# Distributions of cognates in Europe based on the Levenshtein Distance

Job Schepens

Department of Artificial Intelligence

Radboud University Nijmegen

Job Schepens, 0436321

jobschepens@student.ru.nl

Bachelor Thesis

Supervisor: Prof. Dr. A.F.J. Dijkstra

Supervisor: Dr. F.A. Grootjen

Date: Summer 2008

# Contents

## Abstract

We applied the Levenshtein distance on a professional translation database (extracted from Euroglot professional 5.0) in order to identify distributions of cognates in 6 European languages. Using the Rosetta schemes of Grootjen (2008) for database interaction, we classified translation pairs as cognates if a score for orthographic overlap based on the Levenshtein distance was above a motivated threshold. Semantic overlap was determined using the conceptual structure of the database. Differences between cognate distributions across languages were found to be similar to validation studies on language similarity ordering. In addition, numbers of translations, proportions of form-identical to form-similar cognates, and proportions of form-identical false friends to form-identical cognates were compared between languages. We show that these new techniques from artificial intelligence can facilitate the selection of stimulus materials for psycholinguistic cognate and false friend research, and can assess language similarity ordering between the analyzed languages: English, German, French, Spanish, Italian, and Dutch.

## Introduction

Although Sumerian is the oldest written language known (the Kish Tablet is dated 3500 BC), we still use words from this language. For instance, the proper noun *Iraq* is believed to originate from the Sumerian name *Uruk* (a region in Iraq), implying that the form and meaning of this word are maintained in many modern languages. Another example of a word with a long history would be the noun *sugar*, which is believed to originate from the Sanskrit word *sharkara*.

Words like *sugar,* which have many form-similar appearances across languages, are known as cognates in linguistic and psycholinguistic research. Cognates can be defined as translation pairs with a high orthographic overlap. Cognates can be form-similar or form-identical. For instance, the Dutch – English translation pair *sigaret – cigarette* is an example of a form-similar cognate, and *president – president* is an example of a form-identical cognate. Cognates must also have a very similar meaning across languages, but the meaning overlap does not have to be perfect. More specifically, not all of the readings of a word in a source lexicon have to be the same as the readings of a translation of that word in the destination lexicon. For instance, the Dutch – English translation pair *bank – bank*, shares the meanings of *sandbank* and *financial institution,* but the English *bank* also means *waterfront,* whereas the Dutch word does not have this meaning. The dimension with respect to the semantic similarity of cognates is subject of much psycholinguistic research.

In the present study, we are interested in the orthographic and semantic dimensions of words in order to recognize cognates from a linguistic database. In the psycholinguistic literature, cognate research often goes together with research on false friends. False friends form a category of translation-pairs like cognates, but they only share form-overlap across languages and not semantic overlap. False friends are often translated erroneously, because their translation is expected to be the word with the same form in the other language. For instance, the Dutch – English false friends *integer – integer,* are orthographically identical whereas their meanings do not overlap. The Dutch word means *honourable*, and the English word means *whole* or *numeral*. Together with cognates, false friends make up the category of interlingual homographs, words with identical (or similar) orthography. Cognates in modern language can originate from more primitive languages. For example, the words from different languages that denote important concepts like *sun* or *moon* are often cognates in modern languages from the same language families. Another reason for the presence of cognates is that words may be borrowed from other languages. Words like those are often called loanwords. Examples are Dutch words like *computer* (from English) and *cadeau* (from French, meaning 'present'). Within languages with different spelling systems, the cognate's form appearance may change, resulting in form-similar cognates instead of identical cognates.

### Goals of the present study

While these typical examples of word origins interest linguists, psycholinguists use words like these to study language processing in the mind. Linguists can use distributions of cognates with respect to orthographic similarity to determine how and to which extent

languages have changed over time. It may also be of interest to linguists, to assess the cross-linguistic similarity across languages in this way. The present study aims at interest from both fields by discussing new tools from artificial intelligence to relate words from different languages and to produce useful stimulus materials for cross-linguistic and bilingual studies. These tools are based on computer schemes that enable relating words from different languages to each other. The applied schemes are named Rosetta, after the famous Rosetta Stone that offered a way to relate different ancient writing systems and languages to each other. The Rosetta schemes provide a programmatic interface to the Euroglot data.

More specifically, we wish to identify distributions of cognates across different languages by applying the so-called Levenshtein distance to assess the orthographic similarity of cognates and by applying automatic translation to determine their semantic similarity. Furthermore, the process of identifying distributions of cognates, will also enable us to extract such words from the database themselves. Lists of cognates and false friends are useful to select more advanced stimulus materials for psycholinguistic studies. In addition, the collected distributions of cognates, will also allow researchers to control their stimulus materials with respect to orthographic similarity. Furthermore, we will consider the number of translations of words between languages, which gives researchers the possibility to account for polysemy in future stimulus lists.

In the remainder of this introduction, the effects of cognates on language understanding and production and resulting theories of language representation are discussed.

**Psycholinguistic cognate research**

There is an extensive literature on cognate effects in bilingual language processing. There has been a host of bilingual reading studies showing a facilitatory effect of cognate processing relative to words that exist in only one language. The empirical findings have led to different psycholinguistic theories of cognate representation. In this section I will give a short overview of important findings and proposed theories.

Friel and Kennison (2001) provide an overview in their paper of the effects of cognates in various experimental tasks. It turns out to be easier to acquire cognate translations relative to non-cognates when participants have to learn words in a new language (De Groot & Keijzer, 2000). For instance, when participants have to generate an association in two languages, cognates were easier to generate as associates, and associates were more often cognates than non-cognates (Van Hell & De Groot, 1998a). Cognates are also more easily categorized (Dufour & Kroll, 1995). In lexical decision tasks (where the participant must decide if character strings are words or non-words), cognates are usually responded to faster than non-cognates (Caramazza & Brones, 1979). This effect has been shown for cognates presented in a second language as well as for cognates in a first language (Van Hell & Dijkstra 2002). Priming effects have also been found for cognates, whereas non-cognate priming effects are non-existent (Kirsner, Smith Lockhart, King, & Jain, 1984).

Many proposals with respect to the organization of linguistic knowledge organization in bilingual memory and the lexical access during language processing are based on these findings. De Groot and Nas (1991) propose that cognates share a common conceptual representation and non-cognates have their own conceptual representation for each language,

because in their cross-language semantic priming experiment priming effects were only significant for cognates. Another theoretical view holds that cognates do not only share conceptual representations, but also share lexical representations (Sánchez-Casas et al., 1992). Furthermore, it is proposed that every word is represented in a cluster for its common root morpheme. This way, not only are all words with common morphology from one language stored together, but also the possible cognates from other languages (Kirsner, Lalor, & Hird, 1993). Other views on cognate representation are still localist connectionist or distributed connectionist in nature. All in all there is not yet one common theory of cognate representation in the brain.

**Involved Issues**

Evidence about cognate representation has come, to a large extent, from lexical decision tasks involving cognates, false friends, and translation-pairs, that were especially rated before the experiment by bilinguals from the population later tested. The ratings are important to match the semantic similarity and form similarity between test words (e.g., cognates) and control words (usually words that exist in only one language). In addition, the distribution of form similarity of the stimulus words should correspond to that between translation-pairs in the languages themselves. Such a distribution is dependent on the language combination used in the task, which presupposes an analysis of languages in order to control the distribution of cognates, false friends and number of translations. One reason for the present study was to test out new methods for finding such distributions and for comparing them between languages.

The development of methods to obtain cognate, false friend, or translation equivalent distributions across languages requires the consideration of several important issues. The main selection procedure should automatically score every translation-pair on a similarity metric. Thus, a valid metric is needed to norm translation pairs on orthographic similarity. Another issue is to extract all possible translation-pairs from a translation database. The numbers of cognates, false friends, and translations should be counted for every language combination. The Rosetta schemes that we apply will allow access to the basic types of information contained in this database, which enables the processing of every translation-pair in the language combination. This makes it possible to count and analyze every translation-pair one by one. It should be tested if automatic translation is a valid method to approach each theoretical problem addressed. The basic types of information from the database used for automatic translation are: expressions, readings, concepts, and relations to concepts. A description of the database that we made use of and of the basic types of information in the database is given in the next section. Next, we discuss a series of theoretically interesting issues in six sections.

## Database description

For the purposes of automatic translation and analyzing complete lexicons we used the professional translation database Euroglot. Euroglot is a translation database produced by Linguistic Systems B.V., Nijmegen, Netherlands. It has successfully been used for professional translation purposes (they provide a list of references on their website). The database is based on a conceptual translation mechanism, which is used to translate expressions via their relation to language independent concepts. In our study, we have been using an extract of Euroglot Professional 5.0. This database is available for the languages Dutch, English, French, German, Spanish and Italian, so we analyzed each combination of these languages. The average number of expressions in the extracts was around 72000 per language, the standard deviation was around 7000 expressions. The different sizes (and the number of translations as well) varied across language combinations, as seen in Table 1. Note that the database files we used for this study were data extractions from Euroglot itself, so these numbers do not apply to the original database.

| Language | Size |
|---|---|
| Dutch | 76000 |
| English | 74000 |
| French | 63000 |
| German | 81000 |
| Italian | 65000 |
| Spanish | 62000 |

**Table 1: Languages with the exact numbers of expressions in the database extractions.**

There is a specific xml-file for each language in the database, where each file has the same structure with different information in it. The structure consist of the fields (for our interest) expression, reading, concept, and relation. There is other information stored in Euroglot we did not take into account, such as syntactic category. For the present study, we decided to use every translation pair and we did not control for syntactic category. As a consequence, also proper nouns such as country names were analyzed. With the four basic types of information available, the automatic translation procedure can process every translation pair one by one and analyze them directly with respect to orthographic similarity.

The database structure can be visualized as in Figure 1. Each language file contains of a set of expressions and a set of concepts. Each expression has a set of unique reading numbers, each one being a specific meaning of the corresponding concept of that reading. Each concept from the set of concepts in a language file has a set of unique reading numbers, each referring to a specific reading of an expression in the set of expressions. This structure is the key for translation purposes. An expression together with its set of readings and accompanying concept numbers and relation numbers make a ‘word’ in the database structure. A word thus contains every relevant field of information for an expression.

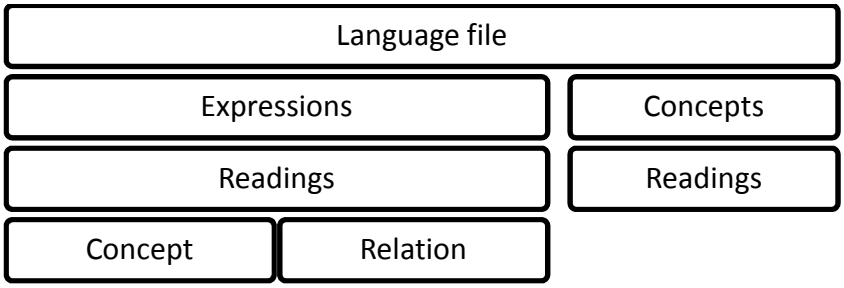


**Figure 1: Structure of the database**

The basic fields in the database are described by explaining the automatic translation procedure. An ambiguous word like *bank* has multiple readings, such as *financial institution* and *waterfront*. Every reading is connected to one concept via a relation number. A concept number is language independent, while a reading-number is unique for a specific reading of an expression. A relation-number specifies the meaning of a reading to a concept. This way, the readings to a concept specify different expressions associated with that concept, and the different relation numbers among all readings specify different meanings the concept has. Readings can have multiple relations to a concept, when a reading governs multiple meanings of that concept. For example, when retrieving expressions associated with the concept of *financial institution* in English, one would get the expression *bank* amongst others. The reading to *financial institution* of *bank* has specific relation numbers to the concept which mean (using expressions) *cash dispenser* and *agricultural loan bank* amongst others, both belonging to the concept of *financial institution,* but represented by different relation numbers.

