Bilingualism: Language and Cognition 15 (1), 2012, 157–166 © Cambridge University Press 2011 doi:10.1017/S1366728910000623

RESEARCH NOTES

Distributions of cognates in Europe as based on Levenshtein distance*

JOB SCHEPENS
TON DIJKSTRA
FRANC GROOTJEN
Donders Institute for Brain, Cognition and Behaviour, Donders Centre for Cognition, Radboud University Nijmegen, The Netherlands

(Received: April 27, 2010; final revision received: October 22, 2010; accepted: November 5, 2010; First published online 11 August 2011)

Researchers on bilingual processing can benefit from computational tools developed in artificial intelligence. We show that a normalized Levenshtein distance function can efficiently and reliably simulate bilingual orthographic similarity ratings. Orthographic similarity distributions of cognates and non-cognates were identified across pairs of six European languages: English, German, French, Spanish, Italian, and Dutch. Semantic equivalence was determined using the conceptual structure of a translation database. By using a similarity threshold, large numbers of cognates could be selected that nearly completely included the stimulus materials of experimental studies. The identified numbers of form-similar and identical cognates correlated highly with branch lengths of phylogenetic language family trees, supporting the usefulness of the new measure for cross-language comparison. The normalized Levenshtein distance function can be considered as a new formal model of cross-language orthographic similarity.

Keywords: Levenshtein distance, cognate, orthographic similarity, language family

In the inhospitable mountain ranges of mid-Turkey, the following sentence in Babylonian-Assyrian cuneiform script was found on a clay tablet near the village of Boghazköy: *Nu ninda-ma ezzateni, wadar-ma ekkuteni* (Ceram, 1966, p. 75). The English translation of this Hittite sentence, dating from the 14th century BC, is "Now you will eat bread, further you will drink water". In 1915, the linguist Friedrich Hrozný was the first person able to translate this sentence on the assumption that Hittite was an Indo-European language. Knowing that *ninda* was the word for bread, he recognized two other words, because they are similar to their translation equivalents in languages that are presently around: *ezzateni*, related to "to eat", and *wadar*, related to "water".

Words like these, which have form-similar translations in other languages, are known as COGNATES in linguistic and psycholinguistic research. For the purpose of the present paper, cognates are defined as translation equivalents with high orthographic overlap (for phonological overlap effects in cognates, see Dijkstra, Grainger & Van Heuven, 1999). They can be either identical or similar in their printed form. For instance, the Dutch–English translation pair *sigaret – cigarette* is an example of a form-similar cognate, and *president – president* is an example of an identical cognate. Cognates may also vary somewhat in terms of their semantic similarity, which does not have to be complete. In particular, not all readings of a word in a source lexicon need to be the same as those of its translation equivalent in a destination lexicon. The Dutch–English translation pair *bank – bank* shares the meaning of "financial institution", but the English word *bank* also means "waterfront", a meaning the Dutch word does not possess; on the other hand, the Dutch word *bank* also refers to a sofa or a bench. In sum, orthographic and semantic dimensions of translation equivalents are clearly important if one wants to identify cognates in the vocabularies of pairs of languages.

In the present study, we discuss new tools from artificial intelligence that may help to automatically identify large numbers of cognates across language pairs. We consider this enterprise of considerable interest to both psycholinguists and linguists.

Psycholinguists have extensively used cognates to study language processing by bilinguals. In many reaction time studies, involving a variety of experimental paradigms, cognates were responded to faster than control words that exist in only one language. This COGNATE FACILITATION EFFECT has consistently been found in studies on bilingual word recognition in the visual modality (for reviews, see Dijkstra, 2005; Friel & Kennison, 2001; Voga & Grainger, 2007). The effect has also been observed in the auditory modality (Marian & Spivey, 2003) and

* In our study, we used the standard input–output functions of the following translation database: Euroglot professional 5.0 (2008), developed by Linguistic Systems B.V. We are grateful to Walter van Heuven, Gerard Kempen, Frank Leoné, Steven Rekké, Bastiaan du Pau, and two anonymous reviewers for their thoughtful comments on an earlier version of this paper.

Address for correspondence:
Job Schepens/Ton Dijkstra, Donders Centre for Cognition, Radboud University Nijmegen, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands
j.schepens@let.ru.nl

in word production (Costa, Caramazza & Sebastián-Gallés, 2000; Kroll & Stewart, 1994). Stronger facilitation effects may arise if the cognates in question exist in three languages rather than in two (Lemhöfer, Dijkstra & Michel, 2004). Recent evidence indicates that cognate effects can be modulated by sentence context (Duyck, Van Assche, Drieghe & Hartsuiker, 2007; Schwartz & Kroll, 2006; Van Hell & De Groot, 2008). The empirical findings have led to various psycholinguistic theories of cognate representation (Davis, Sánchez-Casas, García-Albea, Guasch, Molero & Ferré, 2010; Dijkstra, Miwa, Brummelhuis, Sappelli & Baayen, 2010). The stimulus materials used in psycholinguistic studies are usually collected manually from databases and then rated with respect to orthographic, semantic, and phonological properties before the experiment takes place by bilinguals from the population that is later tested (Caramazza & Brones, 1979; Dijkstra et al., 1999; Van Hell & Dijkstra, 2002). Bilinguals' ratings ensure that cognates and control words are properly matched with respect to the relevant processing dimensions.

