the Basque-English curve of Figure 4 is considerably more convex than the Basque-French and Basque-Spanish curves plotted in Figure 6. The corresponding values of $c$ are given in Figure 7. These figures would seem to indicate that Basque is no closer to Spanish than to French, despite absorbing many Spanish words in recent times. More striking is the fact that the convexity of the Basque-French curves (corresponding to two different Basque-French dictionaries) is no greater than the convexity of many language pairs in which both languages belong to the Indo-european family, such as French-German or English-Polish. This is possibly explained by the fact that 75% of Basque words are of Celtic, Latin, Spanish, French or Gascon origin [114]. It is well known that the Basque language was more influenced by Latin than were languages of other regions in the Roman empire, such as, for example, Celtic, Greek or Albanian [124].

Figure 8 gives the values of convexity $c$ for French-X dictionaries, for various languages X. In the tables in Figues 5, 7, 8, when more than one dictionary was studied, the value of convexity given is a weighted average depending on the sample size from each dictionary. The lowest values of convexity in the table in Figure 8 are attained for the Romance languages, followed by the Germanic and Slavic languages before the most distant languages from French (among those studied), Arabic and Hungarian. As has been mentioned above, Basque shows an undeniable, if distant, semantic relationship with French.

Figure 9 gives the value of convexity for X-Spanish dictionaries for various languages X. Again convexity seems to accurately capture the order of semantic distance from Spanish, since the romance languages have the highest negative convexities. The exceptionally high positive value for Arabic would appear to be anomalous. It is possibly due to grouping of arabic words sharing the same root in the same entry, producing an abnormally low number of words with a single translation into Spanish.

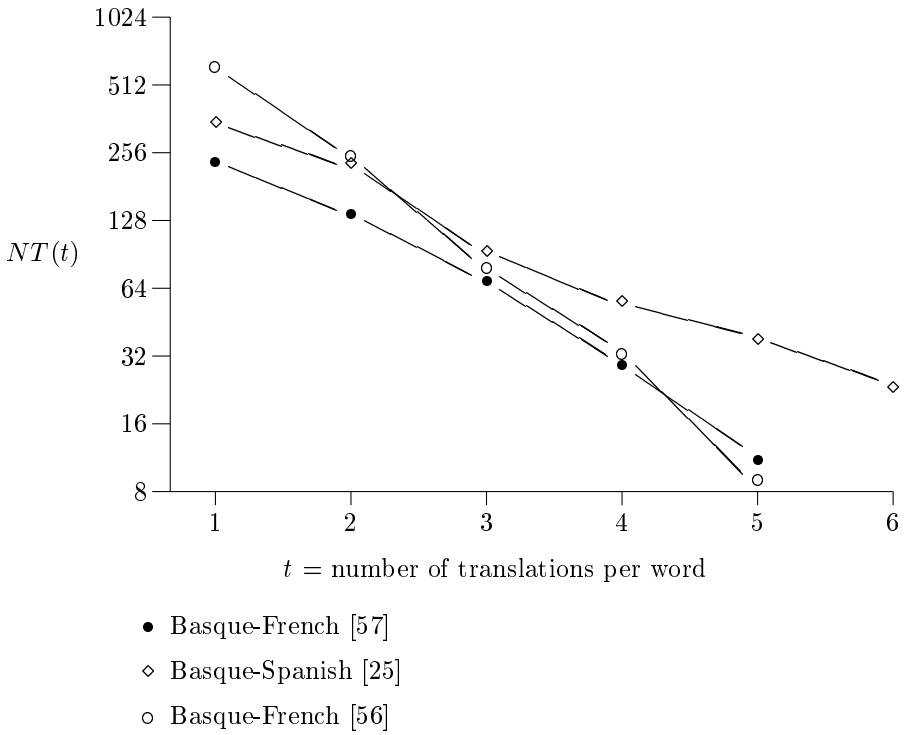


Figure 6: Number of words $NT(t)$ in Basque with $t$ translations in French/Spanish.

| language X | convexity $c$ |
|---|---|
| French [56, 57] | 0.0114 |
| Spanish [25] | 0.0322 |
| English [3] | 0.0975 |

Figure 7: Values of convexity $c$ for Basque-X dictionaries.

| language X | convexity $c$ |
|---|---|
| Spanish [61] | -0.0503 |
| Catalan [29] | -0.0460 |
| Portuguese [26, 60] | -0.0182 |
| Italian [33] | -0.0155 |
| English [54, 67, 80] | -0.0114 |
| Dutch [76] | 0.0122 |
| Basque [56, 57] | 0.0114 |
| Russian [77] | 0.0286 |
| German [51] | 0.0537 |
| Hungarian [36] | 0.0851 |
| Arabic [71] | 0.0978 |

Figure 8: Values of convexity $c$ for French-X dictionaries.

| language X | convexity $c$ |
|---|---|
| Catalan [18] | -0.0544 |
| French [61] | -0.0503 |
| Portuguese [24] | -0.0416 |
| Italian [37] | -0.0132 |
| Dutch [92] | -0.0046 |
| English [49] | -0.0045 |
| Basque [25] | 0.0322 |
| Arabic [1] | 0.2589 |

Figure 9: Values of convexity for X-Spanish dictionaries.

# 3 Monolingual dictionaries

In this section we turn our attention momentarily to monolingual dictionaries in order to be able to present (in Section 4) a theoretical justification of our semantic distance measure between languages.

We have presented in a previous paper [102] a stochastic model of the evolution of polysemy as recorded by monolingual dictionaries. Bilingual dictionaries record not only polysemy in the original language, but also synonymy in the target language together with differences in the segmentation of semantic space by the two langauges. However, we conjecture that polysemy and synonymy distributions are sufficiently similar for all languages to enable us to use the value of convexity $c$ as a measure of the semantic distance between languages.

When a lexicographer lists the different senses of a polysemous word in a monolingual dictionary, there is a certain amount of subjective choice in the number of senses listed. Different objective rules have been proposed for identifying polysemy, based on etymology, statistical analysis of collocations in corpora, the existence of zeugma (such as **there is a pen is on the table and one outside for the sheep*), the existence of different synonyms (such as *present-now*, *present-gift*), antonyms (*right-wrong*, *right-left*) or paronyms (*race-racing*, *race-racist*), and the existence of ambiguous questions (such as the canine/male ambiguity of the word *dog* brought out by the question 'Is it a dog?') [120, 99, 103, 104]. However, a quick comparison of four monolingual English dictionaries shows that there is still considerable variation in the number of senses listed per word. Figure 10 gives the number of senses listed for 10 randomly-chosen words in different English dictionaries of comparable size.

Despite this inter-dictionary variation, we observed in a previous paper [102] that the plot of the number of words $N(s)$ having $s$ senses has a similar shape for practically all monolingual dictionaries of the same language. As with $NT(t)$, the number of words having $t$ translations in a bilingual dictionary, $N(s)$ is a near-exponential function of $s$. But whereas the shape of the plot of $\log NT(t)$ against $t$ ranges from concave to convex depending on the distance between the two languages, the shape of $\log N(s)$ plotted against $s$ is always concave. Figure 11 gives the value of convexity $c$ for thirteen different English dictionaries and Figure 12 the value of $c$ for sixteen monolingual dictionaries of various other languages (French, Spanish, Basque, Catalan, German, Italian). These values were calculated from a random sample of approximately 1000 words from each dictionary.