Semantic overlap between expressions is represented by having readings to the same concept that share the same relation number to that concept in the database. When determining the translations of an expression in a different language, the relation numbers from each relevant concept are retrieved in the other language, and compared to the relation-numbers of the reading to that concept in the first language. When a match is found, the expressions of the readings with a matched relation make a translation pair. This way, a specific translation pair may be found multiple times if both expressions have multiple shared concepts and if shared relations exist for more than one of these shared concepts (multiple shared relations within a shared concept were not counted multiple times). For example, *bank – bank* is counted once for the shared meanings *financial institution* and *waterfront,* and not for *cash dispenser* and *agricultural loan bank,* which belong to the concept of *financial institution.* The issues with this method and a validation thereof, are discussed on page 14 under Study 2 – Semantic Similarity.

For the basic types of information to become available for automatic translation, a specific structuring procedure was executed before language combinations were be analyzed. The two resulting objects of this procedure were smart representations of the lexicons in the form of hash tables. These are data structures used to access specific information from large collections faster. The first object was a collection of all the *expressions* in a language mapped to their corresponding *words*. The second object was a collection of all the *concepts* in a language mapped to lists of *words* which semantically relate to every *concept*. For the implementation in Java of this procedure and others see page 35 under Implementation.

## Studies

This section is separated in six studies that each concern a specific issue having to do with the representation of special words like cognates, translation pairs, or false friends.

The first study to be reported was concerned with determining a form similarity metric for word pairs of different languages. In this study, we examined the usefulness of applying the Levenshtein distance as a psycholinguistic metric of orthographic distance. This study was also concerned with the threshold used to distinguish between cognates and words with too few common characters.

The second study was about semantics in the database. We questioned the use of the semantic structure of the database as a valid mean to determine semantic overlap. The validation was based on a comparison with translation pairs identified by Tokowicz et al. (2002).

Our third study was concerned with the question if the observed cross-linguistic similarity distributions of word pairs in different languages would be in line with other measures of language distance. For this purpose, we constructed a language similarity ordering based on the numbers of cognates in each language combination and compared this ordering to measures by Gray and Atkinson (2003) and to intuitions of language users.

In the fourth study to be discussed we compared the number of translations between language combinations, and subsequently related these numbers to a potential collector's bias in the linguistic database.

In the fifth study, we compared proportions of identical cognates with false friends for different language combinations. A new language similarity ordering, dependant on false friends was compared to the other measures of language distance.

The final study to be reported was about the differences between proportions of form-identical cognates to form-similar cognates across language combinations. Here we studied the dependence of a language similarity ordering on the inclusion/exclusion of form-similar cognates.

### Study 1 – Orthographic Similarity of Translation Pairs

**Goal:** *To classify translation pairs with respect to their cross-linguistic orthographic similarity, assuming a minimal degree of orthographic overlap.*

To be able to classify the translation pairs that the database delivers into cognates and non-cognates, a valid metric for form similarity is needed (note that the translation pairs will already have a certain semantic overlap). The orthographic metric should be able to distinguish expressions with high orthographic overlap (form-similar homographs) from expressions with low orthographic overlap, independent of word length. For instance, the cognate-pairs *relative-relatief* and *idea-idee* should intuitively obtain a similar score, because both pairs share 25% of their characters. The counterintuitive counterargument would be that the second pair shares 100% less different characters than the first. The orthographic metric

should be formalized so that it can be applied in an algorithm. In any case, the measures should correlate with intuitions from bilingual language users.

Cognates used for experiments in the psycholinguistic literature are often rated by the experimenter himself or via similarity rating studies. However, these methods cannot be formalized and are biased towards concrete expressions (Friel & Kennison, 2001). Furthermore, these methods are time-consuming, so they are not applicable for the complete lexicons used for our studies. Tokowicz et al. (2002) also used rating tasks to measure the form similarity between translation pairs. They suggest the use of continuous norms, because of the continuous nature of form-similarity ratings in their experiments.

**Methods.** In information theory, there are two popular metrics for evaluating strings on form similarity, the Hamming distance and the Levenshtein distance. The Hamming distance counts the minimal number of substitutions needed to edit one string into the other. The Levenshtein distance does also take into account insertions and deletions.
Thus, the Levenshtein distance will produce distances smaller or equal to the Hamming distance. We point out here that cognates like *flutist-fluitist* takes advantage from this property. When only counting substitutions, *flutist* would be transformed in *fluitist* by substituting every character after the first three characters: the fourth character *t* becomes an *i*, the fifth character *i* becomes a *t,* etcetera, resulting in a distance of 5. When minimizing between insertions, deletions, and substitutions needed to transform the one string into the other, the resulting distance would be only 1 (one insertion). It is not trivial that the Levenshtein distance can be used as a good approximation to results obtained in rating studies, as is discussed next. With the Levenshtein distance, semi-continuous norms are applied to measure form similarity, in agreement with the research of Tokowicz et al. (2002). Some other recent studies have also made use of the Levenshtein distance, for instance, Heeringa (2004) used the Levenshtein distance to compare dialects.

Our implementation of the Levenshtein distance runs in $O(mn)$ time where $m$ and $n$ are the lengths of the source string and the destination string. However, it should be possible to run it in $O(m)$ time. The procedure is divided in three steps. First, the values in first row and the first column of a $m$ by $n$ matrix $A$ are initialized with the corresponding column and row numbers. Second, the rest of the values are computed in an iterative way, until every entry has a value. A value $A_{i,j}$ is determined by taking the minimal value of $A_{i-1,j} + 1$, $A_{i,j-1} + 1$, $A_{i-1,j-1} + cost$. These three values are deletion, insertion, and substitution, respectively, where cost is 0 when character $i$ is equal to $j$ and 1 otherwise. Third, the value in the $A_{mn}$entry is returned, because this is the minimal number of edits needed to transform the source string into the destination string.

The resulting distance is still sensitive to the word lengths of the given strings. Because we want the metric to be independent, we adopted a formula that normalizes the Levenshtein distance and corrects it for word length. The formula is given by Equation 1.

$$score = \frac{length - distance}{length}$$

$$length = \max(length\ of\ source\ expression, length\ of\ destination\ expression)$$
$$distance = \min(number\ of\ insertions, deletions\ and\ substitions)$$

**Equation 1: normalized score with a correction for word length**

This formula corrects and normalizes the Levenshtein distance using the maximum of the lengths of both expressions. We chose the maximum and not the mean or the minimum, because this choice also normalizes the score and the other options do not. The bounds for every variable in the formula are known, so we can determine the number of possible values of the formula. The maximum of both lengths is an integer between 1 and 8, because only words with length smaller than 8 are evaluated. The Levenshtein distance thus is an integer between 0 and 8. Note that the Levenshtein distance can never be longer than the length of the longer expression. With these constraints, it is observed that the score can take on 23 different values between 0 and 1. Note that maximum 13 of the values can be the result of different combinations of length and distance, because there are 36 unique combinations of length and distance. Furthermore, the score is more sensitive (i.e., differentiated) for longer words. While word-pairs with a maximum of 8 characters can have 9 different values, word-pairs with a maximum of 2 characters can have 3 different values. An important property of the formula is that it scores relative to maximum word length. For instance, expressions sharing 3 out of 4 characters get .75 and expressions sharing 3 out of 8 characters get .375, while expressions sharing 6 out of 8 characters do get .75. A plot of the function for its possible values can be seen in Figure 2, where the floating edge of the brown surface visualizes that short words are rated relatively high.

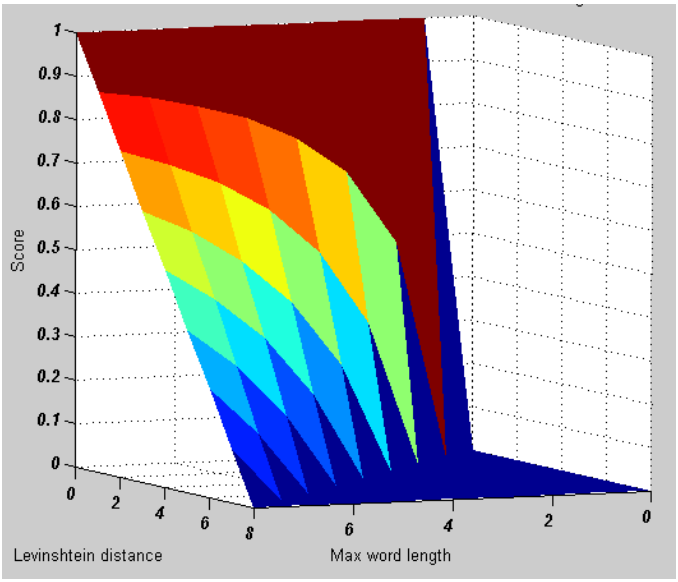


**Figure 2: Score as a function of Levenshtein distance and max word length.**

Two other issues for determining orthographic similarity are discussed now. The processing of diacritic marks and the uppercase of German nouns were issues to decide upon. When processing a translation pair, the diacritic marks would be maintained, while the letter

cases are adjusted. We assume that a character with a diacritic mark is a unique character itself, although subjects tend to see them as more similar. Before scoring a translation pair, the pair is set to lower case, mainly to correct for German nouns. However, this operation is done for each language combination, to take into account the orthographic similarity between an uppercase character and the corresponding lower case character.

**Results.** Example translation pairs obtained by applying this metric with a threshold set on 0.5 after using the semantic structure from the translation database, are included in Appendix 1. Using this value for the threshold, we obtained the best results in our validation studies (discussed in the next section). Because we saved the lexicons from the database in hash tables, which are not alphabetically ordered, the first 1000 cognates saved for each language combination are not the first cognates encountered when one searches a dictionary from start. Saved cognates were just alphabetically ordered afterwards. In this way, the example cognates in the list are only a small part of the obtained cognates. As can be seen in cognates like *vagebond – vagabond* and w*alhalla – Valhalla*, the number of cognates obtained is very high and the obtained list contains words that would not be found if searched for manually. Furthermore, by applying this formal approach, the resulting words could directly be classified as cognates, which is not possible if done manually. Automatic scores for the two example cognates vagabond and Walhalla are both 0.875, because the cost for substitution is 1 (thus the Levenshtein distance is 1) and the maximum word length for both words is 8.

**Validation.** The validation of the metric was based on two studies: Dijkstra, Brummelhuis, and Baayen (submitted) and Tokowicz et al. (2002). First, we performed a quick adoption of 116 form-similar translation pairs that were rated by test subjects in a cognate study by Dijkstra et al. (2004). With a threshold set on 0.5, every cognate that was rated 5/7 or higher on orthographic similarity (47 cognates) was recognized, with the exception of *hope – hoop* and *circle – cirkel*. Experimental subjects rated these two words higher than the automatic metric did. A possible explanation for *circle – cirkel* could be that the *c* is often pronounced like a *k* (a spelling characteristic). For *hope – hoop,* the weight of the insertion of the extra *o* is possibly considered relatively low, because there is an *o* already there. Again this is due to a spelling characteristic of Dutch and English. Other considerations could be a difference between weight of character change between the end or inside of words (Font, 2001), a difference in weight between change in vowels or consonants, or an interchange between characters.