Linguists are interested in cognates because they help in determining how and to what extent languages have changed over time. Cognates are used in many studies that compare cross-language similarity across languages (see Gray & Atkinson, 2003). Studies involving only small lists of cognates have already proved to be successful in the prediction of historical relations between language combinations. The limited numbers of items in such lists are often based on cognacy judgments by experts or experimental subjects. In contrast, we performed an innovative cross-language comparison of cognates derived from semi-complete lexicons of multiple languages.

In the present study, psycholinguistic similarity ratings and language family judgments are compared to cognate similarity measures that are produced by new tools from artificial intelligence applied to a translation database. These tools are based on computer schemes that enable cross-language analysis of translations in many different languages. The applied schemes provide a programmatic interface to a translation database, enabling a detailed interaction with the basic types of stored conceptual and orthographic information.

First, to identify cognates and their distributions across different languages, we must find word pairs that are translation equivalents, i.e., that share their conceptual structure across the language pair under consideration. In this first, conceptual, step, we can make use of the structure incorporated in the translation database, which ensures the correct retrieval of translations that are similar in meaning to the input word. The database provides one or more translation equivalents in the target language of the source word in the input language. In the first study to be reported below, the conceptual structure of the database is compared to experimentally acquired ratings with respect to SEMANTIC SIMILARITY from Tokowicz, Kroll, De Groot and Van Hell (2002). This will allow an evaluation of the applicability of formal techniques to automatically derive translation equivalents.

Second, to identify translation equivalents that are cognates, we must assess the cross-language ORTHOGRAPHIC SIMILARITY of each word pair. For this purpose, the so-called Levenshtein distance (discussed below) is computed to assess the orthographic similarity of collected translation equivalents (Yarkoni, Balota & Yap, 2008). In the second study below, norms derived by means of this formal metric are again compared to ratings from Tokowicz et al. (2002) and to those from Dijkstra et al. (2010).

Third, the resulting lists of identical or similar (cognate-like) translation pairs are useful for (psycho)linguistic research. They can be used to study the properties of the COGNATE DISTRIBUTIONS for particular language combinations. In the third study below, the identified cognates across language pairs are compared to language relatedness measures reported in the literature (in particular, the phylogenetic language family trees from Gray & Atkinson, 2003). At the end of the paper we evaluate the distribution of form-similar cognates in relation to similarities between languages as a whole.

Study 1: Measuring the semantic similarity of word pairs across languages

As a first step to identify cognates, we retrieved translation equivalents from a translation database (Euroglot professional 5.0) that contains 20,278 English, 19,519 German, 19,464 Dutch, 15,710 Spanish, 14,759 French, and 12,855 Italian 3–8 letter words. To verify that the conceptual structure of this database was consistent with that of language users, we automatically extracted translation equivalents across languages and assessed them using ratings of semantic similarity collected in psycholinguistic studies.

The conceptual structure of the database used distinguishes different readings of a word. When two readings of two words in different languages relate to a common language-independent concept, a translation is retrieved. In our study, we accepted only exact conceptual matches of translations and omitted other relations to the specific concept. For instance, in the translation database, the words *Mambo* and *Samba* have different relations to the same concept ("Latin dances"). Although the forms are similar, they were not classified by us as cognates, because their relationship to their shared concept was not identical.[1]

[1] Numbers of words, meanings of words, and word-to-meaning mappings differ across languages according to language-specific variations and the dictionary used. Because we did not want to incorporate these differences, we retrieved the set of most overlapping

https://doi.org/10.1017/S1366728910000623 Published online by Cambridge University Press

When we retrieved translation equivalents for the word pairs used by Tokowicz et al. (2002), these corresponded to a large extent (776 out of 951 word pairs, amounting to 81.5%) to those word pairs with the highest semantic similarity ratings (rated 5/7 or higher).[2] For example, the word pair *father – vader* received a rating of 5 out of 7 and was also retrieved using the conceptual structure of the database. The differences in ratings and retrieval of translations from a database could reflect that the word-to-concept mapping in laymen (as expressed in judgments) is more diffuse or less precise than that of experts on language (as implemented in the database). For instance, a translation pair like *gemeen – cruel* is absent in the database, because, according to experts, it does not share the exact same relation(s) to the shared concept. Other word pairs in the database, like *gemeen – mean* and *wreed – cruel*, are, in fact, better translation pairs than those obtained by layman judgments. On the whole, we conclude that the conceptual structure of the database can be used with confidence to classify translation pairs as potential cognates.

Study 2: Measuring the orthographic similarity of word pairs across languages

A second step in classifying translation pairs into cognates and non-cognates involves assessing the orthographic similarity of word pairs. This requires a valid similarity metric that can distinguish expressions with high orthographic overlap from expressions with low orthographic overlap, independently of word length. For instance, word pairs like *relative – relatief* and *idea – idee* should intuitively obtain a similar orthographic similarity score, because both pairs share 75% of their characters. The counterintuitive argument would be that the second pair shares 100% fewer different characters than the first, because the first pair differs by two characters and the second pair differs by one character. The metric should be formalized so that it can be applied in a computational simulation of cognacy judgments.