A possible theoretical explanation for the concave shape of the $\log N(s)$ against $s$ curve was given in [102] based on a model in which the senses of a word $w$ are grouped into concepts. Each concept (group of senses) is capable of generating, for example by metaphor or metonymy, new senses for the word $w$. These potential new senses are the same for all senses corresponding to the same concept, but are essentially different for two different concepts. A concrete example will make this clearer. The word *passage* has several senses which

| | LDCE [69] | Larousse [62] | OALED [78] | OID [81] |
|---|---|---|---|---|
| *await* | 2 | 2 | 2 | 1 |
| *chine* | 2 | 0 | 1 | 4 |
| *deviate* | 1 | 1 | 1 | 2 |
| *flatten* | 2 | 3 | 2 | 1 |
| *honestly* | 2 | 3 | 3 | 0 |
| *malign* | 2 | 2 | 2 | 3 |
| *paella* | 1 | 1 | 1 | 1 |
| *putty* | 3 | 1 | 2 | 4 |
| *sex* | 6 | 3 | 4 | 5 |
| *telemetry* | 1 | 0 | 1 | 0 |

Figure 10: The number of senses listed in different English dictionaries for a random sample of 10 words.

| DICTIONARY | $c$ | $\kappa$ | $m$ |
|---|---|---|---|
| Oxford Learner's [78] | -0.0268 | 1.23 | 1.56 |
| Larousse English [62] | -0.0145 | 1.27 | 1.67 |
| Shorter OED [86] | -0.0445 | 1.28 | 2.26 |
| Collins Concise [7] | -0.0157 | 1.29 | 2.22 |
| Collins Learners [6] | -0.0257 | 1.29 | 1.74 |
| New Shorter OED [74] | -0.0138 | 1.30 | 2.26 |
| Nelson [72] | -0.0224 | 1.31 | 1.72 |
| LDCE [69] | -0.0134 | 1.32 | 2.04 |
| Oxford Illustrated [81] | -0.0183 | 1.32 | 2.46 |
| Collins School [12] | -0.0346 | 1.36 | 1.64 |
| Johnson [28] | -0.0769 | 1.55 | 1.55 |
| OED [79] | -0.0865 | 1.76 | 1.76 |
| Webster's [93] | -0.0833 | 1.98 | 1.98 |

Figure 11: The convexity $c$, average number $\kappa$ of concepts per word and average number $m$ of senses per word for different monolingual English dictionaries.

can be grouped together since they all refer to either the act of passing from one place to another (often over/across/through etc. some obstacle). Another sense of *passage*, meaning a piece of music or writing, corresponds to a different concept. If *passage* were to become a common term for the transfer of files from one computer to another, this would be by metaphor from the first concept; if *passage* were to become a common term for a few lines of a computer program, this would be by metaphor from the second concept.

The probability that a new sense entering the language receives the name $w$ is considered to be proportional to the number of concepts corresponding to $w$, rather than to the number of senses listed in any particular dictionary. The average number of concepts of a word with $s$ senses is $1 + \alpha(s - 1)$, for some constant $\alpha$ such that $0 \leq \alpha \leq 1$. This formula is derived from the fact that each word corresponds to at least one concept and each of the $s - 1$ complementary senses has a probability of $\alpha$ of corresponding to a new concept.

For any dictionary, the value of $\alpha$ can be estimated by a computer search for the best fit, in a least-squared sense, between the predicted and observed values of $N(s)$ [102]. It was found that the average number of concepts per word, calculated as $\kappa = 1 + \alpha(m - 1)$ where $m$ is the average number of senses per word, was remarkably constant for different monolingual dictionaries of the same language. This corresponds to our intuition that concepts are a property of language, whereas the division of the semantic coverage of a word into distinct senses by a lexicographer involves arbitrary choices. The values of the average number of concepts per word, for various different dictionaries, are given in Figure 11 for English and in Figure 12 for other languages (French, Spanish, Basque, Catalan, German, Italian). The reader is invited to consult [102] for details of exactly how these values were estimated.

Although there is some inter-language and intra-language variation, the average number of concepts per word nevertheless usually lies in the range 1.16 to 1.36. The exceptions are dictionaries such as the OED [79] and Websters [93] whose aim is exhaustiveness rather than being an accurate mirror of the common lexicon of educated speakers of the language. These two dictionaries contain a large number of obscure neologisms, mainly from literary sources, which simply never entered mainstream English. As examples of unlikely-looking English words from the OED, we can cite *fashionability*, *fashionabledom*, *fashionative*, *fashionary* and *floccinaucinihilipilification*. By pruning approximately 60% of the single-sense words in these dictionaries we can obtain a plot of $\log N(s)$ against $s$ which has a similar shape to the other eleven English dictionaries we studied. A similar effect is possibly responsable for the anomalously high value of the average number of concepts per word for Johnson's dictionary [28] and for the two comprehensive Spanish dictionaries [20, 21].

The values in the tables in Figures 11 and 12 indicate a possibly universal law of language that the number of concepts per word is approximately 1.25. This can be compared with the average number of senses listed per word in monolingual dictionaries which is about 1.8 (see the third column of the table

| DICTIONARY | $c$ | $\kappa$ | $m$ |
|---|---|---|---|
| Le Robert Junior (French) [68] | -0.0301 | 1.16 | 1.34 |
| Académie Française (French) [30] | -0.0265 | 1.19 | 1.64 |
| Hachette Pocket (French) [35] | -0.0244 | 1.20 | 1.53 |
| Le Petit Robert (French) [66] | -0.0094 | 1.23 | 1.83 |
| Le Grand Robert (French) [65] | -0.0368 | 1.36 | 1.79 |
| Everest (Spanish) [50] | -0.0042 | 1.21 | 1.75 |
| Gran Larousse (Spanish) [48] | -0.0214 | 1.26 | 1.77 |
| Clave Uso (Spanish) [5] | -0.0087 | 1.28 | 1.66 |
| RAE (Spanish) [20] | -0.0548 | 1.54 | 1.92 |
| Español Actual (Spanish) [21] | -0.0442 | 1.62 | 1.69 |
| Basque School (Basque) [43] | -0.0215 | 1.17 | 1.35 |
| Basque Learners (Basque [45] | -0.0105 | 1.18 | 1.36 |
| Basque Modern (Basque) [44] | -0.0247 | 1.21 | 1.38 |
| Llengua Catalana (Catalan) [19] | -0.0121 | 1.33 | 1.77 |
| Duden (German) [38] | -0.0128 | 1.16 | 1.20 |
| Zingarelli (Italian) [95] | -0.0114 | 1.32 | 1.84 |

Figure 12: The convexity $c$, average number $\kappa$ of concepts per word and average number $m$ of senses per word for different monolingual dictionaries (French, Spanish, Basque, Catalan, German, Italian).

in Figures 11 and 12). In the following section we propose a generalisation of our model of the evolution of monolingual dictionaries to a model of the evolution of bilingual dictionaries. This will not only provide a possible theoretical explanation of the shapes of the $NT(t)$ against $t$ curves given in Section 2, but will also allow us to estimate the reliability of the values of convexity $c$ given in Figures 5, 7, 8, 9.