In addition, we adopted the 794 translation-pairs with length between 3 and 8 characters used by Tokowicz et al. (2002) for a similar test. These translation pairs are in large part the same as those in studies by De Groot (1992) and Dijkstra et al. (1999). These translation pairs also included additional translations for each word in the list, determined by Tokowicz et al. were participants had to name their first spontaneous translation. For the translation pairs that were rated 5/7 or higher, 150 out of 193 translation pairs were classified as cognates (77,7%). The cognates which were not recognized can be found in Appendix 2. With a threshold set on 0.6, 129 out of 193 translation pairs were classified as cognates. This was considered to be a loss of too many cognates, so we adopted the final threshold value of 0.5.

A larger overlap of cognates selected in empirical studies and in automatic studies could be obtained by using AI techniques from information retrieval, such as inclusion-exclusion. However, this point is only of relative importance to the validation of the metric itself. The mentioned techniques could be used to determine the best fitting correction for word length in Equation 1, where we have now made an intuitive choice for this.

**Conclusion.** We have identified many cognates by means of a formal metric for orthographic similarity, assuming semantic similarity of translation pairs in a translation database. Although some cognates from the empirical studies used for validation were not identified, the numbers of cognates found in this study were much larger than those made available in the validation studies. In all, this finding suggests that researchers could make confident use of the type of automatized cognate selection procedures we have described.

**Study 2 – Semantic Similarity**

**Goal:** *To classify translation pairs (both cognates and noncognates) with respect to their cross-linguistic semantic structure in the database, i.e., the specific shared semantic relations or features of the translation equivalents.*

Psycholinguists would like to classify every translation of each word in different language combinations in order to select stimulus materials for cross-linguistic or bilingual experiments. It is not immediately obvious how to decide which words should be classified as translations. The structure of the translation database provides information with respect to related words, but the meaning of words seldom is totally the same between languages. For instance, the cognate *bank* between English and Dutch shares only shares some of its multiple meanings, so that the reading of bank as in *waterfront* is not shared with the Dutch word *bank* at all. However, intuitively, *bank* should certainly (also) be considered a cognate.

For our purposes, we would like to have a formalized translation method that automatically returns not too many and not too few translations for each word in the lexicon. Studying the psycholinguistic literature, we see that this method is totally different from traditional methods. In the field, semantic similarity is traditionally determined in various rating tasks. Friel and Kennison (2001) compared the easier semantic similarity rating task with the more consuming randomized translation elicitation task. Both tasks require test subjects to distinguish word-pairs between cognates and false friends. In the translation elicitation task, monolinguals had to name the translations of foreign words in randomized order. If correct translations are observed, the word-pair would be considered semantically similar, otherwise not. Because the automatic translation method retrieves translation pairs that are already identified by experts (who put them in the database in the first place), this method would be perfect to use for stimulus materials in such tasks. The automatic translation method described next is certainly not replaceable by the opinions of test subjects, but the resulting cognates and false friends could be used to guide the selection of stimulus materials, and would help to identify more cognates than otherwise possible.

**Methods.** An automatic translation algorithm was developed that is able to iterate a search through languages, lexicons, expressions in these lexicons, readings of these expressions, and relations to concepts of these readings. An advantage of this automatic translation procedure is its symmetric property, so that each language combination needs to be processed only once, instead of an iteration in two directions. Of course, the observed relations between words greatly depend on the semantic structure of the database and are limited by its size. Therefore, it is very important that the database is consistent and very secure. In practice, there will always be some noise in the observed numbers of relations and they will probably be underestimations, because they are derived from a database that must necessarily be an incomplete reflection of real, every-day language use.

We decided to classify translation pairs like *bank – bank* as cognates for each different shared reading. A shared reading would be determined by comparing the relations to the concepts of each reading of both words, also securing that these relations point to the same concept. The relation numbers represent the relation of the specific reading to the concept,

which is in itself represented by its relations to readings. *Bank – bank* shares relation numbers to the concept of *financial institution,* so when translating the English *bank* (as in *financial institution*) to Dutch, one would get *bank* as a translation, instead of *cash dispenser,* since *cash dispenser* shares relation numbers with *geldautomaut.* There is a difference between multiple shared readings and multiple shared relations. Each shared reading is specific for a concept, so multiple shared readings between words govern multiple concepts. It is also possible that a source word with specific reading has multiple relations to some concept and another destination word with specific reading also has these relations to that concept, thus sharing multiple relations. An example of a translation pair sharing multiple relations would be the English-French translation pair *glutton-glouton,* which shares no more than six relations. Word pairs sharing meanings by relations were classified as only one translation pair, whereas word pairs sharing multiple readings were classified as multiple translation pairs. Conforming this idea, *glutton-glouton* is only stored once, whereas *bank-bank* is stored over four times (in the sense of *sandbank, cash dispenser, branch bank,* and *banking*). Other meanings of the English word *bank* are not shared with the Dutch expression *bank (slope, capsize, shore* and *border).* Otherwise, the English *bank* does not (exactly) hold the Dutch meaning *couch* of *bank.*

With respect to the classification of translations (by matching relation numbers and concepts), this method could be quite specific. Retrieved translations are exactly matched and other relations to the specific concept are omitted. For instance, in the translation database the words *Mambo* and *Samba* have different relation numbers to the same concept ("*Latin dances"* so to say). Although the forms are similar, they will not be classified as cognates, because their relation numbers do not match. The opposite holds for false friends. False friends of a word are retrieved from the set of homographs by removing all translations. Thus, *mambo* and *samba* are classified as relatively dissimilar false friends. In fact, their orthographic score would be 0.6, so this pair could be counted as a form-similar false friend. However, to keep things simple further on, we only counted identical false friends in the comparisons of word type quantities.

**Results.** The proposed method was used to extract every translation pair from the database. The same example translation pairs as for orthographic similarity will illustrate the method for classifying translation pairs (Appendix 1). As can be seen in the list, the Dutch word *vopo* (derived from Volkspolizei) is classified for each reading separately, so that *vopo – vopo* (as in policemen) and *vopo – vopo* (as in police) are two separate cognates. According to the database, the Dutch do not make a difference between *vopo* and *VOPO*, while the English do. Such a case results in two more cognates, because the meaning of the upper case *vopo* is the same as the lower case *vopo*. Using this formal method for classifying translation pairs automatically using a translation database, it is possible to identify many cognates that are not easily identified when using traditional methods. Resulting lists of translation pairs with scores on orthographic similarity can be found by contacting the author.

**Validation.** To validate the automatic translation method, we compared the identified cognates with the items in Tokowicz et al (2002). The cognates produced by means of the database should be a superset of these, in particular their items with a high orthographic and

semantic similarity rating. From the 1004 translation pairs on their website with semantic and orthographic similarity ratings, 794 translation pairs have word lengths between 3 and 8 characters. Because the norms of both ratings are continuous, the similarity criterion for what is a cognate is not clearly present. Using our own criterion, we found that 768 of these translation pairs has a similarity rating of 5/7 or higher. For this validation study, we checked every translation pair from the database on its presence in Tokowicz' list. If a pair was present in our database, it was excluded from Tokowicz' list. The remaining list of items consisted of 136 word pairs. Of these, 106 pairs still had a semantic similarity rating above 5/7. So, 86.2% of the translation pairs rated 5/7 or higher were classified using the semantic structure from the database.

The unclassified translation pairs are included in Appendix 3. The 11 most semantically similar word pairs of this list have been sent back into the translation database in order to find an explanation for why these translation pairs, according to test subjects, should not have been considered as translation pairs according to the experts who constructed the translation database. The explanation for each word can be found in Table 2.

| Dutch | English | Rating | Explanation |
|---|---|---|---|
| dorpje | village | 7.00 | *dorpje* is a diminutive |
| geloof | religion | 7.00 | *geloof – faith* and *religion – religie* |
| lammetje | lamb | 7.00 | *lammetje* is a diminutive |
| mist | mist | 7.00 | *mist – fog* and *mist – nevel* |
| steegje | alley | 7.00 | *steegje* is a diminutive |
| verraad | betrayal | 7.00 | *verraad – treason*, *betrayal* is not in the database |
| pop | puppet | 6.88 | *pop – doll* and *puppet – poppenkastpop/marionette* |
| vrouw | female | 6.88 | *vrouw – wife/Mrs/queen* and *female – vrouwtje/vrouwelijk* |
| ede | oath | 6.75 | *onder ede – on oath* and *oath – vloek/eed* |
| gemeen | cruel | 6.75 | *gemeen – mob/rabble/common/mean/biting* and *cruel – wreed* |
| graaf | duke | 6.75 | *graaf – count/earl* and *duke – hertog* |
| huurder | renter | 6.75 | *huurder – tenant, renter* is not in the database |
| kijken | watch | 6.75 | *kijken – look/see* and *watch – gadeslaan/waken* |
| snoer | wire | 6.75 | *snoer – line/cord* and *wire – kabel* |
| spoor | rail | 6.75 | *spoor – track/rails* and *rail – rail/spoorrail* |
| bandiet | crook | 6.62 | *bandiet – bandit* and *crook – gannef/dief* |
| jammer | pity | 6.62 | *jammer – a pity* and *pity – medelijden* |
| pokken | pox | 6.62 | *pokken – smallpox/variola, pox* is not in the database |
| voorkeur | favour | 6.62 | *voorkeur – preference* and *be in favour of – voelen voor* |
| waard | worth | 6.62 | *het waard zijn – be worth* and *worth - waarde* |

**Table 2: Semantic structure for unclassified translation pairs in Euroglot**

Explanations for the differences seem to be a result of the way experts think about constructing the semantic structure of the database, while test subjects are generally not that precise in their ratings or language use. A translation pair like *mist – mist* is absent, because (according to experts), it does not share the exact same relation(s) to the shared concept. It is important to note that the database is quite detailed, and that we considered only primary translations. Words that have relatively many different meanings, also have more refined relations to their concepts and possibly different words to 'capture' them in another language.

So, for these words to be classified correctly, secondary translations also have to be considered. This explains why they were identified as translation pairs in Tokowicz' study.

**Conclusion.** Translation pairs were classified according to the semantic structure of a translation database. To a large extent (86,5%), this database was found to reflect the semantic structure that is also present in the semantic similarity rating study by Tokowicz et al. Therefore, this automatic translation method can be used with confidence to classify translation pairs as a means of identifying cognate distributions across language combinations.

**Study 3 – Cross-Linguistic Similarity**

**Goal:** *To determine a language similarity ordering with respect to distributions of cognates across language combinations. This language similarity ordering should be supported by language evolution studies and the intuitions of language users.*

With the proposed methods for determining orthographic and semantic cross-language similarity, we identified distributions of cognates across language pairs. In this study, we determined if these distributions can be used as measures for cross-linguistic similarity and linguistic diversity. The methods we applied and the resulting distributions will be discussed first. Next, we will compare the similarity results to language evolution studies and to common intuitions on language similarity. It may be hypothesized that an ordering of the observed cognate quantities over language pairs reflects the language distance between the languages involved, because cognates often have shared language origins. In other words, if one language pair shares 10,000 cognates and another language pair shares only 7500 cognates, the second pair may be considered to be less similar than the first. Also, if the second pair shares more translation pairs but fewer cognates, cross-linguistic similarity could be decreased furthermore relatively to the first language pair, which shares many cognates in less translation pairs. One can base a language similarity ordering on several item characteristics, such as cross-language form similarity, numbers of identical cognates, relative proportions of cognates and false friends, etcetera. In this section we will consider language similarity and cognate distribution. In the next three sections, we will discuss the dependence of a language similarity ordering on the number of translations, the number of false friends, and the proportions of form-similar to form-identical orthographic similar items for cognates.