Resulting norms for orthographic overlap should correlate with ratings by bilingual language users. Cognates used for psycholinguistic experiments are traditionally rated by the experimenters themselves or are obtained via similarity rating studies. However, these methods cannot be formalized and are biased towards concrete expressions (Friel & Kennison, 2001). Furthermore, these collection methods are time-consuming, so they are not applicable in a simulation with a large set of word pairs. In Dijkstra et al. (2010), the use of continuous similarity norms was indispensable to assess the effects of variations in cross-language orthographic similarity on cognate effects in experimental studies. However, all stimuli in this study had to be retrieved by hand rather than automatically.

In information theory, two popular string metrics are available to evaluate strings on orthographic similarity: the Hamming distance and the Levenshtein distance. The Hamming distance (Hamming, 1950) counts the minimal number of substitutions needed to edit one string into another of equal length. The Levenshtein distance (Levenshtein, 1966) counts the minimal number of substitutions, insertions, and deletions to edit one string into another of any length. For word pairs that have equal word length, the Levenshtein distance produces only distances smaller or equal to the Hamming distance. A cognate pair like *guitar – gitaar* takes advantage of this property. In the Hamming distance, the characters ⟨u⟩, ⟨i⟩, and ⟨t⟩ of *guitar* are substituted for the corresponding three characters in *gitaar*, resulting in a distance of 3. If we assign a standard cost of 1 to each of the edit operations (insertion, deletion, and substitution), the resulting Levenshtein distance is 2 (one deletion of ⟨u⟩ and one insertion of ⟨a⟩). Standard edit operation costs ensure that the Levenshtein distance is a METRIC over the set of strings. A METRIC is mathematically defined as a special distance function that has particular properties: non-negativity, being zero only if strings are equal, symmetry, and triangle inequality.

The Levenshtein distance metric produces high values for long words and low values for short words. Because similarity needs to be comparable between word pairs of different lengths, we adjusted the Levenshtein distance as given in Equation 1.

$$score = 1 - \frac{distance}{length}$$

$$length = max(length\ of\ source\ expression,\ length\ of\ destination\ expression)$$

$$distance = min(number\ of\ insertions,\ deletions\ and\ substitutions)$$

Equation1. Levenshtein distance normalized for word length.

This division of the Levenshtein distance by the maximum length of both words provides a distance function for orthographic similarity that is relative to word length. The operation normalizes the distance metric at the same time; identical words return a Levenshtein distance of 0, resulting in a similarity score of 1, words with no overlap return a distance equal to the length of the longest word, resulting in a similarity score of 0. This normalized Levenshtein distance (NLD) simulates orthographic similarity between 0 and 1 in a

translations only. Incorporating only word pairs with identical relations in word-to-meaning mappings, ensured that only meaning equivalents across languages were retrieved.

[2] Incorporating non-identical relations between word-to-meaning mappings resulted in a retrieval rate of 91.0% of the semantically similar word pairs from Tokowicz et al. (2002).

https://doi.org/10.1017/S1366728910000623 Published online by Cambridge University Press

semi-continuous way. In simulations where only words of length from 3 to 8 are included, the distance function can take 23 different degrees of orthographic similarity.

Recent studies have made successful use of Levenshtein distance to simulate orthographic similarity (Heeringa, 2004; Kondrak & Sherif, 2006; Yarkoni et al., 2008). Yarkoni et al. computed the ORTHOGRAPHIC LEVENSHTEIN DISTANCE 20 (OLD20) for all words in a monolingual lexicon (including words of different lengths). OLD20 measures the average distance over the 20 closest neighbors according to the classic Levenshtein distance metric. The authors demonstrated that OLD20 was a significantly better predictor of both lexical decision and pronunciation performance in three large data sets than standard orthographic neighborhood density measures (a neighbor is a word that differs in just one letter position from a target word). There was a stronger interaction of the new measure with word frequency and stronger effects of neighborhood frequency as well. OLD20 neighborhood measures correlated strongly with word length (.71 for monosyllabic words, .87 for mono- and multimorphemic words), resulting in relatively high distance measures for long words.

In our study, we used the Levenshtein distance function for bilingual research purposes to approximate traditional measures of orthographic similarity, i.e., experimentally acquired ratings. However, we normalized the function in order to take summations of all cognates with differences in word length. The semi-continuous norms of the distance function were applied to the word pairs from different languages included in Tokowicz et al. (2002) and Dijkstra et al. (2010). These studies yielded similarity ratings on, respectively, 1003 and 318 word pairs. The ratings from the two studies correlated, respectively, .88 and .96 with the distance function.[3] We applied an inclusive threshold of .5 to the computed similarity measures to estimate the classification accuracy of the new distance function. Of the highly form-similar word pairs from the first study (rated 5/7 or higher), 183 out of 193 were correctly classified as cognates (91%). Of low form-similar word pairs (rated 4/7 or lower), 728 out of 735 were correctly classified as non-cognates (99%). All 78 high form-similar word pairs from the second study were correctly classified as cognates (100%). Here, 184 out of 205 low form-similar word pairs were correctly classified as non-cognates (90%).