# 4 The evolution of bilingual dictionaries

This section presents a tentative theoretical explanation for the correlation between our measure of convexity and the semantic distance between languages.

Consider a situation in which a language $L$ splits into two distinct languages $L_A$ and $L_B$ at some time $T$. We postulate, given the apparently universal property of the concave shape of the plot of $\log N(s)$ against $s$ (demonstrated by the negative values of $c$ in Figures 11 and 12), that the language $L$ would also have this property at time $T$. Since the two languages $L_A$, $L_B$ were effectively identical at time $T$, we hypothesise that a bilingual dictionary at this time would look like a monolingual dictionary of language $L$.

Before the branching point $T$, the evolution of polysemy in language $L$ can be modelled as a Markov process [102] in which one of two events occurs at each step:

- Eventl: a new single-sense word enters the language, or
- Event2: a new sense is added to the list of senses of a word already present in the language.

An important feature of this Markov model is that the probability that a word $w$ gains a new sense (whether by metonymy, metaphor or any other mechanism) is proportional to the number of concepts represented by $w$ rather than its number of senses. Considering that each word represents at least one concept and that each new sense has the same probability $\alpha$ of corresponding to a new concept, the expected number of concepts represented by a word with $s$ senses is $1+\alpha(s-1)$. The values of $\alpha$ and $\kappa$ (the average number of concepts per word, as given in Figures 11 and 12) are related by the following formula: $\kappa = 1 + \alpha(m-1)$, where $m$ is the average number of senses per word.

We can consider that the relative probabilities of Eventl and Event2 remain constant during the period before $T$. In fact the probability of Eventl is easily shown to be simply $1/m$ [102], giving values ranging from 0.4 to 0.75 for the 29 monolingual dictionaries listed in Figures 11 and 12. However, after the branching point $T$, the relative probability of Event2 increases considerably. A bilingual dictionary always lists more translations than a monolingual dictionary of comparable coverage lists senses. This is due to the different segmentation of semantic space by the two languages, as we have seen (see Section 1) through

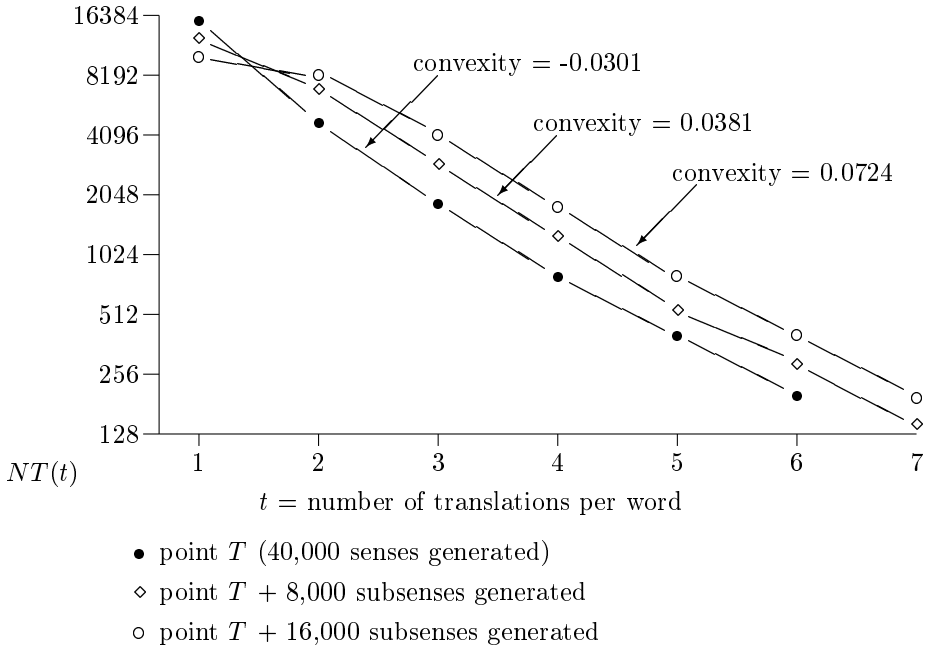


Figure 13: The number of words $NT(t)$ with $t$ subsenses (translations) after the random generation of 40,000 senses in stage 1 and up to 16,000 subsenses in stage 2 of the simulation.

examples such as the Japanese word *aoi* (blue,green) and the French verb *briller* which has eleven translations into English, depending on the subject of the verb. What is considered as a single sense when compiling a monolingual dictionary becomes a set of distinct subsenses, each with a different translation, when compiling a bilingual dictionary. A bilingual $L_A$-$L_B$ dictionary must effectively record each subsense corresponding to the intersection of the segmentations of semantic space by the two languages, since it must list each possible translation. This implies that the relative probability of Event2 increases after $T$, since Event2 is now: a new translation is added to the list of translations of a word. We can assume that the rate at which new words enter language $L_A$ after $T$ is the same as the rate at which new words entered language $L$ before $T$.

We simulated a 2-stage Markov process in which stage 1 corresponds to the evolution of a monolingual dictionary of language $L$ before $T$, and stage 2 corresponds to the evolution of the bilingual $L_A$-$L_B$ dictionary after $T$. The only difference between the two stages is that the relative probability of Event2 was greater after $T$ than before $T$.

Plotting the values of $NT$ after generating 0, 8000 and 16000 subsenses (translations) in stage 2, produced the range of curves shown in Figure 13. The

parameters of stage 1 (namely $u$ the relative probability of Eventl and $\kappa$ the average number of concepts per word) were chosen so that the $NT$ curve at point $T$ had the same shape as the $N(s)$ curve for monolingual dictionaries. We used the values $u = 0.6$ and $\kappa = 1.27$. We were able to simulate the generation of a complete dictionary of about 24,000 words and 40,000 senses. The parameters of stage 2 were chosen so as to obtain curves of approximately the same slope and convexity as the $NT$ curves of bilingual dictionaries. We used the value $u = 0.1$ for stage 2. Note that we kept the value of $\kappa$ constant throughout. In Figure 13 the value of the convexity is given next to each curve. These curves are similar to those plotted in Figures 2,3,4, 6 and hence provide a possible explanation for our experimental observations. The fact that convexity increases from -0.0301 to 0.0724 (and beyond) as the duration of stage 2 increases corresponds to the results of the experimental trials reported in Section 2: as languages $L_A$ and $L_B$ diverge the convexity corresponding to an $L_A$-$L_B$ dictionary increases.

As an alternative hypothesis, consider a situation in which language $L_A$ borrows a large number of words from language $L_B$. What would be the effect on the $NT$ curve? We assume that when a word is borrowed from language $L_B$, it is a new word in language $L_A$ with only a single sense. For example, the word *baguette* has four senses in French (according to [66]) only one of which has accompanied the word into English, namely the 'loaf of french bread' sense. Under this assumption, borrowing of words has absolutely no effect on the $NT$ curve. The actual form of the word is irrelevent. This explains, for example, why the Japanese-English curve has a high convexity despite the fact that Japanese has borrowed over 3000 words from English. We conclude that the greatest similarity between the underlying structure of two languages occurs when they share a common parent, since borrowing often means taking only one meaning of a word.