**Methods.** Cross-linguistic similarity can be assessed using different methods, such as determining distributions of cognates (this study), considering the evolution of languages (Gray and Atkinson, 2003), comparing the grammar of languages, comparing meaning overlap between concepts for different languages, and collecting intuitions on language similarity. A language similarity ordering based on Gray and Atkinson is used in the present study to validate the ordering obtained by cognate distributions. In addition, a little language questionnaire was sent out to Dutch-English bilingual students to obtain intuitions about a language similarity ordering.

To identify the distributions of cognates, we translated each word from each source lexicon to each destination lexicon (15 language combinations). A score on orthographic similarity was calculated for each of these translation pairs (total quantities can be seen in Table 4). Every score was saved in a table along with the length of the words of the translation pair, in order to observe separate scores for all minimum word-lengths. We chose minimum word length, and not maximum, mean, source or destination length because this measure gave the best distribution across word lengths. Source or destination length is not specific to a combination of two languages, and maximum and mean both had distributions that were shifted too much towards larger word lengths. The table was used to further determine if the score for orthographic similarity was not biased in preferring translation pairs of a specific word length (see Table 3).

To visualize the cognate distributions in a continuous way, a moving window representation was used (see Figure 3). The best trade-off between smoothness and keeping the data intact was found for a moving window of size 0.05. For every value in the graph, a new value was computed by taking the mean over values that were less distant than 0.05 points. This was not done for scores of 1.0, because numbers of identical cognates are more in demand. The numbers of identical cognates in the graph are therefore the same as in Table 7. The graph uses a logarithmic y-axis with number of cognates to account for the increase of identical cognates in the far right of the graph.

To visualize the cognate distributions in a way that differences between language families can be observed, we inverted the axis of Figure 3, resulting in Figure 4. This time the observed numbers of cognates were stored in bins instead of represented in a moving window. The figure consists of 8 bins, but 4 are not visible since only scores from 0.5 and higher are visualized. The range of a bin is determined by dividing 1 by the number of intended bins. The number of cognates in a bin is determined by summing all numbers of cognates with a score that falls in the range of the bin. Also in this figure, the numbers of identical cognates were retained for clarity.

**Results.** The distribution of cognates across word lengths in Table 3 is comparable across languages. Of course, the numbers of translations and cognates differ, but generally, the more translations there are for a certain word length, the more cognates are found.
The numbers of cognates (Table 4) show that Dutch-German is the most similar language combination of all language combinations. This is also seen in Figure 3, were the red line lies clearly above all others. Another closely related language (Italian-Spanish) is the second most similar language combination. Although the difference in total number of cognates (1423 cognates) is quite distinctive, it is observed in Figure 3 that for some scores there are yet more form-similar cognates in Italian-Spanish compared to Dutch-German.

From the resulting ordering of cognate numbers, it can be observed that closely related languages share more cognates than faraway related languages, with the exception of English-French, -Spanish and -Italian. In Figure 4, these languages also appear further to the right of the graph, running through the closely related languages. Note that only English-French has a number of identical cognates comparable to closely related languages, whereas English-Spanish and English-Italian have a number of identical cognates comparable to other faraway related languages, as seen in Figure 4. A resulting language similarity ordering could be like the order in Table 3, that was determined by sorting the cognate quantities.

| | du-en | | du-fe | | du-ge | | du-it | | du-sp | |
|---|---|---|---|---|---|---|---|---|---|---|
| length | total | cognates | total | cognates | total | cognates | total | cognates | total | cognates |
| 3 | 3504 | 374 | 2104 | 217 | 2053 | 464 | 1635 | 152 | 1801 | 141 |
| 4 | 9078 | 1182 | 4834 | 601 | 4910 | 1359 | 4159 | 505 | 5018 | 458 |
| 5 | 8839 | 1660 | 6110 | 1099 | 5888 | 2237 | 5118 | 883 | 6386 | 896 |
| 6 | 8660 | 2099 | 7348 | 1612 | 7712 | 3378 | 6054 | 1365 | 7417 | 1297 |
| 7 | 7020 | 2060 | 6088 | 1685 | 6721 | 3430 | 5506 | 1601 | 6044 | 1477 |
| 8 | 3056 | 1062 | 2760 | 932 | 3527 | 1903 | 2574 | 907 | 2782 | 848 |

**Table 3: Table with numbers of translations and cognates in Dutch language combinations for each possible minimal word length.**

| language combination | cognates | intuitions | evolution |
|---|---|---|---|
| dutch-german | 12908 | 1.95 | 20 |
| italian-spanish | 11485 | 1.95 | 26 |
| english-french | 9286 | 7.64 | 204 |
| french-spanish | 9120 | 3.32 | 34 |
| french-italian | 8871 | 4.00 | 26 |
| dutch-english | 8609 | 5.55 | 42 |
| english-spanish | 7837 | 9.41 | 204 |
| english-german | 7750 | 7.45 | 36 |
| english-italian | 7430 | 10.05 | 184 |
| dutch-french | 6269 | 8.68 | 200 |
| french-german | 5725 | 9.86 | 194 |
| dutch-italian | 5564 | 12.73 | 180 |
| dutch-spanish | 5298 | 11.77 | 200 |
| german-italian | 5187 | 11.45 | 174 |
| german-spanish | 4794 | 11.73 | 194 |

**Table 4: Language similarity orderings based on, respectively, cognate numbers, intuitions of Dutch-English language users, and language evolution.**

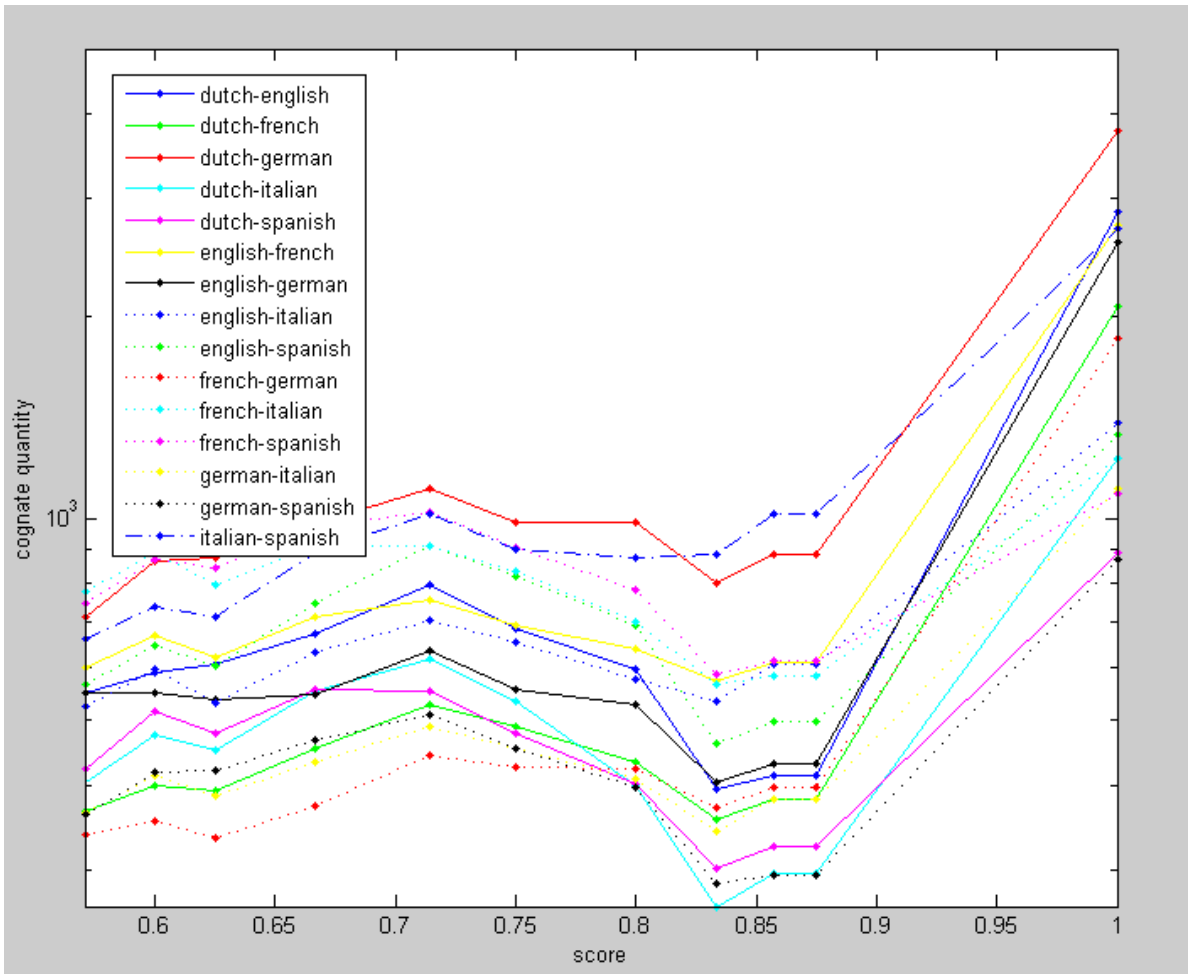


**Figure 3: Cognate distributions on a logarithmic y-axis and a moving window of 0.05. Numbers of cognates are plotted against each score between 0.5 and 1.0.**

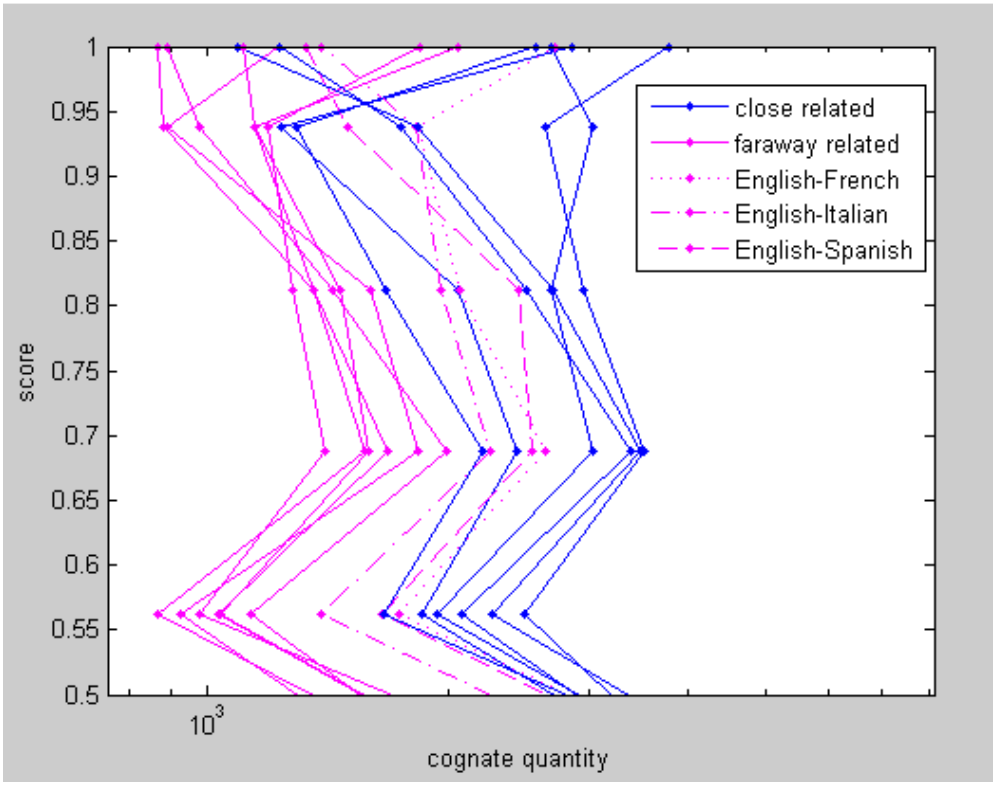


**Figure 4: Score as a function of numbers of cognates. Closely related languages are combinations of Romance or Germanic languages, faraway languages are combinations between Romance and Germanic languages.**

**Validation.** In order to assess whether the acquired language similarity ordering makes sense, a little language questionnaire was constructed to collect intuitions about language similarity. This questionnaire simply drew on the intuitions of respondents by asking to write down a number between 1 and 15 next to each language combination. The 22 respondents were mainly students at the Radboud University Nijmegen. Participants were asked to try using every number only once. The mean of the rating for each language combination was calculated and ordered with the other means. The resulting intuitive language similarity ordering (Table 4) is generally similar (correlation of 0.91) with the automatic language similarity ordering. However, an important difference is that intuitions imply that English-French is a rather dissimilar pair, while the corresponding number of cognates for this language combinations is relatively high in the automatic analysis. We would like to suggest that our Dutch-English bilinguals underestimate the degree to which French and Latin have affected the English language. As is well-known, there was the famous Norman-French victory at Hastings in 1066 by William the Conqueror, who made the defeated Harold the last English-speaking king for nearly 300 years.