Two false negatives were *cotton – katoen* (see Supplementary Materials Part 1) and *stone – steen* (see Supplementary Materials Part 2).[4] The ⟨c⟩ in the first word pair is pronounced as /k/, which is usually represented in Dutch by the letter ⟨k⟩. Because we apply the Levenshtein distance to orthographic input, the distance measures do not take into account grapheme-to-phoneme mappings and are therefore unable to approximate human similarity judgments in such cases. Secondly, in Dutch, a long vowel in a monomorphemic word is usually represented by two equal vowel letters (so *steen* is pronounced as /staɪn/, not /sten/). In English, a vowel is usually added later (*stone* is pronounced as /stəʊn/, not /stɒn/). Such regular phoneme-to-grapheme mapping rules are not incorporated in Levenshtein distance measures, while similarity judgments may incorporate these rules. It seems that phoneme-to-grapheme mappings are used by the test subjects as more evidence for orthographic similarity; the test subjects apparently incorporate sound evidence in orthographic similarity ratings.

To conclude, the orthographic similarity norms obtained with the normalized Levenshtein distance can be applied with confidence (correlations across independent studies of .88 and .96) to obtain reliable measures (classification rates between 91% and 100%) of orthographic similarity for given word pairs. This allows the determination of orthographic similarity for translation pairs in large databases in an automatized way that is much faster than traditional methods.

Study 3: Orthographic similarity distributions of cognates

By applying the proposed normalized Levenshtein similarity metric to translation equivalents in six languages, we can now identify distributions of cognates across language pairs in an automatized analysis. Only words with a word length between 3 and 8 characters were evaluated to limit computation time. Character case was ignored (e.g., for German nouns).

For a particular language combination, we computed the number of cognates of different word lengths. We used an exclusive threshold of .5 on the Levenshtein function, to exclude word pairs that share exactly half of their letters. A score of .5 can result from 19 combinations of word lengths, while a score of .571 can result from only seven combinations. The resulting numbers are represented in Table 1.[5] As can be seen in this table, the proportions of cognates (relative to extracted translation equivalents) in word pairs of a given length are not identical across length categories. Without normalization of the Levenshtein distance for words of different lengths, a larger number of cognates would have been found for shorter words. In contrast, as Table 1 shows, after normalization many

[3] The non-normalized Levenshtein distance measures correlated .73 with ratings from Tokowicz et al. (2002) and .93 with ratings from Dijkstra et al. (2010).

[4] Supplementary Materials are available online, on the Journal's website, accompanying the online version of the present article (see journals.cambridge.org/bil).

[5] The number of cognates between language combinations computed in this way did not correlate with the number of translations between language combinations ($r = .10$, ns).

https://doi.org/10.1017/S1366728910000623 Published online by Cambridge University Press

Table 1. *Numbers of translations (trans.) and cognates (cogn.) for language combinations involving Dutch, as computed for each possible minimal word length (in letters).*

	Dutch–English		Dutch–French		Dutch–German		Dutch–Italian		Dutch–Spanish	
Length	Trans.	Cogn.	Trans.	Cogn.	Trans.	Cogn.	Trans.	Cogn.	Trans.	Cogn.
2	3504	374	2104	217	2053	464	1635	152	1801	141
3	9078	1182	4834	601	4910	1359	4159	505	5018	458
4	8839	1660	6110	1099	5888	2237	5118	883	6386	896
5	8660	2099	7348	1612	7712	3378	6054	1365	7417	1297
6	7020	2060	6088	1685	6721	3430	5506	1601	6044	1477
7	3056	1062	2760	932	3527	1903	2574	907	2782	848

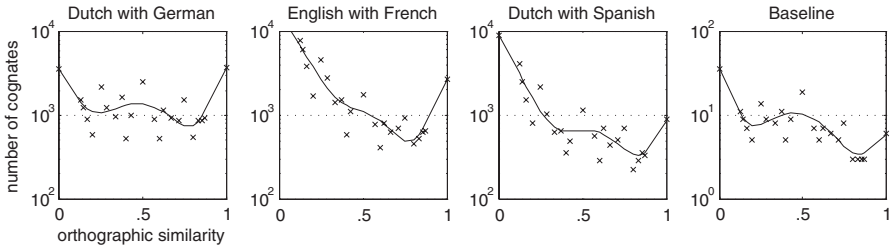

Figure 1. Differences in orthographic similarity distributions plotted on a logarithmic y-axis. The smooth lines are fitted to the data points by 4th degree polynomials.

cognates are identified in translation pairs that consist of long words. Further studies need to look into the nature of long and short cognate word pairs, for example, in terms of word frequency and borrowing history. The normalization of the Levenshtein distance is a necessary condition to obtain these insights.

Next, in order to compute the distributions of cognates, we summed the number of cognates of different word lengths for each language combination. Figures 1, 2, and 3 show the resulting distributions with different degrees of cross-language orthographic similarity.

