For the final version, see:
Vanhove, Jan & Raphael Berthele. 2015. Item-related determinants of cognate guessing in multilinguals. In Gessica De Angelis, Ulrike Jessner & Marijan Kresić (Eds.), *Crosslinguistic influence and crosslinguistic interaction in multilingual language learning*, 95-118. London: Bloomsbury.

Item-related determinants of cognate guessing in multilinguals

Jan Vanhove♠ and Raphael Berthele

Supplementary materials (data, computer code and questionnaires) available from
http://dx.doi.org/10.6084/m9.figshare.763246

1 Introduction

Formal similarities to a known language make it considerably easier to learn a new one. More specifically, the learning curve is smoothed if language learners are able to make use of the cross-linguistic similarities provided by cognate pairs (e.g. Carton, 1971; De Groot & Keijzer, 2000; Lotto & De Groot, 1998; Odlin, 1989; Ringbom, 2007). These are etymologically related translation equivalents that often bear a certain degree of formal similarity, e.g. Dutch *appel* and English *apple* or Swedish *försvinna* and German *verschwinden* 'to disappear'. It is especially in comprehension that cognate relations, if perceived, can bring about positive interlingual transfer (Ringbom, 2007). In fact, formal similarities between related languages can be so pervasive that they give rise to *receptive multilingualism* (Braunmüller & Zeevaert, 2001), a constellation in which readers or listeners are able to understand a language variety (L*x*) without having 'officially' learnt or acquired it. Additionally, a series of learning materials aiming to foster receptive skills in related languages rely on the presence of such cross-linguistic similarities (e.g. Hufeisen & Marx, 2007; Klein & Stegmann, 2000; Müller et al., 2009).

For cross-linguistic cognate relationships to be of actual help, it is of course essential that foreign language learners or readers and listeners engaging in receptive multilingualism be aware of them (see Otwinowska-Kasztelanic, 2011, for a concise discussion)—a condition which is not always met automatically. In order to investigate the factors that affect how well participants can identify cross-linguistic cognate relationships, several studies have made use of what we will call *cognate guessing tasks*. These are purposefully reductionistic tasks in which participants are presented with isolated L*x* stimuli with cognates in known languages and are asked to translate them. Since the role of con- and cotextual cues is eliminated, participants need to engage in *interlingual inferencing* (Carton, 1971) and draw on their linguistic and meta-linguistic knowledge in order to guess the meaning of the L*x* stimuli. Cognate guessing tasks consequently yield insights into the factors that determine the probability with

♠ Corresponding author: Jan Vanhove, Department of Multilingualism, University of Fribourg, Rue de Rome 1, CH-1700 Fribourg, Switzerland, jan.vanhove@unifr.ch.

which interlingual transfer and inferencing processes produce the correct outcome (i.e. a correctly identified cognate relationship) when the roles of target language and extralinguistic knowledge are reduced to a minimum.[1]

Such factors fall into two broad categories. On the one hand, the accurate recognition of cognate relationships in an L*x* depends on participant-related variables. Of particular interest for our present purposes, multilinguals often draw on their knowledge of several languages when confronted with a text in an unknown language (e.g. Gibson & Hufeisen, 2003; Singleton & Little, 1984; Wenzel, 2007). With respect to cognate guessing tasks specifically, a large multilingual repertoire is often associated with better task performance, particularly if it includes well-developed competences in language varieties that are related to the L*x* (Berthele, 2008, 2011; Berthele & Lambelet, 2009; Swarte, Schüppert & Gooskens, 2013; but see Van Bezooijen, Gooskens & Kürschner, 2012, for a study failing to find such a multilingualism effect). On the other hand, the likelihood with which cognate relationships can be identified is also affected by item-related variables, e.g. the formal similarity between an L*x* word and a cognate in a known language.

However, the participants' knowledge of foreign languages is not systematically taken into account when modelling cognate guessing success in terms of item-related variables: typically, these measures are computed with respect to the L1 alone. This state of affairs can presumably partly be ascribed to two facts. First, the number of variables quickly becomes too large when several predictors have to be considered with respect to several potential supplier languages (the 'small *n*, large *p*' problem). Second, these variables are often characterised by properties that are unfavourable for statistical analyses, particularly by strongly skewed distributions and a high degree of collinearity (see Fox & Weisberg, 2011, Ch. 6). Researchers consequently need to take rather arbitrary decisions about which variables to include with respect to which potential supplier languages so that the set of predictors remain of a manageable size. Relatively recent developments in statistics, however, are better suited to cope with multidimensional datasets characterised by skewed and collinear variables and offer new opportunities to analyse such datasets in a principled way.