We further examined a language similarity ordering adopted from Gray and Atkinson (2003). This language similarity ordering is based on a language tree constructed to predict divergence times in the evolution of language. The length of the branches of the language tree were proportional to maximum likelihood estimates of evolutionary change. Evolutionary change was estimated using a database with 2,449 cognates across 87 languages, with prior models of lexical evolution based on detailed constraints on language grouping. The ordering is seen in Table 3, where the quantities are measured branch lengths between languages in the language tree. Although the cognate quantities are largely consistent (correlation of 0.72) with an ordering with respect to language evolution, some differences are present. Again, as a historical explanation for the large similarity of English and French in our study, we suggest that many words from French were borrowed by English, and vice-versa. Because both validation studies do not count infrequently used words (which loanwords frequently are), a difference was to be expected. To compare the orderings more precisely, cognate quantities should be controlled for borrowings. Furthermore, a collector's bias is to be assessed, because also the number of translations for English-French was in top-position across language combinations. Among the distorting factors in our ordering, there could be a too strict automatic translation procedure and a too loose metric for estimating form-similarity.

**Conclusion.** Using quantities of cognates present in the translation database, we constructed a language similarity ordering that is generally consistent with intuitions on language similarity and an ordering based on language evolution studies. English-French was considered more dissimilar by both validation studies, English-Spanish and English-Italian in a less obstruct way. Differences between orderings may be explained by word frequency issues, historical events, and a collector's bias in the database. Our ordering is based on semi-complete lexicons (in contrast to the validation studies), and may therefore be of use for linguists interested in cross-linguistic similarity and diversity.

## Study 4 – Number of Translations

**Goal:** *To determine the dependence of a language similarity ordering on the degree of polysemy between languages by determining the number of translations between each language combination.*

The reader may already have noticed that the numbers of translations between language combinations differ considerably across language combinations. In this section, we discuss how this aspect relates to language similarity ordering. The actual numbers of translations can be used to determine the degree of polysemy between language combinations in the database. If using the numbers of translations for our language similarity ordering, these numbers should also generalize to polysemy between language combinations in general, i.e. the semantic structure of the database represents the polysemy between languages. A detailed database evaluation is needed to confirm this hypothesis.

The number of translations of an expression has been shown to affect translation performance, and also semantic similarity between expressions decreases if multiple translations of an expression exist (Tokowicz et al., 2002). Therefore, stimulus materials should be controlled for the number of translations of expressions used in experimental studies.

The specific numbers of translations for individual expressions as observed by the automatic translation procedure in the database can be summed to determine the degree of polysemy between language combinations in the database. Thus, besides individual numbers of translations for each cognate, also total numbers of translations are useful to keep accounts of. In Tokowicz et al. (2002), the numbers of translations of 562 Dutch words into English were obtained by using a translation method, in which participants produce their first spontaneous translation of a given word.

**Methods.** We used the relations in the translation database to identify the number of translations per item. These individual numbers were summed to retrieve the total number of translation between a language combination. We measured the number of translations by counting every translation in the database smaller than 8 letters, making no distinction between frequently and infrequently used translations. The total number of translations between two languages can be simply expressed by $T$ in Equation 2.

$$T = \sum_{k=0}^{K} \sum_{l=0}^{L} \sum_{m=0}^{M} \sum_{n=0}^{N} \begin{pmatrix} if(l.Concept = n.Concept) \\ AND \\ if(l.RelationArray \text{ contains a value in } n.RelationArray \end{pmatrix}$$

$$K = Source\ language\ expressions$$
$$L = Readings\ of\ expression_k$$
$$M = Destination\ language\ expressions\ relating\ to\ the\ concept\ of\ reading_l$$
$$N = Readings\ of\ translation_m$$

**Equation 2: Principle method for determining the number of translations in a language combination.**

For every translation pair that shares a specific relation to a specific concept, the number of translations counter was updated by one. For instance, *bank-bank* shares an exact same relation (*cash dispenser*) to the exact same concept (*financial institution*). The numbers of translations for individual cognates were saved in the lists of identified cognates.

The total numbers of translations were used to compute mean numbers of translations (column 3 in Table 5) for indefinite words in the source lexicons. These means are a rough reflection of the polysemy across languages in the database. A mean was computed by dividing the total number of translations by the number of analyzed words from the source language. These means are only applicable for mean number of translations of source language expressions, since means were computed using analyzed words form the source language. Mean number of translations for destination language words were not computed, because each analyzed destination expression would have been stored in a hashtable to be able to check every further destination expression on its occurrence.

To further examine the dependence of cross language similarity on the number of translations, proportions of cognates to translations were computed. Dividing cognates by translations, one determines in how many translations a cognate appears. If the language similarity ordering is corrected using these proportions, languages with high proportions would become less similar, because more cognates appear in less translation pairs.

| language combination | number of translations | mean number of translations | cognates: translations |
|---|---|---|---|
| english-french | 60000 | 2.7 | **0.16** |
| english-spanish | 59000 | 2.7 | 0.13 |
| english-german | 53000 | 2.4 | 0.15 |
| english-italian | 47000 | 2.1 | 0.16 |
| french-spanish | 43000 | 2.7 | 0.21 |
| dutch-english | 40000 | 1.9 | 0.21 |
| german-spanish | 38000 | 1.8 | 0.13 |
| french-german | 38000 | 2.3 | 0.15 |
| italian-spanish | 37000 | 2.6 | **0.31** |
| french-italian | 37000 | 2.2 | 0.25 |
| german-italian | 31000 | 1.6 | 0.16 |
| dutch-german | 31000 | 1.5 | **0.42** |
| dutch-spanish | 30000 | 1.4 | 0.18 |
| dutch-french | 30000 | 1.4 | 0.21 |
| dutch-italian | 25000 | 1.2 | 0.22 |

**Table 5: Total numbers of translations, mean numbers of translations, and numbers of cognates relative to the total numbers of translations across languages.**

**Results.** It can be seen that the mean numbers of translations differ much across language combinations. For instance, there are almost two times as much translations of English expressions to French compared to Dutch-French. Because such observed differences cannot be explained easily, we chose not to correct numbers of cognates with the total numbers of translations. Among the ruffling factors is a degree of noise in the used database. The number

of relations between language combinations is probably influenced by a collector's bias. Although the database has been built up by a Dutch company, probably taking Dutch as a reference point, language combinations with English show more translations than others as seen in the sorting of Table 5. Future analysis might provide a more detailed database evaluation to investigate this issue in detail. Furthermore, the differences between the proportions of cognates to translations are not easily explained either: It is not clear to what extent these proportions are language dependent. For instance, the proportions, printed bold in Table 5, show that high numbers of cognates (as seen in Table 4) can mean that there are actually few cognates with respect to the numbers of translations between the languages. However, since the number of translations of English-French could be odd, we will not use the differences between these proportions for a language similarity ordering.

**Validation.** A questionnaire to assess the degree of polysemy between languages was considered too hard for untrained linguists. Linguists should be approached to validate the ordering of polysemy across languages found in Table 5. One might expect that cultures that are more similar will have more words for the same or overlapping concepts. However, it may be the case that cultural and therefore language differences are too small to measure for the present series of closely related west-European countries. A translation database with more distant languages might be needed to answer the interesting question to what extent language distance, cultural distance, and conceptual distance might be related. As for now, it is not known to what extent the variance in observed numbers of translations is explained. One explanation could be the way in which concepts are represented by a different number of words across languages. Another explanation is the way the semantic structure in the database does not satisfy the semantic structure of languages.

The language similarity ordering based on proportions of cognates to translations confirms the suggestions to validate the observed numbers of translations. Based on numbers of cognates, the correlations were 0.91 (intuitions) and 0.72 (evolution), based on proportions, the correlations were 0.72 and 0.64, compared to the two validation measures.

**Conclusion.** We have determined mean numbers of translations across language combinations. However, there are unexplained differences between these means. A validation in the form of a comparison to other studies is needed to safely make use of these values. It did not appear to be useful to put up a language similarity ordering using proportions of cognates to translations because of these differences.

## Study 5 – Proportion False Friends to Cognates

**Goal:** *To determine the dependence of a language similarity ordering on the number of false friends between language combinations*

Like cognates, false friends are a special word type that is of interest to both psycholinguists and linguists. False friends are harder to understand, because they combine one orthographic form with two different meanings (Klein & Doctor, 1992). The occurrence of false friends is generally assumed to be the result of coincidental form overlap, for instance in terms of lexical or sublexical orthotactics or phonotactics. On the one hand, if the existence of false friends would be completely due to chance, their proportion would be similar across languages; on the other hand, their occurrence might be an indication of compatible or even similar orthotactic or phonotactic rules in the languages considered. In this case, the number of false friends would signal a form of language similarity.

Because false friends are useful stimuli in psycholinguistic studies, we wanted to record the occurrences of false friends in the database analysis. Furthermore, we wished to compare the proportions of identical false friends to identical cognates, because a language similarity ordering could be assumed to depend on both. Cognates might affect the ordering because they share their origins and false friends because they might be coincidently unrelated.

**Methods.** Only identical false friends were analyzed to make the analysis less complex, although it is also possible to analyze form-similar false friends with the current implementation. Form-identical false friends were retained from the set of form-identical homographs by excluding the translations. These homographs were found by looking up every expression from every source lexicon in the destination lexicon. The resulting expressions were restricted to have lengths between 3 and 8 characters.

When looking up the form-identical homographs of a given expression, the first character was set to uppercase when there was no form-identical homograph found for lowercase, and to lower case when no form-identical homograph was found for uppercase. This way, looking up German homographs would return these typical nouns that have uppercase first characters in German. However, applying this wrinkle led to small inconsistencies violating the expectation that the number of identical cognates plus the number of identical false friends, sums to the number of identical homographs (see Table 6). The numbers of identical cognates in the table were adjusted so that cognates with multiple shared meanings were only counted once[1].