The first three panels of Figure 1 give typical examples of polynomial fits to the orthographic similarity function. Higher curves for a certain orthographic similarity value indicate that more cognates were identified for that value. Identical cognates are displayed at the extreme right of each panel. The fourth panel provides a theoretical baseline orthographic similarity distribution, assuming that all word length combinations are equally likely. The baseline and obtained orthographic similarity distributions were fitted to the data by quartic approximation functions. The bump in the example distributions can be interpreted as a side effect of the number of possible word length combinations for word pairs from different languages, which is highest for a score of .5. As can be seen, the distribution of Dutch–German cognates is highest for all high similarity values.

Figure 2 shows that languages belonging to the same language family (e.g., Germanic or Romance languages) share more cognates than languages belonging to different families. The exceptions are combinations of English with French, Spanish, or Italian. This similarity of English to Romance languages derives to considerable extent from the consequences of the famous Norman-French victory at Hastings in 1066 by William the Conqueror. By loosing this battle, Harold became the last English-speaking king for nearly 300 years, and in the following centuries many Romance words were introduced into English.

In Figure 3, the comparison of English to Germanic and Romance languages is represented in two different panels. The relative height of the curve in the right panel indicates the strong relation of Romance languages to English. However, note that only the combination English–French has a number of IDENTICAL cognates (2,727) comparable to languages from the same family (average is 2,374), whereas English–Spanish (1,330) and English–Italian (1,389) have a number of identical cognates comparable to languages from different families (average is 1,495).

Figure 4 shows that a Principle Component Analysis (PCA) on the observed numbers of cognates between language pairs reveals at least two strong components. These two components together explained 61% of the variance in the data. Because all three Romance languages plus English are projected on component 1 and all three Germanic languages plus French are projected

https://doi.org/10.1017/S1366728910000623 Published online by Cambridge University Press

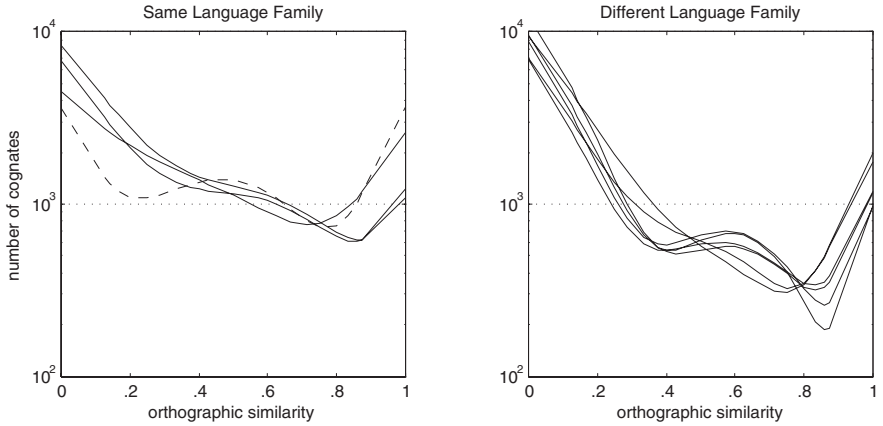

Figure 2. Orthographic similarity distributions plotted for similar (more related) and different (less related) languages separately. *Note:* Language pairs involving English are excluded (see Figure 3). The variability in the left panel for Same Language Family is due to the difference between similarity curves for Germanic (dashed) and Romance language families (see Figure 1 for comparison).

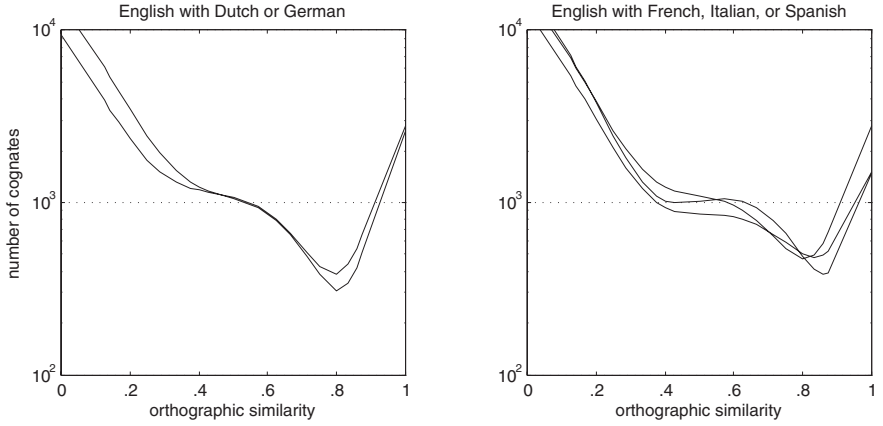

Figure 3. English has characteristics of both the Germanic and Romance language family.

on component 2, these components reveal the language groups of Germanic and Romance language families. There was an orthogonal projection of English to French in components 3 and 4, indicating the difference between English and French.