In this paper, we make use of one such recent development, *random forests* (Breiman, 2001), in order to (re)analyse data from two independent but similar studies in which multilinguals took part in written cognate guessing tasks. Our aim is to assess the impact of a selection of item-related characteristics on the likelihood with which the meaning of L*x* cognates can be guessed correctly. For both studies, German-speaking Swiss participants (L1s: Swiss German dialect and Standard German) with some knowledge of at least French and English (typically the first and second foreign language,

1 The word *cognates* in a paper on multilinguals tends to evoke associations with studies about lexical access processes in bi- and multilinguals that exploit the properties of cognates (e.g. Costa, Caramazza & Sebastián-Gallés, 2000; Dijkstra & Van Heuven, 2002). Such studies are concerned with the automatic activation of lexical items in the bi- or multilingual brain and make use of participants with extensive prior knowledge in the target language. In cognate guessing tasks, by contrast, participants have no prior knowledge of the target language, and whatever automatic lexical activation effects the stimuli may bring about are often buried under a thick layer of decision processes.

respectively[2]) were asked to translate written stimuli from other Germanic languages (Danish, Dutch, Frisian and Swedish). Many of these stimuli have translation-equivalent cognates in French and English, warranting the inclusion of item-related variables with respect to these languages in the analysis. On the basis of the literature (see Berthele, 2011; Van Heuven, 2008; Vanhove, 2014; see Section 2 for further references), we identify four types of such variables whose impact on cognate guessing success can be investigated in the present paper. We then operationalise each type of variable with respect to Standard German, French and English.[3] Rather than assuming that cognate guessing success can most parsimoniously be modelled by including variables with respect to one, two or all three of these common languages, this decision is taken in an exploratory, data-driven way. By analysing data from two similar but independent studies using different participants and different stimuli, we nevertheless ensure that our exploratory findings are empirically robust. In the spirit of 'open science', the data and computer code used for the analyses have been made available at http://dx.doi.org/10.6084/m9.figshare.763246.

2 Explanatory variables considered

2.1 Overall formal distance

In order to make an 'interlingual identification' (Weinreich, 1953), participants need to perceive the cross-linguistic similarity between the stimulus and the potential transfer base. While the perception of cross-linguistic similarity is not a function of objective formal similarity alone (Kellerman, 1977, 1983), it is generally speaking more likely to occur when the formal overlap between the stimulus and the potential transfer base is high. The degree of such formal overlap is consequently a key variable in receptive multilingualism.

In receptive multilingualism studies, the formal overlap between an L*x* word and one of its cognates in a known language is typically measured by means of the Levenshtein algorithm (see examples below). This algorithm is used to compute the smallest number of deletions, insertions and substitutions necessary to transform one string into another. The Levenshtein distance is defined as the total minimal operation cost; higher Levenshtein distances indicate less formal overlap. An example is given in Figure 1, which shows how the Dutch string *oorzaak* 'cause' can be transformed into its German cognate *Ursache* with a minimal total transformation cost of 7. These raw Levenshtein distances are then normalised in order to account for differences in word length between cognate pairs, typically by dividing them by the length of the longest possible least-cost alignment (as recommended by Heeringa, 2004: 130–132). For the example in Figure 1, this yields a length-normalised Levenshtein

2 Traditionally, French has been the first language beside Standard German taught in schools in German-speaking Switzerland. Since 2006, English has been introduced as the first foreign language in the eastern cantons.

3 While dialect speakers seem to be able to rely on their knowledge of dialects as well as of the standard language in receptive multilingualism (Berthele, 2008; Gooskens et al., 2011), there is no agreed upon orthography for the various Swiss-German dialects. It is therefore not clear how the variables with respect to the stimuli's dialectal cognates should be computed.

distance of 7 ÷ 9 = 0.78. In what follows, ‘Levenshtein distance’ refers to the length-normalised distance measure.