However, the history of English demonstrates that languages do not necessarily descend from a single parent language and that languages may, in fact, converge [111]. When languages co-exist in a population during a long period of time, a phenomenon of mass borrowing can occur in which the language $L_A$ takes on some of the polysemy existing in language $L_B$. This kind of relationship between $L_A$ and $L_B$ will tend to decrease the value of convexity. This probably explains the values of convexity for English-French, Basque-French and Basque-Spanish.

We have seen that the model presented in this section, of a parent language $L$ splitting into two languages $L_A$ and $L_B$, does not accurately model all cases. Furthermore, the well-known criticisms [100, 101, 108] of lexicostatistics [116, 122] also apply here. In particular, pairs of languages may diverge at different rates, meaning that two languages which broke away from each other a long time ago may be more similar than two languages whose split occurred more recently. In our model, time is measured not in years but in the number of new subsenses (translations) introduced, which is obviously dependent on the rate of change of society as a whole (technological innovations, political upheavals,

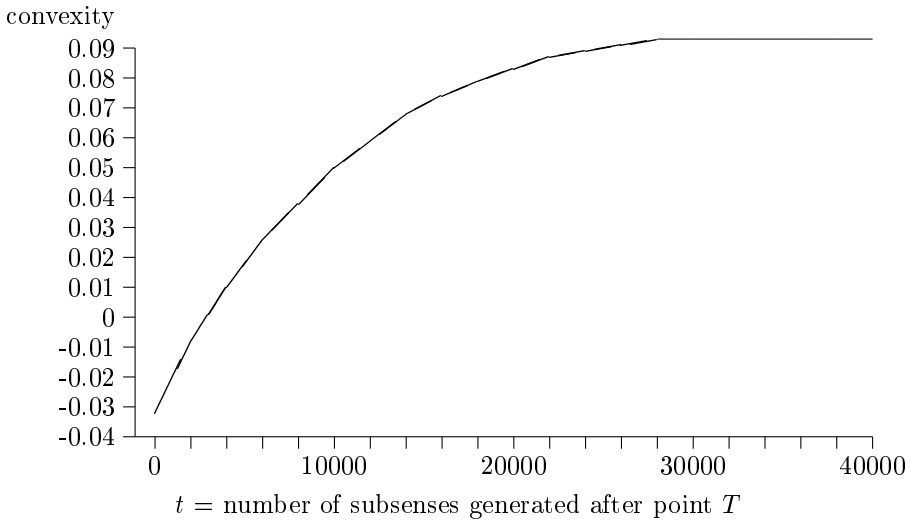


Figure 14: Convexity as a function of the number of subsenses (translations) generated in stage 2 of the simulation.

contact with other cultures, etc.).

We should also bear in mind that the value of convexity is not a linear function of the number of subsenses introduced into a language. Figure 14 shows a plot of convexity against number of subsenses generated in stage 2 of our computer simulation (in which a language $L$ splits into languages $L_A$ and $L_B$ as described above). Each point represents an average of 10 trials. The rate of increase of convexity decreases with time until convexity gradually flattens off tending towards a limit value.

# 5 Reliability of estimating convexity

An obvious question concerns the reliability of the values of convexity reported in Section 2 obtained from samples of approximately 1000 words. In order to estimate the random variation due to sampling, we took random samples of various sizes, ranging from 250 to 2000 words, of the output of our computer simulation (as described in Section 4) of the generation of a 24,000-word bilingual dictionary. For each sample size $s$, we generated 20 different random samples. The standard deviation of the convexity $c$ calculated from these 20 samples provides a good estimate of the reliability of the value of $c$ obtained from a sample of $s$ words. Figure 15 is a plot of the standard deviation of convexity against sample size. The values of the standard deviation were found to be independent of the

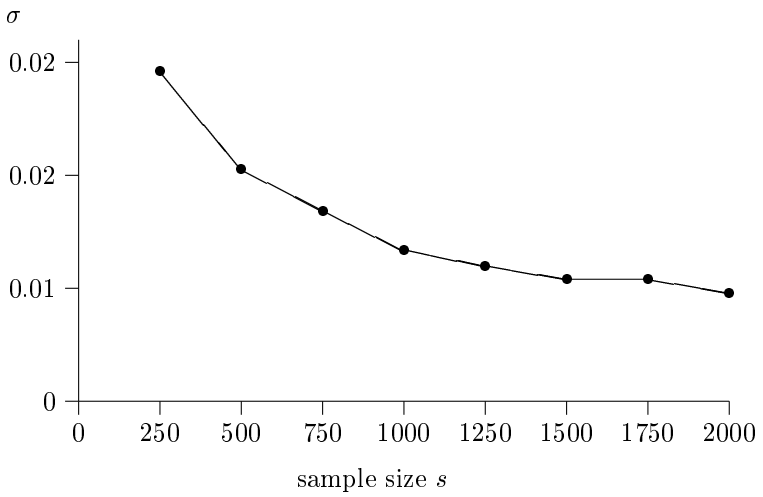


Figure 15: Standard deviation $\sigma$ of convexity as a function of sample size $s$.

parameters of the simulation and hence of the actual value of the convexity.

We can note that the variations observed in the values of convexity $c$ for the same pair of languages but calculated from different dictionaries are entirely consistent with the values of the standard deviation of $c$ given by Figure 15. Specifically, the values of convexity (calculated from a sample of $N$ words) were -0.0188 ($N = 1064$), -0.0091 ($N = 1024$), 0.0033 ($N = 560$) for the three French-English dictionaries [54, 67, 80]; 0.0733 ($N = 789$), 0.0461 ($N = 413$) for the two English-Estonian dictionaries [40, 41]; 0.0175 ($N = 1003$), 0.0313 ($N = 489$) for the two Basque-French dictionaries [56, 57]; -0.0159 ($N = 471$), -0.0189 ($N = 1112$) the two French-Portuguese dictionaries [26, 60]. We observed that such variations tend to decrease for larger sample sizes.

A similar remark holds for the variations in the values of convexity given in Figures 11 and 12 for the monolingual dictionaries, providing we treat as outliers those dictionaries [20, 21, 28, 79, 93] whose aim is exhaustivity and which include many words which are not part of the mainstream language (as discussed in Section 3).

As an extra check on the reliability of our results, for certain dictionaries we performed experiments on both halves of the same bilingual dictionary with the following results: $c = 0.0505$ ($N = 631$) for English-Polish and $c = 0.0509$ ($N = 858$) for Polish-English [11], $c = 0.0854$ ($N = 716$) for French-Arabic and $c = 0.1043$ ($N = 569$) for Arabic-French [71], $c = -0.0098$ ($N = 750$) for French-Basque and $c = 0.0313$ ($N = 489$) for Basque-French [57], $c = -0.0567$ ($N = 734$) for French-Catalan and $c = -0.0226$ ($N = 330$) for Catalan-French [29]. Again the differences in convexity were consistent with random variations.

However, it should be noted that this is not the case for languages with complex morphology. For example, fusional languages, such as German, contain many words which correspond to a multiword expression in other languages. This tends to increase the value of $NT(1)$ and hence decrease the value of convexity. We effectively encountered this phenomenon with the fusional languages Hungarian [58], Estonian [42] and Finnish [47]. The values of convexity for these three X-English dictionaries were all approximately $-0.01$ which would imply an inexplicable semantic closeness between English and these three members of the Uralic family of languages [110]. The problem of fusional languages can be easily avoided by calculating convexity from the values of $NT$ obtained from the other half of the same dictionary (for example, English-Hungarian instead of Hungarian-English). This is, indeed, how we calculated the reported values of convexity for fusional languages. Note, however, that we have no method for calculating the semantic distance between two fusional languages.