Considering the figures together, most languages share relatively many cognates with only one other language. Dutch shares most cognates with German, Italian shares most cognates with Spanish, English shares most cognates with French, and vice versa. In addition, a particular Romance or Germanic language generally shows a high overlap within the whole Romance or Germanic language family cluster. Interestingly, and in contrast to the general case, only English shares many cognates with all other languages.

These points are clarified by the patterns of cognate numbers in Tables 2–4. Tables 2 and 3 show, respectively, the proportion of all cognates or only identical cognates that a particular language shares with another language, relative to its language pair with maximum number of

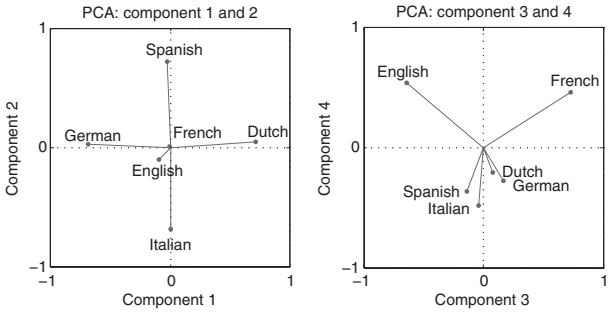

Figure 4. Principal Component Analysis (PCA) of number of cognates across language combinations.

cognates. For instance, English shares a maximum of 9,286 cognates with French and 8,609 cognates with Dutch. Thus, Dutch has 93% of cognates in common with English relative to French and English. There is one clear exception to the similarity relationships between and within languages from the same and different language

https://doi.org/10.1017/S1366728910000623 Published online by Cambridge University Press

Table 2. *Cognates in different language combinations represented as proportions of the maximum numbers of cognates.*

	English	French	Dutch	German	Italian	Spanish
English	0	1	.93	.83	.80	.84
French	1	0	.68	.62	.96	.98
Dutch	.67	.45	0	1	.43	.41
German	.60	.44	1	0	.40	.37
Italian	.64	.77	.48	.45	0	1
Spanish	.68	.79	.46	.41	1	0

Table 3. *Identical cognates in different language combinations represented as proportions of the maximum numbers of identical cognates.*

	English	French	Dutch	German	Italian	Spanish
English	0	.95	1	.90	.49	.46
French	1	0	.76	.67	.45	.40
Dutch	.76	.55	0	1	.33	.23
German	.68	.49	1	0	.29	.23
Italian	.51	.46	.46	.41	0	1
Spanish	.49	.40	.33	.32	1	0

family. At least 80% of English–French cognates is always present for the comparison of any language with English. In contrast, French shows 62%, while other languages show only 41%, 40%, 45%, and 41% as the minimum percentage of shared cognates expressed in terms of their maximum similarity pair. The low proportions of cognates shared between less related languages did not appear in the proportion of cognates relative to English.

We hypothesized that numbers of automatically identified cognates from a translation database can predict language distances as observed in phylogenetic studies. We therefore compared the computed similarity orderings to a language similarity tree from Gray and Atkinson (2003). As presented in Table 4, this language similarity ordering is based on a language tree constructed to predict divergence times in the evolution of languages. The branch lengths of this tree are proportional to maximum likelihood estimates of language relatedness. Relatedness was estimated by Gray and Atkinson using a database with 2,449 cognates across 87 languages, applying prior models of lexical evolution based on detailed constraints on language grouping. The two accounts were largely consistent ($r = .72$, $p < .001$, for 13 comparisons). This automatic identification of form-similar cognates improved consistency with Gray and Atkinson as compared to identification of identical cognates alone (leading to $r = .48$, $p < .1$, for 13 comparisons). In general, differences between our account and that by Gray and Atkinson appear to be due to the similarity of English to Romance languages.

In all, the computed language similarity measures correspond quite well with measures from phylogenetic studies. In contrast to those validation studies, the computational ordering is based on semi-complete lexicons, which may be of theoretical and practical relevance to linguists interested in cross-language similarity and diversity. Like the study by Gray and Atkinson, our computational study enhances the theoretical insight into cross-language orthographic similarity. In addition, the normalized Levenshtein distance leads to the identification of form-similar cognates that can be used as stimulus materials in linguistic and psycholinguistic experiments.

Table 4. *Language pair similarity orderings based on total number of cognates.*

	Total number of cognates	Form-identical cognates only	Language evolution
Dutch–German	12908	3785	20
Italian–Spanish	11485	2698	26
English–French	9286	2727	204
French–Spanish	9120	1091	34
French–Italian	8871	1232	26
Dutch–English	8609	2865	42
English–Spanish	7837	1330	204
English–German	7750	2576	36
English–Italian	7430	1389	184
Dutch–French	6269	2063	200
French–German	5725	1850	194
Dutch–Italian	5564	1232	180
Dutch–Spanish	5298	889	200
German–Italian	5187	1108	174
German–Spanish	4794	869	194

Discussion

In our study, we approximated the orthographic similarity of translation pairs from different languages by applying a normalized version of the popular Levenshtein distance metric. The distance function was applied to a translation database in order to study the distribution of cognates in six European languages. In the following paragraphs, we relate the orthographic similarity distributions of cognates to other, more general orthographic, phonological, and semantic aspects of translation equivalents. We consider in some detail orthographic similarity in relation to spelling differences across scripts from different languages, number of meaning mappings, and word frequency characteristics.