1	2	3	4	5	6	7	8	9	
o	o	r	z	a	a	k			
	u	r	s	a		c	h	e	
D	S		S		D	S	I	I	= 7

Figure 1. Example of a Levenshtein distance computation. (D: deletion, I: insertion, S: substitution)

Levenshtein distances between L*x* words and their cognates in a related L1 have been found to be respectable predictors of cognate guessing accuracy in the spoken modality (e.g. Gooskens, Kürschner & Van Bezooijen, 2011; Kürschner, Gooskens & Van Bezooijen, 2008; Van Bezooijen & Gooskens, 2005; but see Berthele, 2011). As for the written modality, Berthele and Lambelet (2009) administered a cognate guessing task featuring 29 isolated Romansh and Romanian words to 140 French- and Italian-speaking Swiss students. They found that the number of correct translation attempts per word was correlated with the orthographic Levenshtein distance between the L*x* word in question and its French or Italian cognate ($r = -0.32$). In a similar task involving 28 isolated Danish and Swedish words presented to 163 Swiss German participants, Berthele (2011) failed to find a significantly negative association between the Levenshtein distances between the Scandinavian words and their standard German cognates ($r = 0.14$). However, such a correlation was present when the Levenshtein distances with respect to English, the participants’ L2 or L3, were considered ($r = -0.42$). Lastly, Vanhove and Berthele (2013) presented 181 Germanic (Danish, Dutch, Frisian and Swedish) words to 98 Swiss German participants. They found that while Levenshtein distances between the L*x* words and their L1 (German) counterparts were negatively associated with the stimuli’s intelligibility, the fit of their model was markedly better when they also took into consideration the participants’ L2 and L3 (French and English). They did this by computing Levenshtein distances with respect to French and English as well and taking as the predictor variable whichever of three distances computed (L*x*–German, L*x*–French, L*x*–English) was the lowest for each stimulus. Together, Berthele’s (2011) and Vanhove and Berthele’s (2013) findings indicate that even when the participants’ L1 is closely related to the L*x*, L*x* cognate guessing may be modelled more accurately when taking into account the possibility that the participants make use of multiple supplier languages (for the notion of supplier language, see Williams & Hammarberg, 1998).

In addition to orthographic formal similarity, cognate guessing success in the written modality may also be affected by the participants' conception of the cognates' spoken form. Several authors (e.g. Berthele, 2011; Möller & Zeevaert, 2010; Wenzel, 2007) noted that participants tend to self-pronounce written L*x* stimuli on the basis of familiar or assumed grapheme–phoneme correspondences in their search of a plausible translation. Evidently, such self-pronunciations do not necessarily correspond to the stimuli's actual target language pronunciation. This allows for a substantial degree of inter- and intra-individual variability: not all participants will self-pronounce a given stimulus in the same way and one participant may consider several plausible articulations. As a result, the effect of 'assumed phonetic distance' was deemed to be too complex to model adequately in our quantitative analyses.

2.2 The importance of consonants

In its crudest form, the Levenshtein algorithm weighs all operations equally, i.e. a vowel–vowel substitution has the same operation weight as a consonant–consonant substitution. However, a few studies suggest that L*x* intelligibility is more severely affected by consonantal differences than by vocalic differences between the L*x* and the L1, L2, …, L*n*. Gooskens, Heeringa & Beijering (2008) computed Levenshtein distances between a spoken standard Danish text and its renderings in 17 other Scandinavian language varieties. They then computed the share of these distances that was due to vowel insertions, deletions and substitutions (vocalic Levenshtein distance) and to consonant operations (consonantal Levenshtein distance). In the example in Figure 1, four operations are associated with vowel slots and three with consonant slots; the length-normalised vocalic and consonantal Levenshtein distances would be $4 \div 9 = 0.44$ and $3 \div 9 = 0.33$, respectively. Gooskens et al. (2008) found that the intelligibility of the 17 Scandinavian language varieties to speakers of standard Danish correlated more strongly with the consonantal than with the vocalic Levenshtein distances ($r = -0.74$ *vs* $r = -0.29$), although the correlation with the overall phonetic Levenshtein distances was stronger still ($r = -0.86$). Similarly, Berthele (2011) found consonantal differences vis-à-vis German and English to be more detrimental to the intelligibility of Danish and Swedish targets to Swiss Germans compared to vocalic contrasts for both spoken and written stimuli. These results corroborate qualitative findings by Möller (2011), who reported that German-speaking students show greater tolerance towards vocalic than towards consonantal differences in a written cognate guessing task. A greater degree of flexibility vis-à-vis vocalic discrepancies has also been reported in so-called word reconstruction tasks in the L1 (Moates & Marks, 2012).