In general, basic knowledge of the characteristics of the languages under study is important to avoid any systematic bias in the estimate of convexity. For example, the fact that Basque has several dialects [125] and that Greek often has two words for the same notion (one in demotic or 'common Greek' and the other in katharevusa or 'purist Greek' [84]) both may tend to increase the number of translations per word in English-Basque or English-Greek dictionaries. Note that, for these particular languages we avoided any potential problem by counting the number of translations into English. It goes without saying that convexity $c$ should be calculated from data obtained from general-purpose bilingual dictionaries. In experimental trials on small dictionaries for tourists [32, 59] we found the value of $c$ to be close to 0 ($-0.0151$ and 0.0106, respectively) and independent of the distance between the two languages. For technical French-English dictionaries [27, 34, 31, 53] the value of $c$ was always a large negative number ($-0.0600$, $-0.0432$, $-0.0397$, $-0.0382$, respectively), whereas $c$ was a large positive number (0.0731) for a French-English slang dictionary [52], and in all cases was unrelated to the value of $c$ ($-0.0114$) obtained from general-purpose French-English dictionaries [67, 80, 54].

One standard bilingual dictionary [16] also gave what would appear to be an anomalous value of convexity (-0.0027). This Serbo-Croat - English dictionary lists approximately twice as many translations per word as most of the other bilingual dictionaries studied (3.94 compared with an average of 2.29 for the other dictionaries). We therefore ignored this dictionary.

In the light of the relationship between the standard deviation of convexity and sample size, illustrated by Figure 15, we re-examined the values of convexity reported in Section 2. The values of convexity $c$ can be replaced by a 66% confidence interval $[c - \sigma, c + \sigma]$, where $\sigma$ is the standard deviation of $c$. The value of $\sigma$ was read off the plot of Figure 15 as a function of the sample size. Figure 16(a) shows the 66% confidence intervals for each of the X-English dictionaries of Figure 5. The interval for French-English is much tighter than for other languages, since the value of $c$ is the average of values obtained from

sampling three different dictionaries. Figures 16(b),(c),(d) show the 66% confidence intervals for the values of convexity for X-French, X-Basque, X-Spanish dictionaries, respectively, for various languages X.

# 6 Linguistic conclusions from the experiments

We conclude from the non-intersection of the confidence intervals in Figure 16(a) that Basque is semantically the furthest language from English among all the languages studied, and that English is further from Estonian, Greek, Hungarian, Japanese, Mongolian, Polish, Russian, Tibetan and Turkish than from French, German, Portuguese and Spanish. However, we cannot say with any confidence which of these latter four languages is semantically closest to English, since their confidence intervals all overlap.

Many of the conclusions we can draw from Figure 16(b) are hardly surprising. For example, French is semantically closer to Spanish than to English, but French is closer to English than to Arabic or Hungarian. It is, however, noteworthy that French appears to be semantically closer to Basque than to German. Figure 16(c) illustrates without any doubt that Basque has an undeniable, albeit distant, relationship with both French and Spanish.

As a direct application of our results, consider a situation in which a document must be translated from Hungarian into German. If no Hungarian-German translator is available, the obvious solution is to pass via another language such as English or French. The results presented in Figure 16 indicate that the translation would be more reliable if the choice of intermediate language was English. On the other hand, for a Basque-Portuguese translation, we conclude that the best choice of intermediate language is French.

Of course, in indirect translation, since the intermediate-language version of the text is only intended to be read by a translator, it could be annotated with grammatical or semantic information to reduce ambiguity, as in the following English sentences.

- I saw a mother [NOUN] bear [VERB] with its baby.
- The [*government*] minister left the [*political*] party.

In machine translation (MT), some workers have even developed a completely new intermediate language for use in multi-lingual MT systems [106]. Indirect machine translation will no doubt become more and more common as individual access to the internet becomes a truly global phenomenon.

# 7 Conclusion

Analysis of the histogram of the number of words in language $A$ with $n$ translations in language $B$, for $n$=1,2,3, etc, provides a new statistical method in

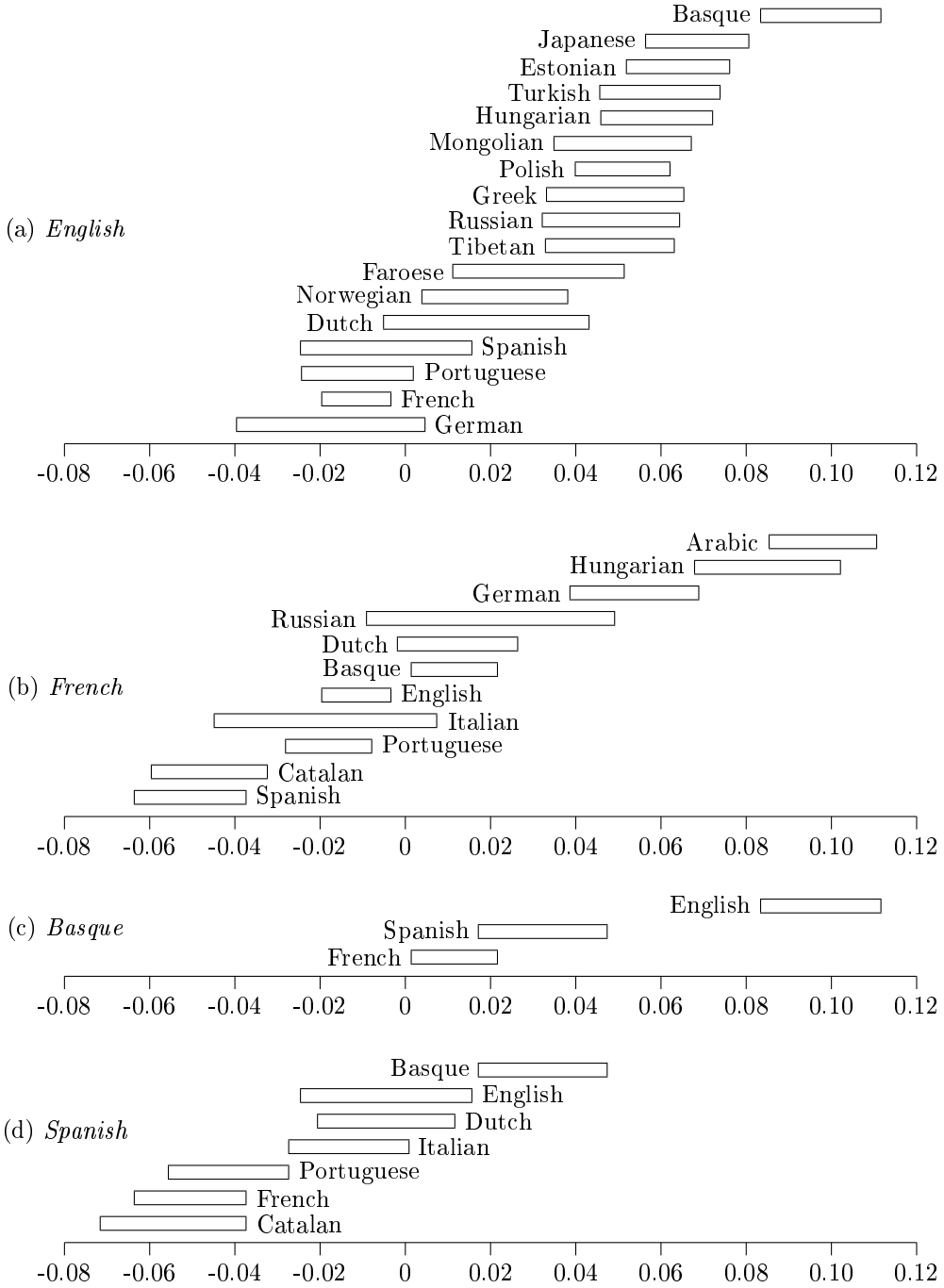


Figure 16: 66% confidence intervals for values of convexity for language pairs (a) X-English, (b) X-French, (c) X-Basque, (d) X-Spanish.