The remaining inconsistencies arise because on the one hand, before the process of determining orthographic distance of cognates, cognates were set to lower case (the complete word), while on the other hand, in the process of looking up homographs, only the first

[1] The numbers of cognates used for the language similarity ordering were determined by counting cognates sharing multiple meanings, for every shared meaning (see for examples Study 2 – Semantic Similarity).

character of a word was set to lower case. As a consequence, words with upper case characters at places other than the first, which have identical homographs in other languages that do not have these uppercase characters, are not returned when looking up homographs. A consideration to overcome these inconsistencies was to use lexicons containing only lowercase words. But by doing this one would get conflicts between words like *mars* and *Mars,* which have different meanings. Either way, the numbers of homographs in Table 6 are small underestimations of the homographs present in the database. Therefore, some false friends may not have been counted as well since false friends are retained from the set of identical homographs.

| Language combination | unique cognates | false friends | homo graphs | false friends: cognates |
|---|---|---|---|---|
| english-french | 2207 | 644 | 2840 | 0.069 |
| english-german | 2276 | 522 | 2712 | 0.067 |
| dutch-english | 2463 | 522 | 2971 | 0.061 |
| french-german | 1637 | 314 | 1894 | 0.055 |
| dutch-french | 1823 | 305 | 2120 | 0.049 |
| english-italian | 1243 | 281 | 1518 | 0.038 |
| german-italian | 1031 | 190 | 1201 | 0.037 |
| dutch-german | 3232 | **448** | 3560 | 0.035 |
| german-spanish | 793 | 164 | 946 | 0.034 |
| english-spanish | 1083 | 258 | 1335 | 0.033 |
| dutch-italian | 1115 | 172 | 1279 | 0.031 |
| dutch-spanish | 805 | 157 | 955 | 0.030 |
| italian-spanish | 2036 | **327** | 2360 | 0.028 |
| french-italian | 1073 | 242 | 1311 | 0.027 |
| french-spanish | 884 | 197 | 1075 | 0.022 |

**Table 6: Proportions of identical false friends to identical cognates. High proportions indicate relatively many false friends. Values in column 2 and column 3 should add up to column 4.**

**Results.** The numbers of false friends and proportions of false friends to cognates are found in Table 6. The language combination English-French has the largest number of false friends, and the highest proportion when divided by the number of cognates. It is remarkable that language combinations with English are found high in the ordering. This result may be ascribed to characteristics of the English language that are similar to both Romance and Germanic languages, for instance, in terms of spelling. As such, that might induce more coincidental form overlap. The proportions of form-identical false friends to from-identical cognates show that English is a rather different language because it does not appear low in the ordering of Table 6: English has relatively many false friends with other languages. On the other hand, the numbers of false friends of the two most similar languages (printed in bold) appear low in the ordering. Furthermore, the relatively low numbers of false friends according to our own intuitions may be caused by the way we classify semantic similarity. It is

suggested that translation pairs like *bank – bank* can also be classified as false friends since not all meanings are shared by both words.

**Validation.** For this study we have used the two validation measures to compare a language similarity ordering based on proportions of false friends to cognates with the ordering from study 3, based on numbers of cognates. The ordering based on proportions of false friends to cognates was less consistent with the validation studies (correlations of 0.00 and 0.01), as compared to the language similarity ordering based on numbers of cognates (correlations of 0.91 and 0.72). This analysis indicates that the numbers or proportions of false friends should not be used as a measure of similarity, other ways in which false friends can be used for cross-linguistic similarity are to be studied.

**Conclusion.** We have counted false friends between language combinations in the translation database in order to determine the dependence of a language similarity ordering on numbers of false friends (and to identify false friends). A more detailed study should be done with respect to the relation of false friends to a language similarity ordering. This study should consider reasons for the existence of false friends (coincidence, spelling overlap) and should also examine the relationship between false friends and (identical) cognates in more detail.

## Study 6 – Proportion Form-Similar to Form-Identical Cognates

**Goal:** *To determine the dependence of a language similarity ordering on orthographic similarity.*

The last study is about the dependence of a language similarity ordering on the inclusion of form-similar cognates. Cognates may have similar forms across languages because they were adopted from a shared common root language or because they were useful borrowings or loan words. Depending on time and writing systems, they stayed identical in alphabetic form or underwent certain changes in orthography (spelling and capitalization). We think that language combinations with relatively many form-similar cognates have changed more than languages with relatively many form-identical cognates. If that is correct, language change should be predictable on the basis of the proportion of form-similar versus form-identical cognates. Such proportions are studied in this section.

Language change is also important for a language similarity ordering, because this ordering depends primarily on the number of words with similar form and meaning. From an evolutionary perspective, language distance might depend on how long ago certain languages branched off. And the proportion of form-similar versus form-identical cognates might also, to a certain degree, be dependent on these same branches. To evaluate these notions, the proportions of form-similar to form-identical cognates are compared to the validation studies used earlier.

**Methods.** Occurrences of both form-similar and form-identical cognates were counted across languages. When translation pairs scored 1.0 on form similarity, the counter for form-identical cognates was updated. If this score was above the threshold of 0.5 (except for 1.0), the counter for form-similar cognates was updated. As before, we only counted cognates with lengths between 3 and 8, and also counted cognates for every shared meaning. The proportions of form-similar to form-identical cognates for different language combinations are found in Table 7.

| language combination | similar | identical | similar: identical |
|---|---|---|---|
| dutch-english | 5744 | 2865 | 2.0 |
| dutch-french | 4206 | 2063 | 2.0 |
| english-german | 5174 | 2576 | 2.0 |
| french-german | 3875 | 1850 | 2.1 |
| dutch-german | 9123 | 3785 | 2.4 |
| english-french | 6559 | 2727 | 2.4 |
| italian-spanish | 8787 | 2698 | 3.3 |
| dutch-italian | 4332 | 1232 | 3.5 |
| german-italian | 4079 | 1108 | 3.7 |
| english-italian | 6041 | 1389 | 4.3 |
| german-spanish | 3925 | 869 | 4.5 |
| english-spanish | 6507 | 1330 | 4.9 |
| dutch-spanish | 4409 | 889 | 5.0 |
| french-italian | 7639 | 1232 | 6.2 |
| french-spanish | 8029 | 1091 | 7.4 |

**Table 7: Proportions of form-similar to form-identical cognates. High proportions reflect relatively many form-similar cognates.**

**Results.** The proportions found for form-similar to form-identical cognates are rather different across language combinations. It appears that French-Italian and French-Spanish are deriving their high places in the original ordering from the relatively large number of form-similar cognates. Therefore, an ordering based on only identical cognates would not take into account the high similarity, although not form-identical, between such language combinations. An ordering based on proportions of form-similar to form-identical cognates seems to be of profit to languages that have many loanwords in other languages, since loanwords often turn up as form-identical cognates. Assuming that this is true, Germanic languages seem to borrow more (identical) words from each other and from French as compared to the Romance languages.

**Validation.** The observed proportions probably do not encompass language change and language distance, as such. The divergence of semantic overlap is also of importance for language change, since not only orthography diverges in time but also the meaning of words. For instance, the Dutch-German translation pair *smart-Schmertz* (meaning *sorrow*) has changed with respect to orthography, whereas the identical homograph *smart* appears in English having a completely different meaning. For linguists, the observed proportions are still of interest to research the predictability of language change by the degree of form-overlap in translation pairs, although the importance of divergence in meaning has to be assessed as well.

The new ordering was not consistent with validation studies. The correlation of this ordering based on proportions with intuition studies was 0.06, and the correlation with evolution studies was 0.09. An ordering based on identical cognates correlated 0.61 with

intuition studies, and 0.48 with evolution studies, so identical cognates are of more importance for a language similarity ordering than using only similar cognates. Still, the sum of them is clearly most consistent with the validation studies.

**Conclusion.** The understanding of language distance one obtains on the basis of these proportions, depends strongly on the language combination considered. One possibility is that differences in the spelling systems of the languages involved affect the number of form-similar versus form-identical cognates. Further study could investigate this point by considering the proportion of form-similar versus form-identical false friends (e.g., the Dutch-German false friend *knap-knapp*, meaning 'wise' or 'pretty' in Dutch, and meaning 'tight' in German). Because cognates share also meaning, the proportions form-similar to form-identical cognates depend also on semantic change.

Furthermore, we have found that numbers of form-similar cognates are important for a language similarity ordering.

## Discussion

By means of a translation database we were able to automatically identify distributions of cognates with respect to form-similarity in various European languages. In the first two studies we have addressed specific issues concerning measures for orthographic similarity and semantic similarity. The measures were carefully compared to validation studies. In Study 3 we have discussed the results obtained when applying these two measures to the translation database, and constructed a language similarity ordering based on the numbers of cognates. In the remaining studies we have discussed the numbers of translations and false friends, as well as proportions of form-similar to form-identical cognates. Each account was related to the language similarity ordering constructed in Study 3, and compared to validation studies.

The present work has been an exploration of the possibilities with new available techniques from artificial intelligence. Therefore we focussed on the interests that this work has for researchers in different fields. We have successfully made use of the Rosetta Schemes, so we conclude that these can provide researchers with useful information contained in the professional translation database we have used. Researchers can be psycholinguists who want to construct stimulus materials for their experiments. Stimulus materials can be controlled to have the desired distribution of orthographic similarity. Also the number of translations for each item is available, so that it is possible to control for polysemy. And, as we have done ourselves, it is possible to extract false friends. Another interest for our work is for linguists who are interested in cross-linguistic similarity. By counting the numbers of cognates, false friends and translations between semi-complete lexicons, an accurate and detailed language distance can be determined. These numbers and the distributions with respect to form-similarity can be used to study language similarity and language change.

The Levenshtein distance used for measuring orthographic similarity was validated by comparing translation pairs with orthographic similarity ratings from Tokowicz et al. (2004). 77,7% of these translation pairs were classified using a threshold on the metric we constructed with respect to the orthographic similarity ratings. The setting of the threshold did have a high influence on the number of correctly classified cognates. We suggested that the best fitting threshold and the best fitting correction for word length can be estimated more precisely using techniques from information retrieval.

The automatic translation procedure used for classifying translation pairs was validated by counting how many of experimentally approved translation pairs (adopted form Tokowicz et al., 2004) were found using the semantic structure in the translation database. We found a classification rate of 86,2% recognized translation pairs. Since translations were restricted to share the same relation with source expressions, not all semantically similar translation pairs can be found by definition. So to improve the performance of automatic translation, more translations can be retrieved by also allowing semantically similar translations to be classified. In practice this means that only head translations were considered, whereas also additional translations and synonyms were represented in the database. However, another point is to be made on the ratings of translation pairs that we adopted, because we found that most of the unrecognized translation pairs were explained by having little semantic overlap. Therefore, the classification rate also represents the degree in which high ratings really are semantically similar.

Applying the metric for orthographic similarity to the translation pairs found by the automatic translation procedure, we were able to keep accounts of several specific word types. The observed numbers of cognates across language combinations were ordered and compared with language similarity orderings based on intuitions of Dutch-English bilingual language users and a language evolution study by Gray and Atkinson (2003). Our findings correlated 0.91 with the intuitions ($p=0.00001$) and 0.72 with the evolution study ($p=0.001$).

The differences between the observed language similarity ordering and the validation studies can be explained by differences in the way cross-linguistic similarity is determined. Most notably, both validation studies did not deal with semi-complete lexicons. Judging the intuitions of our bilingual language users, it is plausible that they only considered the set of frequently used words from the lexicons we used. Also in the language evolution study, only frequently used were used to determine language origins. According to Pagel et al. (2007), word frequency accounts for 50% of lexical replacement (divergence of characters over time between translation pairs). Therefore, a language similarity ordering, based on frequently used words (which evolve at slower rates than infrequently used words), reflects more of the shared origins in the analyzed languages, compared to our study. The extensive lexical wordbase we used, reflects also lexical replacement of the infrequently used words, since it consists of semi-complete lexicons. It would be interesting to order languages on numbers of high frequency cognates and compare that with our present findings. Such a cross-linguistic similarity measure with respect to frequently used words would predict to what extent language users can actually understand each other, because lexical replacement of infrequently used words does not affect the understanding of speech. Anyhow, the language similarity ordering we constructed is different compared to studies based on high word frequency, because infrequently used words can also be cognates (i.e. loanwords). For example, the reason why English-French, -Spanish and -Italian appear high in our language similarity ordering is attributed to the many loanwords that the English language adopted from French during a French invasion in the middle ages.