https://doi.org/10.1017/S1366728910000623 Published online by Cambridge University Press

Orthographic overlap

Cognate distributions over an orthographic similarity continuum showed a common pattern that differed only slightly across language combinations. It indicated the presence of a large number of cognates appearing as form-similar translation pairs besides identical 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 historical development and writing systems, they kept an identical alphabetic form or underwent changes in spelling or capitalization. Language combinations for speakers that are geographically more distant may have fewer cognates because the chances of mutual influence are smaller; language combinations with relatively many form-similar cognates may have changed more than languages with relatively many form-identical cognates. As a consequence, the relative number of form-similar and identical cognates might be correlated with divergence time in language family trees. In the present study, French–Italian and French–Spanish (two Romance language pairs) attained high numbers of cognates because form-similar cognates were incorporated. According to the corresponding relative cognate proportions in Tables 3 and 4, similarity based on form-identical cognates is indeed less salient within Romance languages and English. Similarity within Germanic languages depends to a larger extent on identical cognates than within Romance languages.

Orthographic–phonological mappings

We have defined cross-language similarity in terms of orthographic overlap between cognates. Another similarity component of cognates is phonetic overlap, which is a primary clue for their identification (Mackay & Kondrak, 2005). The presented orthographic similarity measures may reflect phonetic similarity according to phoneme-to-grapheme mappings, which differ across languages. For example, the mapping rules between phonemes and graphemes are more complex for English than for Italian. Said differently, English is considered to have a "deeper" orthography than Italian. It is therefore important that future research considers to what extent cognate similarity measures could benefit from considering spelling to sound regularities. The resulting differences in cross-language similarity with respect to orthography and phonetics may be informative with respect to language history and spelling reforms. The feasibility of this possibility has been proven by researchers like Gooskens and Heeringa (2004), who investigated the use of Levenshtein distance to measure the psychoacoustic perceptual distance between speakers of 15 different Norwegian dialects.

Orthographic change

To study cognate distributions, we examined the orthographic similarity of translation equivalents across language combinations. However, there are also word pairs with orthographic similarity that do not share their meaning across languages. For example, the Dutch–German false friend *knap – knapp* means "wise" or "pretty" in Dutch but "tight" in German. To identify distributions of such FALSE FRIENDS with respect to orthographic similarity might be of additional interest for examining changes in orthographic similarity over time. Because false friends may result from coincidental form-overlap within the orthotactic and phonotactic systems of the languages concerned, such distributions may be informative with respect to the spelling systems on which the observed distributions of cognates depend as well.

Possibly, differences between numbers of false friends across language pairs can be explained by such systematic spelling differences. In the present study, we showed that English has characteristics of both Romance and Germanic languages. We might therefore expect more coincidental form-overlap between language pairs with English than between language pairs that belong to either the Romance or Germanic family. Preliminary results confirmed the expectation that English shares a relatively high number of identical false friends with French (644), and with both German and Dutch (522). Orthographic similarity distributions of false friends across languages might be even more informative with respect to this issue.

Number of mappings

The cognate distributions that were identified here for various language combinations successfully predicted characteristics of language relatedness. Semantic overlap between word pairs was assumed to be present when at least one identical semantic mapping existed in the semantic structure of the translation database. We assumed that these semantic mappings would be associated with high semantic similarity ratings.

However, it might be the case that some language combinations are characterized by a larger number of semantic mappings across cognates than others, due to cross-language differences in concept equivalence (compare English–Dutch to English–Chinese). In addition, the number of mappings may not be symmetrical across the language pairs considered (see Tokowicz et al., 2002). As a consequence, the cross-language similarity patterns captured by our cognates will sometimes be less refined than in reality. We are currently investigating to what extent identification of cognates is sensitive to number of mappings and how it interacts with word frequency.

Word frequency characteristics

In the present study, differences between the computed language similarity orderings and the reviewed phylogenetic studies could be accounted for in part by differences in the way cross-language similarity is determined. Whereas our computational study included semi-complete lexicons, linguistic experts and the average language user might base their judgments on an implicit set of most frequently used words. For instance, in Gray and Atkinson (2003), only high frequency words were used to estimate divergence times.

According to Pagel, Atkinson and Meade (2007), word frequency accounts for 50% of lexical replacement over time (i.e., divergence of characters over time between translation pairs). Therefore, a language similarity ordering based on frequently used words (diverging at slower rates than infrequently used words) might reflect more of the shared origins in the analyzed languages. Because our language corpora consist of semi-complete lexicons, they also contain infrequently used words that have a fast lexical replacement rate. We are currently investigating to what extent low frequency words have influenced the language similarities we observed, e.g., the observed multiple family characteristics of English.