comparative linguistics. Taken together with a mathematical model for the divergence of languages, it is a potentially powerful tool in historical linguistics. Samples of only a thousand words are sufficient to estimate the semantic distance between the two languages, provided that for each of these words in language $A$ we have a list of its possible translations into language $B$.

The distance measure between languages thus derived is based only on semantics and hence provides a source of information which is completely independent of other aspects of language, such as phonetics, morphology or grammar. It is a purely objective measure whose accuracy can be improved by simply taking larger samples or by examining different sources (such as other bilingual dictionaries between the same two languages). Further research is required to determine to what extent our dictionary-based measure can be combined with other lexicostatistical approaches based on the comparison of morphology [118, 101], core vocabulary [122, 109] or parallel texts [116, 117].

A subject for future research is the question of whether the distance measure proposed in this paper can be improved by the development of more refined models of the evolution of the semantic interconnections between languages as represented, for example, by the semantic maps derived from many English-French dictionaries [119].

# References


[1] Alhambra Diccionario Arabe-Español Español-Arabe, Maurice G. Kaplanian, Editorial Ramon Sopena: Barcelona, 1996 (*c.* 12,000 words).

[2] *Amador's Handy Dictionary English-Spanish and Spanish-English*, Editorial Ramon Sopena: Barcelona, 1957 (*c.* 50,000 words).

[3] *Basque-English Dictionary*, Gorka Aulestia, University of Nevada Press: Reno & Las Vegas, 1989 (*c.* 30,000 words).

[4] *Cassell's English-Dutch Dutch-English Dictionary*, Cassell: London, 1978 (*c.* 44,000 words).

[5] *Clave Diccionario de Uso del Español Actual*, Ediciones SM: Madrid, Tercera edición, 1999 (*c.* 37,000 words).

[6] *Collins Cobuild Learner's Dictionary*, Harper Collins: London, 1996 (*c.* 24,000 words).

[7] *Collins Concise Dictionary of the English Language*, 2nd edition, Collins: London, 1988 (*c.* 37,000 words).

[8] *Collins Concise German-English Dictionary*, HarperCollins Publishers: Glasgow, 1987 (*c.* 18,000 words).

[9] *Collins Concise Spanish-English English-Spanish Dictionary*, Mike Gonzalez, Collins: Glasgow, 1985 (*c.* 29,000 words).

[10] *Collins German-English English German Dictionary*, 4th edn., HarperCollins: Glasgow, 1999 (*c.* 37,000 words).

[11] *Collins Polish-English Dictionary*, Jacek Fisiak, HarperCollins Publishers Ltd & Polska Oficyna Wydawnicza BGW: Warsaw, 1997 (*c.* 22,000 words).

[12] *The Collins School Dictionary*, Collins: Glasgow, 1989 (*c.* 17,000 words).

[13] *Collins Spanish-English English-Spanish Dictionary*, 2nd edn., Colin Smith, HarperCollins Publishers: Glasgow, 1990 (*c.* 49,000 words).

[14] *The Concise Cambridge Italian Dictionary*, Barbara Reynolds, Cambridge University Press: Cambridge, U.K. (*c.* 24,000 words).

[15] *A Concise English-Hungarian Dictionary*, 14th edn., Lásló Országh & Tamás Magay, Oxford University Press: Oxford, U.K., 1990 (*c.* 36,000 words).

[16] *Croato-Serbian-English Dictionary*, M. Drvodeli'c, Skolska Knjiga: Zagreb, 1961 (*c.* 30,000 words).

[17] *Diccionari Català-Anglès Anglès-Català*, Jordi Colomer, Editorial Pòrtic: Barcelona, 1981 (*c.* 22,500 words).

[18] *Diccionari Català-Castellà*, Enciclopèdia Catalana: Barcelona, 2nd edition, 1995 (*c.* 52,000 words).

[19] *Diccionari de la Llengua Catalana*, Enciclopèdia Catalana: Barcelona, tercera edició actuditzada, 1993 (*c.* 78,000 words).

[20] *Diccionario de la Lengua Española*, Real Academia Española: Madrid, Vigésima Segunda Edición, 2001 (*c.* 74,000 words).

[21] *Diccionario del Español Actual*, M. Seco, O. Andrés & G. Ramos, Aguilar lexicografía: Madrid, 1999 (*c.* 74,000 words).

[22] *Diccionario Inglés-Español*, Arturo Cuyás Armengol, 19th edn., Ediciones Hymsa: Barcelona, 1960 (*c.* 22,000).

[23] *Diccionario Oxford Pocket Español-Inglés Inglés-Español* , 2nd edn., ed. Sharon Peters, Oxford University Press: Oxford, U.K., 2000 (*c.* 11,000).

[24] *Diccionario Português-Espanhol Español-Portugués*, Editorial Juventud: Barcelona, 1995 (*c.* 19,000).

[25] *Diccionario Vasco-Castellano*, Placido Mugica Berrondo, Mensajero: Bilbao, 1981 (*c.* 86,000 words).

[26] *Dicionârio de Francês Portugués*, Olivio da Costa Carvalho, Porto Editora: Porto, 1997 (*c.* 53,000 words).

[27] *Dictionary of Information Technology Vol 2 Français-Anglais*, Jacques Hildebert, La Maison du Dictionnaire: Paris, 1998 (*c.* 8,000 words).

[28] *A Dictionary of the English Language*, Samuel Johnson, fascimile edition, Times Books: London, 1979 (date of publication of original edition 1755) (*c.* 40,000 words).

[29] *Dictionnaire Catalan Français*, Enciclopèdia Catalana: Barcelona, 2nd edition, 1984 (*c.* 32,000 words).

[30] *Dictionnaire de l'Académie Française A-Enz*, Editions Julliard: Paris, 1994 (*c.* 56,000 words in the complete dictionary).

[31] *Dictionnaire des Sciences et Techniques du Pétrole Anglais-Français Français-Anglais*, Magdeleine Moureau & Gerald Brace, Editions Technip: Paris, 1993 (*c.* 15,000 words).

[32] *Dictionnaire Français-Finnois Finnois-Français*, Editions Berlitz: Lausanne, 1974 (*c.* 5,000 words).

[33] *Dictionnaire Général Français-Italien*, Larousse: Paris, 1994 (*c.* 35,000 words).