In the remaining three studies we addressed the dependence of a language similarity ordering on numbers of translations, numbers of false friends and numbers of form-similar cognates, respectively. In the numbers of translations across language combinations, too much unexplained differences were observed, to be of use for a language similarity ordering. A collector's bias was among the suggested reasons for this. We suggested that a more detailed evaluation study of the database used here is desirable. Besides cognates, numbers of false friends were determined because language distance was assumed to depend on both. However, a language similarity ordering based on numbers of cognates proved to be most consistent with the validation studies. As a final account to study factors that predict language distance, we studied the proportions of form-similar to form-identical cognates, which were assumed to predict language change over time. However, it was concluded that differences in spelling systems between languages are more obvious causes to affect these proportions.

Besides distributions of cognates with respect to form-similarity, it would be of interest for determining orthographic change over time to identify distributions of false friends with respect to form-similarity. Since the differences in proportions of form-identical false friends to form-identical cognates were explained by the differences in spelling systems between languages. For example, it was suggested that English has characteristics of both Romance

and Germanic languages which causes more coincidental form-overlap. Because we think that false friends are the result of coincidental form-overlap within the orthotactic and phonotactic rules of the languages, such distributions can reveal the spelling systems were distributions of cognates also depend upon.

As a result of our focus on the interests for the present work, we did not optimize the classification rates of the orthographic similarity measure and the automatic translation procedure with respect to the validation studies. However, such maximizations can be performed if there is still interest for this. Furthermore, our use of the Rosetta Schemes for database interaction was limited in the sense that we only made use of expressions, readings, relations, and concepts. Among the other types of information we suggest that the syntactic categories that are available in Euroglot are very useful, as our lists of cognates now also include proper nouns. In a follow up to the present work we will also want to construct a database for stimulus materials, specialized to the desires of researches. Such a database could then be compared to stimulus materials collected by other researchers, in order to evaluate a possible collector's bias in Euroglot.

For future studies we are intersted in applying classification algorithms to identify language clusters on the basis of the observed findings. We are also interested in simulating orthographic change using MCMC. The distributions with respect to the orthographic similarity of random strings can then be compared to the distributions that we have identified here.

## Acknowledgements

I am thankful to my supervisors Ton Dijkstra and Franc Grootjen, they have invested more time than required in this project. The deliberations with Ton Dijkstra did not only result in a solid research design, but also motivated me to think further about aspects to language evolution and beyond. Also the commentary on an earlier draft of this paper was indispensible. Franc Grootjen started as the pragmatic provider of resources but revealed to be a provider of vital suggestions to the classification procedures and my programming style. I also really appreciate the interest others have showed when I was working on this thesis.

## Implementation

This section governs the various resources and procedures that were used and developed for the performed studies. The resources include the Rosetta schemes, and the database extracts. The procedures include a dictionary reader, the automatic translation procedure, a procedure that calculates the Levenshtein distance, a procedure for maintaining the numbers of various word types, and a procedure that processes words from the database. Figure 5 shows how these procedures interact with the resources.

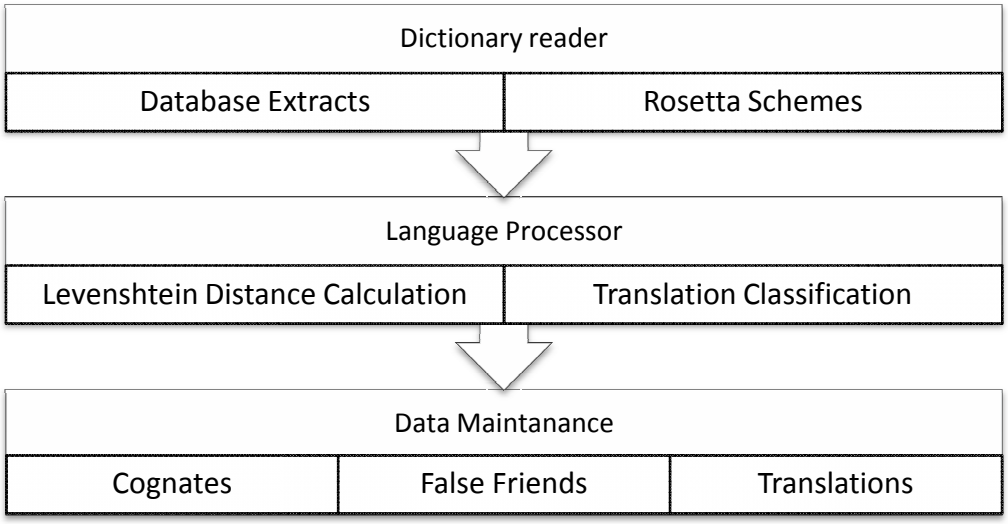


**Figure 5: Structure of the interaction between resources and procedures.**

The Rosetta schemes are a set of classes that put information from the database extracts into useful data structures. By these simple statements

```
DictionaryDocument dictionaryDocument;
File file=new File("English+".xml");
dictionaryDocument=DictionaryDocument.Factory.parse(file,parseOptions);
Dictionary dictionary=dictionaryDocument.getDictionary();
```

an object is made that contains all the basic types of information from the database extracts. Among these basic types are fields that represent information such as: Word, Concept, Reading, Expression and Relation. But also fields that we did not made use of such as: Gender, Morphology, Sound, and Word Type. To print each expression from this dictionary, the following statements have to be executed.

```
Word[] words=dictionary.getWordArray();
for(int i=0;i<words.length;i++)
{
   String expression=words[i].getExpression();
   System.out.println(expression);
}
```

Because all information of our interest is represented in the `Word` class, the `words` were extracted from the dictionary object and practically represented in two specific data structures:

```
private Map<String,Word> wordByExpression;
private Map<String,List<Word>> wordsByConcept;
```

The first object maps the `Word` object to its corresponding expression for each expression in the database. Note that every `Word` is identified by its expression. This object is used by the language processor to be able to iterate over all of these expressions. The second object maps all the `Words` that relate to a specific reading to that concept for each concept in the database. This object is used by the automatic translation procedure to obtain the `Words` from the destination language that relate to the concept of the current source language expression. Together these two objects contain all the information that was needed for the language processor.

Now it is possible for the language processor to simply iterate through every source language expression. In order to be able to calculate the Levenshtein distance, translation equivalents are identified first of all. This is done in the following way:

```
private Map <String, String> translate(Word word, DictionaryReader german)
{
  Map<String,String> expressionByReading = new HashMap<String,String>();
  for(Reading reading:word.getReadingArray())
  {
    Set<Integer> relations = toSet(reading.getRelationArray());
    for(Word translation:german.getTranslation(reading.getConcept()))
    {
      for(Reading translatedReading:translation.getReadingArray())
      {
        if (reading.getConcept().equals(translatedReading.getConcept()))
        {
          for (Relation translatedRelation:translatedReading.getRelationArray())
          {
            if (relations.contains(translatedRelation.getNumber()))
            {
                ExpressionByReading.put(translatedReading.getId(),translati
                on.getExpression());
            }
          }
        }
      }
    }
  }
  return expressionByReading;
}
```

Also before iterating over these translations, the set of form-identical homographs is identified to be able to determine which of these are false friends of the source language expression later on. The process of iteration over the set of translation equivalents, follows the following order of analysis:

1. Determine orthographic similarity
2. Save cognate
3. Remove the translation from the set of homographs (if it is one)
4. Update counters.

In the first step the Levenshtein distance is calculated and the formula is applied. If a subsequent check on the threshold yields to a successful evaluation, the expression-translation pair is classified as a cognate. In step two this cognate is saved along with the matched reading number of the expression and the number of translations the expression has in the destination language. In step three we remove the current translation from the set of homographs in order to end up with a set of homographs that contains no more translations: the set of false friends. Since no form-similar false friends are considered in the current version, this set will contain only 1 or 0 elements. Step 4 keeps account of various word types. After processing a language combination, these counters represent the total numbers of the various word types.

Since the Rosetta schemes were implemented in Java, we chose to develop every procedure in Java, using the Eclipse platform. Because many classes are contained in the Rosetta schemes, the JAR format was used to aggregate them into one file. For current implementations of the procedures to be executed, we used Subversion because the database extracts themselves were not owned by the programmer. A limited database was made available to be able to extract the bugs out. Using subversion, there were 8 updates needed to revise the implementation because some additions were developed and bugs were found after the first version.

| Revision | Reason |
|---|---|
| 8 | Translations were unintentionally set to lowercase, so too many false friends were retained from the set of homographs. |
| 7 | Threshold was lowered from 0.6 to 0.5 to classify more translation pairs as cognates |
| 6 | A check on minimal word length was added. |
| 5 | Bug with a check on maximal word length was found: <= instead of <. |
| 4 | Corrections to the translation procedure were made. |
| 3 | Translation pairs were now checked on their presence in Tokowicz' stimulus materials. Spanish language was made available. |
| 2 | Bug with capitals found when retrieving homographs. |
| 1 | False friends experiment implemented. |

**Figure 6: Revisions to the implementation were communicated using Subversion.**

## References


Dijkstra, A., Brummelhuis, B., Baayen, H. (submitted). How cross-language similarity affects cognate recognition.

Doctor, E., Klein, D. (1992). Phonological word processing in bilingual word recognition. In R. J. Harris (Eds.), *Cognitive processing in bilinguals,* 237-252. New York: Elsevier.

Friel, B., Kennison, S. (2001). Identifying German-English cognates, false cognates, and non-cognates: methodological issues and descriptive norms. *Bilingualism: Language and Cognition, 4(3),* 249-274.

Gooskens, C., Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language variation and change*, *16(3)*, 189-208.

Gray, R., Atkinson, Q. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature, 426(27),* 435-439.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance.* Doctoral dissertation. University of Groningen.

Linguistic Systems B.V. (2008). *Euroglot Professional 5.0.* Website: http://www.euroglot.nl/

Pagel, M., Atkinson, Q., Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature, 449(11),* 717-720.

Serva, M., Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. *Europhysics letters, 81(996).* 68005-6900.

Tokowicz, N., Kroll, J., de Groot, A., van Hell, J. (2002). Number of translation norms for Dutch-English translation pairs: a new tool for examining language production. *Behavior Research Methods, Instruments, & Computers, 34(3),* 435-451.