Conclusion

It is possible to automatically identify large distributions of cognates with respect to form-similarity in various European languages by means of a formalized form-similarity metric such as normalized Levenshtein distance. Applying this metric to a professional translation database, similarity norms were obtained that are comparable to experimentally acquired orthographic similarity ratings (Dijkstra et al., 2010; Tokowicz et al., 2002), and lead to high correlations (around .90) and a large proportion of correctly classified stimuli (over 90%). The obtained distributions were also compared to an account of cross-language similarity based on Gray and Atkinson ($r = .72$). A common pattern in the degree of orthographic similarity of these distributions was observed within languages of the same family. In our analysis, English showed characteristics of multiple language families (Germanic, Romance). Cognate distributions were computed here using semi-complete lexicons, whereas Gray and Atkinson used only a small set of high frequency words.

In all, our study demonstrated the feasibility and advantages of applying techniques from artificial intelligence to psycholinguistic and linguistic research involving multiple languages. First, the application of the normalized Levenshtein distance function resulted in an automatized selection of more and better stimulus materials for cognate studies on bilingual word processing. Second, the Levenshtein distance function yielded accurate and detailed cross-language similarity distributions for multiple languages, thus allowing a comparison to language family trees. As such, the present study has shown that the Levenshtein distance function can compete with existing similarity measures (such as those proposed by Coltheart, Davelaar, Jonasson & Besner, 1977, and Van Orden, 1987) and can be considered as a new formal and computational model of orthographic similarity, useful for future empirical studies in monolingual and bilingual domains as diverse as those dealing with neighborhood effects, spelling systems, and dyslexia.

References

Caramazza, A., & Brones, I. (1979). Lexical access in bilinguals. *Bulletin of the Psychonomic Society, 13 (4),* 212–214.

Ceram, C. W. (1966). *Enge Schlucht und Schwarzer Berg: Entdeckung des Hethiter-Reiches*. Reinbek: Rororo Taschenbuch Ausgabe.

Costa, A., Caramazza, A., & Sebastián-Gallés, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26 (5),* 1283–1296.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (ed.), *Attention and performance VI*, pp. 535–555. Hillsdale, NJ: Erlbaum.

Davis, Ch., Sánchez-Casas, R., García-Albea, J., Guasch, M., Molero, M., & Ferré, P. (2010). Masked translation priming: Varying language experience and word type with Spanish–English bilinguals. *Bilingualism: Language and Cognition, 13,* 137–155.

Dijkstra, A. (2005). Bilingual visual word recognition and lexical access. In J. F. Kroll & A. De Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*, pp. 178–201. Oxford: Oxford University Press.

Dijkstra, A., Grainger, J., & Van Heuven, W. J. B. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language, 41,* 496–518.

Dijkstra, A., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language, 62,* 284–301.

Duyck, W., Van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for nonselective access. *Journal of Experimental Psychology: Learning, Memory and Cognition, 33 (4),* 663–679.

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.

Hamming, R. W. (1950). Error detecting and correcting codes. *The Bell System Technical Journal, 22 (2),* 147–160.

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

Hrozný, F. (1915). Die Lösung des hethitischen Problems. Ein vorlaüfiger Bericht. *Mitteilungen der Deutschen Orient Gesellschaft, 56,* 17–50.

Kondrak, G., & Sherif, T. (2006). Evaluation of several phonetic similarity algorithms on the task of cognate identification. *Proceedings of the Workshop on Linguistic Distances Sydney: Association of Computational Linguistics,* pp. 43–50. [ACL Anthology Network, archive at http://aclweb.org/anthology-new/.]

Kroll, J. F., Stewart, E. (1994). Category interference in translation and picture naming – Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33 (2),* 149–147.

Lemhöfer, K., Dijkstra, T., & Michel, M. C. (2004). Three languages, one ECHO: Cognate effects in trilingual word recognition. *Language and Cognitive Processes, 19 (5),* 585–611.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10 (8*), 707–710. [Russian original (1965) in *Doklady Akademii Nauk SSSR, 163 (4),* 845–848.]

Mackay, W., & Kondrak, G. (2005). Computing word similarity and identifying cognates with pair hidden Markov models. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL),* pp. 40–47. Ann Arbor, MI. [ACL Anthology Network, archive at http://acl.ldc.upenn.edu/W/W05.]

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing. *Bilingualism: Language and Cognition, 6 (2),* 97–115.

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.

Schwartz, A. I., & Kroll, J. F. (2006). Bilingual lexical activation in sentence context. J*ournal of Memory and Language, 55,* 197–212.

Tokowicz, N., Kroll, J. F., De Groot, A. M. B., & Van Hell, J. G. (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.

Van Hell, J. G., & De Groot, A. M. B. (1998). Disentangling context availability and concreteness in lexical decision and word translation. *Quarterly Journal of Experimental Psychology Section a – Human Experimental Psychology, 51 (1),* 41–63.

Van Hell, J. G., & De Groot, A. M. B. (2008). Sentence context modulates visual word recognition and translation in bilinguals. *Acta Psychologica, 128 (3),* 431–451.

Van Hell, J. G., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychonomic Bulletin & Review, 9 (4),* 780–789.

Van Orden, G. C. (1987). A rows is a rose. *Memory & Cognition, 15,* 181–198.

Voga, M., & Grainger, J. (2007). Cognate status and cross-script translation priming. *Memory & Cognition, 35 (5),* 938–952.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15 (5),* 971–979.