[34] *Dictionnaire Technique et Scientifique Français-Anglais*, Editions H. Goursau: Saint-Orens de Gameville, France, 1996 (*c.* 12,500 words).

[35] *Dictionnaire Universel de Poche*, Hachette: Paris, 2000 (*c.* 32,000 words).

[36] *Dictionnaire Hongrois-Français*, Alexandre Eckhardt, 3e édition, Akadémiai Kiadó: Budapest, 1973 (*c.* 34,000 words).

[37] *Dizionario Fraseologico Completo Italiano-Spagnolo e Spagnolo-Italiano*, S. Carbonell, Editore Ulrico Hoepli: Milan, 1986 (*c.* 65,000 words).

[38] *Duden Deutsches Universalwörterbuch*, Dudenverlag: Mannheim, 1996 (*c.* 122,000 words).

[39] *EDAF New Comprehensive English-Spanish Dictionary*, EDAF Ediciones Distribuciones: Madrid, 1972 (*c.* 85,000 words).

[40] *English-Estonian Dictionary*, 4th edn., Johannes Silvet, TEA Kirjastus: Tallinn, 2002 (*c.* 42,000 words).

[41] *English Estonian Student's Dictionary*, K Dictionaries Ltd/OÜ Festart: Tallinn, 2001 (*c.* 22,000 words).

[42] *Estonian-English Dictionary*, Johannes Silvet, Publishing House "Eesti Raamat": Tallinn, 1965 (*c.* 31,000 words).

[43] *Europa Hiztegia - Eskola berrirakoa*, Adorez 6: Bilbao, 1993 (*c.* 24,000 words).

[44] *Euskal Hiztegi Modernoa*, Elhuyar Kultur Elkarten/ Elkar SL: Donostia (San Sebastian), 1994 (*c.* 38,000 words).

[45] *Euskara Ikaslearen Hiztegia* (Basque Learner's Dictionary), Ibon Sarasola, Vox: Barcelona, 1999 (*c.* 26,000 words).

[46] *Faroese-English Dictionary*, G.V.C. Young & Cynthia R. Cleaver, Mansk-Svenska Publishing Co. Ltd.: Peel, Isle of Man, 1985 (*c.* 18,000 words).

[47] *Finnish-English Dictionary*, Aino Wuolle, 6th edition, Werner Söderström Osakeyhtiö: Porvoo-Helsinki, 1956 (*c.* 24,500 words).

[48] *Gran Diccionario de la Lengua Española*, Larousse Planeta: Barcelona, 1996 (*c.* 68,000 words).

[49] *Gran Diccionario Español-Inglés*, Larousse: Paris, 1993 (*c.* 38,000 words).

[50] *Gran Diccionario Everest de la Lengua Española*, Editorial Everest S.A.: Madrid, Segunda Edición, 1995 (*c.* 72,000 words).

[51] *Grand Dictionnaire Français Allemand*, P. Grappin, Larousse: Paris, 1991 (*c.* 29,000 words).

[52] *Harrap's English-French Slang Dictionary*, Harrap: London, 1984 (*c.* 10,000 words).

[53] *Harrap's Informatique Dictionnaire Anglais-Français Français-Anglais*, Harrap: London, 1985 (*c.* 4,000 words).

[54] *Harrap's Shorter French-English Dictionary*, Harrap: London, 1991 (*c.* 35,000 words).

[55] *Hippocrene Standard English-Turkish Turkish-English Dictionary*, Kemal Kılıç, Kristin P. Jones, Ali Bayram, Şűkrű Meriç, Deniz Meriç, Hippocrene Books: New York, 1995 (*c.* 16,000 words).

[56] *Hiztegia II Eüskara-Français*, J. Casenave-Harigile, Editions HITZAK argitaldaria: Ossas-Suhare, France, 1993 (*c.* 25,000 words).

[57] *Hiztegia Euskara-Frantsesa/Dictionnaire Basque-Français*, P. Charriton, Elkar S.L.: Donastia (San Sebastian), 1997 (*c.* 13,000 words).

[58] *Hungarian-English Dictionary*, L. Országh, 2nd edition, Akadémiai Kiadó: Budapest, 1963 (*c.* 11,000 words).

[59] *Langenscheidt's Pocket Japanese Dictionary*, Langenscheidt: Berlin, 2001 (*c.* 11,000 words).

[60] *Larousse Dictionnaire de Poche Français-Portugais*, Larousse: Paris, 1999 (*c.* 10,000 words).

[61] *Larousse Dictionnaire Français-Espagnol*, Larousse: Paris, 1989 (*c.* 45,000 words).

[62] *Larousse English Dictionary*, Larousse-Bordas: Paris, 1997 (*c.* 29,000 words).

[63] *Larousse French-English Dictionary*, unabridged edition, Larousse: Paris, 1995 (*c.* 47,000 words).

[64] *Larousse Gran Diccionario Moderno Español-Inglés English-Spanish*, Ramón García-Pelayo, Larousse: Paris, 1983 (*c.* 67,000 words).

[65] *Le Grand Robert de la Langue Française*, 2nd ed., Le Robert: Paris, (9 volumes) 1992 (*c.* 80,000 words).

[66] *Le Petit Robert*, Dictionnaires Le Robert: Paris, 2000 (*c.* 46,000 words).

[67] *Le Robert & Collins English-French Dictionary*, HarperCollins Publishers: Glasgow, 1978 (*c.* 31,000 words).

[68] *Le Robert Junior*, Dictionnaires Le Robert: Paris, 1999 (*c.* 20,000 words).

[69] *Longman Dictionary of Contemporary English*, Longman: Harlow, Essex, UK, 1978 (*c.* 38,000 words).

[70] *Mongolian-English Dictionary*, Charles Bawden, Kegan Paul International: London, 1997 (*c.* 25,000 words).

[71] *Mounged de Poche Français-Arabe*, 10e édition, Dar El-Machreq: Beirut, Lebanon, 1983 (*c.* 15,000 words).

[72] *The Nelson Contemporary English Dictionary*, Nelson: Walton- on-Thames, Surrey, 1977 (*c.* 20,000 words).

[73] *New Crown Japanese-English Dictionary*, Sanseido: Tokyo, 1968 (*c.* 55,000 words).

[74] *New Shorter Oxford English Dictionary*, Clarendon Press: Oxford, U.K., 1993 (*c.* 78,000 words).

[75] *Norwegian Dictionary*, 2nd edn., Routledge: London, 1994 (*c.* 15,500 words).

[76] *Nouveau Dictionnaire Erasme Néerlandais - Français Français - Néerlandais*, Editions Erasme: Anvers/Bruxelles, 14e édition, 1985 (*c.* 65,000 words).

[77] *Nouveau Dictionnaire Français-Russe*, Librairie du Globe: Moscow, 1994 (*c.* 68,000 words).

[78] *Oxford Advanced Learner's Encyclopedic Dictionary*, Oxford University Press, 1992 (c, 33,000 words).

[79] *Oxford English Dictionary*, 2nd edn., Clarendon Press: Oxford, U.K., 1989 (*c.* 290,500 entries).

[80] *Oxford-Hachette French Dictionary*, Hachette Livre - Oxford University Press: Oxford, 1994 (*c.* 35,000 words)

[81] *Oxford Illustrated Dictionary*, 2nd edition, Oxford University Press: Oxford, U.K., 1975 (*c.* 50,000 words).