# Appendix 1

Dutch – English example translation pairs obtained using the metric discussed in Study 1 – Orthographic Similarity and using the semantic structure from a translation database discussed in Study 2 – Semantic Similarity

| Dutch | English | Score |
|---|---|---|
| typiste | typist | 0.8571 |
| ultra- | ultra- | 1.000 |
| underdog | underdog | 1.000 |
| unionist | unionist | 1.000 |
| update | update | 1.000 |
| urgent | urgent | 1.000 |
| urgent | urgently | 0.7500 |
| vacuüm- | vacuum | 0.7143 |
| vagebond | vagabond | 0.8750 |
| vamp | vamp | 1.000 |
| variant | variant | 1.000 |
| variété | variety | 0.7143 |
| veda | Veda | 1.000 |
| velvet | velvet | 1.000 |
| verdict | verdict | 1.000 |
| vers | verse | 0.8000 |
| vers | verse | 0.8000 |
| vest | vest | 1.000 |
| veto | veto | 1.000 |
| vice- | vice- | 1.000 |
| video | video | 1.000 |
| videotex | videotex | 1.000 |
| viewdata | viewdata | 1.000 |
| villa | villa | 1.000 |
| vinyl | vinyl | 1.000 |
| viola | viol | 0.8000 |
| virus | virus | 1.000 |
| visie | vision | 0.6667 |
| visueel | visual | 0.7143 |
| visueel | visually | 0.6250 |
| vitriool | vitriol | 0.8750 |
| vitriool | vitriol | 0.8750 |
| vocatief | vocative | 0.7500 |
| volleyen | volley | 0.7500 |
| volt | volt | 1.000 |
| vont | font | 0.7500 |
| vopo | VOPO | 1.000 |
| vopo | VOPO | 1.000 |
| vopo | Vopo | 1.000 |
| vopo | Vopo | 1.000 |
| vork | fork | 0.7500 |
| vork | fork | 0.7500 |
| vorm | form | 0.7500 |
| vorm | form | 0.7500 |
| vorm | form | 0.7500 |
| vorm | form | 0.7500 |
| vorm | form | 0.7500 |
| voucher | voucher | 1.000 |
| waden | wade | 0.8000 |
| wafel | wafer | 0.8000 |
| wafel | waffle | 0.6667 |
| wagen | waggon | 0.6667 |
| wagen | wagon | 0.8000 |
| wajang | wayang | 0.8333 |
| walhalla | Valhalla | 0.8750 |
| walhalla | Walhalla | 1.000 |
| walkman | Walkman | 1.000 |
| wallaby | wallaby | 1.000 |
| wapen | weapon | 0.6667 |
| wapen | weapon | 0.6667 |
| warmte | warmth | 0.8333 |
| warmte | warmth | 0.8333 |
| wattage | wattage | 1.000 |
| werkster | worker | 0.6250 |
| western | western | 1.000 |
| westers | western | 0.8571 |
| wever | weaver | 0.8333 |
| whippet | whippet | 1.000 |
| whisky | whiskey | 0.8571 |
| whisky | whisky | 1.000 |
| winter- | winter | 0.8571 |
| winters | wintery | 0.8571 |
| winters | wintry | 0.7143 |
| would-be | would-be | 1.000 |
| zeewater | seawater | 0.7500 |
| zeewind | sea wind | 0.6250 |
| zeloot | zealot | 0.6667 |
| zelote | zealot | 0.6667 |
| zifting | sifting | 0.8571 |
| zigzag | zigzag | 1.000 |
| zionisme | Zionism | 0.8750 |
| zioniste | Zionist | 0.8750 |
| zoeker | seeker | 0.6667 |
| zonlicht | sunlight | 0.6250 |
| zwelling | swelling | 0.8750 |

## Appendix 2

Adopted cognates from Tokowicz et al. (2002) which are not classified as cognates by the automatic metric for orthographic similarity. The orthographic rating by Tokowicz et al. is found in the third column.

| Dutch | English | O-Rating | P-Rating | Score |
|---|---|---|---|---|
| cirkel | circle | 6.38 | 6.88 | 0.5000 |
| mijl | mile | 6.25 | 6.62 | 0.5000 |
| naam | name | 6.12 | 6.88 | 0.5000 |
| vaas | vase | 6.12 | 7.00 | 0.5000 |
| aap | ape | 6.12 | 5.88 | 0.3333 |
| kin | chin | 6.0 | 6.75 | 0.5000 |
| schaap | sheep | 6.0 | 6.88 | 0.5000 |
| zee | sea | 6.0 | 7.00 | 0.3333 |
| zon | sun | 6.0 | 7.00 | 0.3333 |
| jaar | year | 5.88 | 7.00 | 0.5000 |
| pijp | pipe | 5.88 | 6.62 | 0.5000 |
| roos | rose | 5.88 | 7.00 | 0.5000 |
| voet | foot | 5.88 | 6.75 | 0.5000 |
| pool | pole | 5.75 | 4.75 | 0.5000 |
| hitte | heat | 5.75 | 6.88 | 0.4000 |
| elleboog | elbow | 5.62 | 7.00 | 0.5000 |
| sneeuw | snow | 5.62 | 7.00 | 0.5000 |
| vader | father | 5.62 | 6.88 | 0.5000 |
| zeep | soap | 5.62 | 6.88 | 0.2500 |
| bad | bath | 5.5 | 6.88 | 0.5000 |
| daad | deed | 5.5 | 6.12 | 0.5000 |
| thee | tea | 5.5 | 7.00 | 0.5000 |
| steen | stone | 5.5 | 7.00 | 0.4000 |
| kans | chance | 5.5 | 6.88 | 0.3333 |
| hoed | hat | 5.5 | 6.75 | 0.2500 |
| boezem | bosom | 5.43 | 6.71 | 0.5000 |
| maan | moon | 5.38 | 7.00 | 0.5000 |
| zeil | sail | 5.38 | 6.50 | 0.5000 |
| huis | house | 5.38 | 7.00 | 0.4000 |
| stroming | stream | 5.38 | 5.75 | 0.3750 |
| pokken | pox | 5.38 | 6.62 | 0.3333 |
| dij | thigh | 5.29 | 7.00 | 0.2000 |
| hol | hollow | 5.25 | 6.12 | 0.5000 |
| honing | honey | 5.25 | 6.88 | 0.5000 |
| leen | loan | 5.25 | 5.88 | 0.5000 |
| muziek | music | 5.25 | 7.00 | 0.5000 |
| viool | violin | 5.25 | 7.00 | 0.5000 |
| zomer | summer | 5.25 | 7.00 | 0.5000 |
| rijst | rice | 5.25 | 7.00 | 0.4000 |
| tijd | time | 5.12 | 6.88 | 0.5000 |
| aarde | earth | 5.12 | 7.00 | 0.4000 |
| borst | breast | 5.0 | 6.38 | 0.5000 |
| katoen | cotton | 5.0 | 7.00 | 0.3333 |
| ritme | rhythm | 5.0 | 6.88 | 0.3333 |
| jeugd | youth | 5.0 | 6.62 | 0.2000 |

## Appendix 3

Adopted cognates from Tokowicz et al. (2002) which are not classified as translation pairs by the automatic translation procedure for semantic similarity. The first 20 are discussed in the validation section of Study 2 – Semantic Similarity.

| Ducth | English | Rating |
|---|---|---|
| dorpje | village | 7.00 |
| geloof | religion | 7.00 |
| lammetje | lamb | 7.00 |
| mist | mist | 7.00 |
| steegje | alley | 7.00 |
| verraad | betrayal | 7.00 |
| pop | puppet | 6.88 |
| vrouw | female | 6.88 |
| ede | oath | 6.75 |
| gemeen | cruel | 6.75 |
| graaf | duke | 6.75 |
| huurder | renter | 6.75 |
| kijken | watch | 6.75 |
| snoer | wire | 6.75 |
| spoor | rail | 6.75 |
| bandiet | crook | 6.62 |
| jammer | pity | 6.62 |
| pokken | pox | 6.62 |
| voorkeur | favour | 6.62 |
| waard | worth | 6.62 |
| blaam | blame | 6.50 |
| boot | ship | 6.50 |
| geloof | believe | 6.50 |
| kerel | lad | 6.50 |
| loon | salary | 6.50 |
| rail | rails | 6.50 |
| zal | will | 6.50 |
| zorgen | care | 6.50 |
| bewijs | prove | 6.38 |
| bot | rude | 6.38 |
| daling | descent | 6.38 |
| eerbied | honor | 6.38 |
| gala | ball | 6.38 |
| gemeen | crude | 6.38 |
| heer | sir | 6.38 |
| kaartje | postcard | 6.38 |
| kostuum | costume | 6.38 |
| loon | payment | 6.38 |
| neger | negro | 6.38 |
| salaris | wage | 6.38 |
| straat | road | 6.38 |
| troep | trash | 6.38 |
| bankje | bench | 6.29 |
| draadje | thread | 6.25 |
| groot | huge | 6.25 |
| jurk | gown | 6.25 |
| kwaad | anger | 6.25 |
| noodlot | fate | 6.25 |
| vader | dad | 6.25 |
| verkoop | sell | 6.25 |
| vordeel | favour | 6.25 |
| taak | duty | 6.14 |
| afval | waste | 6.00 |
| noodzaak | need | 6.00 |
| recent | current | 6.00 |
| verlies | defeat | 6.00 |
| vloek | spell | 6.00 |
| vuil | dirt | 6.00 |
| zwaar | rough | 6.00 |
| aap | ape | 5.88 |
| aarde | soil | 5.88 |
| bedrog | betrayal | 5.88 |
| dief | crook | 5.88 |
| kloof | canyon | 5.88 |
| leen | loan | 5.88 |
| schotel | saucer | 5.88 |
| stoer | tough | 5.88 |
| basis | basic | 5.75 |
| hard | tough | 5.75 |
| rouw | grief | 5.75 |
| stroming | stream | 5.75 |
| verdelen | part | 5.75 |
| dame | dame | 5.71 |
| beroemd | fame | 5.62 |
| inwoner | citizen | 5.62 |
| molen | windmill | 5.62 |
| ochtend | dawn | 5.62 |
| papier | sheet | 5.62 |
| rondje | circle | 5.62 |
| ruil | trade | 5.62 |
| sprookje | tale | 5.62 |
| vraag | demand | 5.62 |
| vrouw | lady | 5.62 |
| zwaar | tough | 5.62 |
| gangetje | alley | 5.50 |
| gast | chap | 5.50 |
| stok | pole | 5.50 |
| want | glove | 5.50 |
| zonde | pity | 5.50 |
| grond | floor | 5.43 |
| beeld | insight | 5.38 |
| grots | giant | 5.38 |
| wreed | crude | 5.38 |
| eenheid | measure | 5.25 |
| hardheid | cruelty | 5.25 |
| hoed | cap | 5.25 |
| jas | jacket | 5.25 |
| kijkt | watch | 5.25 |
| vriend | chap | 5.25 |
| overjas | cloak | 5.14 |
| moedig | bold | 5.12 |
| oneven | unequal | 5.12 |
| plan | idea | 5.12 |
| rugtas | bag | 5.12 |
| ruzie | riot | 5.12 |
| wachter | watch | 5.12 |
| bij | with | 4.88 |
| meid | chick | 4.88 |
| mode | mode | 4.88 |
| mooi | fair | 4.88 |
| preek | speech | 4.88 |
| regel | sentence | 4.88 |
| blokkade | block | 4.75 |
| drukte | crowd | 4.75 |
| keten | string | 4.75 |
| stand | mode | 4.75 |
| tafel | desk | 4.75 |
| haven | haven | 4.50 |
| hoofd | master | 4.50 |
| ruwheid | cruelty | 4.38 |
| staat | shape | 4.38 |
| fabriek | mill | 4.25 |
| verdriet | pain | 4.25 |
| blok | square | 4.12 |
| domein | property | 4.12 |
| kast | chest | 4.12 |
| schoen | boot | 4.12 |
| onlust | riot | 4.00 |
| hals | throat | 3.88 |
| mening | meaning | 3.62 |
| veld | domain | 3.62 |
| cyclus | circle | 3.25 |
| eenvoud | single | 3.25 |
| offer | offer | 3.25 |
| bleek | fair | 3.00 |
| cirkel | cycle | 2.88 |