[82] *The Oxford Russian Dictionary*, ed. Paul Falla, Oxford University Press: Oxford, U.K., 1993 (*c.* 42,000 words).

[83] *The Oxford Spanish Dictionary*, Oxford University Press: Oxford, U.K., 1998 (*c.* 44,000 words).

[84] *The Pocket Oxford Greek Dictionary*, Julian T. Pring, Oxford University Press: Oxford, U.K., 1995 (*c.* 17,000 words).

[85] *A Portuguese-English Dictionary*, J.L. Taylor, George G. Harrap & Co. Ltd.: London, 1958 (*c.* 56,000 words).

[86] *Shorter Oxford English Dictionary*, Clarendon Press: Oxford, U.K., 1933 (*c.* 63,000 words).

[87] *Simon and Schuster's International Dictionary English-Spanish Spanish-English*, ed. Tana de Gámez, Simon and Schuster: New York (*c.* 79,000 words).

[88] *The Standard Danish-English English-Danish Dictionary*, ed. Jens Axelsen, Cassell Publishers Ltd: London, 1984 (*c.* 65,000 words).

[89] *The Standard Swedish-English English-Swedish Dictionary*, Cassell: London, 1985 (*c.* 38,000) words).

[90] *Tibetan-English Dictionary*, Stuart H. Buck, The Catholic University of America Press: Washington D.C., 1969 (*c.* 15,000 words).

[91] *A Turkish-English Dictionary*, H.C. Hony, 2nd edition, Oxford University Press: Oxford, U.K., 1957 (*c.* 16,000 words).

[92] *van Dale Handwoordenboek Nederlands-Spaans*, Peter Jan Slagter, Van Dale Lexicografie: Utrecht/Antwerp, 1994 (*c.* 61,500 words).

[93] *Webster's Third New International Dictionary*, Encyclopaedia Britannica: Chicago, 1986 (*c.* 156,000 words).

[94] *Williams Spanish-English English-Spanish Dictionary*, Edwin B. Williams, 2nd edn., McGraw-Hill: New York, 1993 (*c.* 48,000 words).

[95] *lo Zingarelli 2001, Vocaborario della Lingua Italiana*, di Nicola Zingarelli, Zanichelli: Bologna, 2001 (*c.* 75,000 words).

[96] John Algeo "Vocabulary" in *The Cambridge History of the English Language*, Vol IV, ed. S. Romaine, Cambidge University Press: Camridge, UK, 1998, pp. 57–91.

[97] John M. Anderson, "Historical Linguistics", *The Linguistic Encyclopedia*, 2nd edition, ed. Kirsten Malmkjaer, Routledge: London, 2002.

[98] Raimo. Anttila, *Historical and Comparative Linguistics*, Current Issues in Linguistic Theory 6, John Benjamins Publishing Company, Amsterdam/Philadelphia, 2nd edn., 1989.

[99] John R. Ayto, "On specifying meaning : semantic analysis and dictionary definitions", in Reinhard R.K. Hartmann, ed., Lexicography: Principles and Practice, Academic Press: London, 1983, pp. 89–98.

[100] Lyle Campbell, *Historical Linguistics: An Introduction*, Edinburg University Press, Edinburg, U.K., 1998.

[101] Lyle Campbell, "How to Show Langauges are Related: Methods for Distant Genetic Relationship", *The Handbook of Historical Linguistics*, ed. Brian D. Joseph & Richard D. Janda, Blackwell Publishing: Oxford, U.K., 2003, pp. 262–282.

[102] Martin C. Cooper, A mathematical model of historical semantics and the grouping of word-meanings into concepts, *Computational Linguistics* 31(2), 2005, pp. 227–248.

[103] D.Alan Cruse, *Lexical Semantics*, Cambridge University Press: Cambridge, UK, 1986.

[104] D.Alan Cruse, "Polysemy and related phenomena from a cognitive linguistic viewpoint", in *Computational Lexical Semantics*, ed. Patrick Saint-Dizier & Evelyne Viegas, Cambridge University Press: Cambridge, UK, 1995, pp. 33–49.

[105] Alan Cruse, *Meaning in Language: An Introduction to Semantics and Pragmatics*, 2nd Edition, Oxford University Press: Oxford, U.K., 2004.

[106] Bonnie J. Dorr, *Machine Translation: A View from the Lexicon*, MIT Press: Cambridge, Mass., U.S.A., 1993.

[107] Joachim Grzega, Bibliography of Onomasiological Works, http://wwwl.ku-eichstaett.de/SLF/EngluVglSW/OnOn-7.pdf.

[108] Istvan Fodor, "The rate of linguistic change; limits to the application of mathematical models in linguistics" in Jacek Fisiak, ed., *Linguistic Change Under Contact Conditions*, Mouton: The Hague, 1965.

[109] Brett Kessler, *The Significance of Word Lists*, CSLI Publications, 2001.

[110] Kenneth Katzner, *The Languages of the World* , 3rd edn., Routledge: London, 1995.

[111] Sydney Lamb, *Language and Reality*, Continuum: London, Chapter 25, 2004.

[112] Leonhard Lipka, *English Lexicology*, Gunter Narr Verlag: Tübingen, 2002.

[113] Sebastian Löbner, *Understanding Semantics*, Arnold: London, 2002.

[114] Michel Malherbe, *Les Langages de l'Humanité*, Editions Seghers: Paris, 1983.

[115] Kirsten Malmkjaer, "Semantics", *The Linguistic Encyclopedia*, 2nd edition, ed. Kirsten Malmkjaer, Routledge: London, 2002.

[116] Witold Manczak, "Glottochronology and the method of comparing the vocabulary in parallel texts" in J. Fisiak, ed., *Linguistic Change Under Contact Conditions*, Mouton: The Hague, 1965, pp. 149–160.

[117] Witold Manczak, "The method of comparing the vocabulary in parallel texts", *Journal of Quantitative Linguistics* 10(2), August 2003, pp. 93–103.

[118] Johanna Nichols *Linguistic Diversity in Space and Time*, The University of Chicago Press: Chicago, 1992.

[119] Sabine Ploux & Hyungsuk Ji, "A model for matching semantic maps between languages (French/English, English/French)", *Computational Linguistics*, 29(2) pp. 155–178.

[120] Robert H. Robins, "Polysemy and the lexicographer" in R. Burchfield, ed., *Studies in Lexicography*, Oxford University Press: Oxford, U.K., 1987, pp. 52–75.

[121] Vahan Sarkisian, "Edward Spencer Dodgson y la problemática vasco-armenia", *Fontes Linguae Vasconum*, No. 86, 2001, pp. 13–32.

[122] Morris Swadesh, "Lexico-statistic dating of prehistoric ethnic contacts", *Proceedings of the American Philosophical Society*, Vol. 96, 1952, pp. 452–463.

[123] Eve E. Sweetser, *From Etymology to Pragmatics*, Cambridge University Press: Cambridge, UK, 1990.

[124] Carlo Tagliavini, *Le origini delle lingue neolatine. Introduzione alla filologia romanza*, Casa Editrice Prof. Riccardo Patron: Bolonia, 1969.

[125] Robert L. Trask, *The History of Basque*, Routledge: London, 1997.

[126] Peter Whittle, *Probability*, John Wiley & Sons: London, 1976.