The implication is that consonants may contribute more to the subjective transparency of cognate relationships than do vowels, possibly due to their greater information value (see Van Heuven, 2008: 53). This paper provides an additional quantitative evaluation of whether consonants should indeed be weighed more heavily in models of written cognate guessing. It does so by considering the consonantal Levenshtein distances between the L*x* stimuli and their cognates in the three potential shared supplier languages. Our logic is that, if our participants are indeed particularly sensitive to the consonant

frames, then these consonantal Levenshtein distances should offer explanatory power over and beyond the full orthographic Levenshtein distances also considered in the model.

2.3 The importance of word beginnings

Participants may not only let guide their cognate guesses more by the consonantal skeleton of the L*x* stimulus than by its vowels; they may also be particularly sensitive to word beginnings in cognate guessing tasks (Berthele, 2011; Möller, 2011; Möller & Zeevaert, 2010; see also Müller-Lancé, 2003; for L1-related findings pointing to a privileged role of word beginnings, see Broerse & Zwaan, 1966; Johnson & Eisler, 2012; Scaltritti & Balota, 2013). Here, too, the implication is that certain word parts contribute more to the perceived similarity of cognate relationships than do others. In this paper, we therefore investigate whether word-initial discrepancies between the stimuli and their cognates can help to account for between-item differences in cognate guessing success over and beyond what can be explained by the full formal distances between the stimuli and their cognates. To this end, we will consider the word-initial distances between the L*x* stimuli an their cognates in all of the three potential supplier languages. Here, too, our logic is that a higher reliance on word beginnings on the part of our participants should result in explanatory power of these word-initial distances over and beyond what is offered by the full orthographic Levenshtein distances.

2.4 Cognate frequency

The recognition of words in known languages is affected by their corpus frequencies, such that high-frequency words are generally easier to recognise than low-frequency words (e.g. Lemhöfer et al., 2008, and references in e.g. Brysbaert et al., 2011). Van Heuven (2008) suggests that L*x* stimuli with high-frequency known cognates may similarly be easier to decode correctly. The impact of cognate frequency on auditory cognate guessing was investigated empirically by Kürschner et al. (2008) and was found to be negligible. Vanhove and Berthele (2013), by contrast, not only found that L1 cognate frequency was positively associated with written cognate guessing task performance but also that a joint measure comprising the aggregate frequency of the stimuli's German, French and English cognates was a better predictor still. This may indicate that multilingual participants are sensitive to frequency information from their different languages when participating in a written cognate guessing task.

3 Method

In order to determine the set of item-related variables that can most parsimoniously model written cognate guessing success in multilinguals, we adopt a highly exploratory approach, as we discuss below. As a guard against spurious findings, we turn to two independent but similar datasets on cognate guessing tasks (Dataset 1: Vanhove & Berthele, 2013; Dataset 2: Vanhove & Berthele, 2015) and reanalyse them separately from one another but along the same lines. The datasets feature different

stimuli and participants and the cognate guessing tasks were administered in a different way. Crucially, however, we assume that the effects of the four variables outlined in Section 2, if at all present, are comparable in the two datasets, both in terms of their direction (positive or negative) and effect sizes. Consequently, if a particular effect is not found in *both* datasets, we will have to conclude that it is not robust with respect to the stimuli chosen or the participants recruited and has to be left out of the final models on the grounds of parsimony.

We first discuss the specifics of both datasets. Then, we explain how we operationalised the variables discussed in the previous section. Lastly, we briefly present the statistical tools used for our analyses.

3.1 Description of the datasets

3.1.1 Dataset 1

The first dataset was collected for a study by Vanhove and Berthele (2013) that aimed to assess to what extent findings regarding cognate guessing are dependent on the L*x* in question.

Participants. Ninety-eight eligible participants were recruited to participate in a cognate guessing task.[4] All were native speakers of a Swiss German dialect, did not have knowledge of any Germanic language apart from Swiss German dialects, Standard German and English, and were not language experts (such as linguistics students or interpreters). Before the cognate guessing task, participants filled out a short questionnaire and rated their reading and listening skills in English, French and any other languages according to the six-level assessment grid for the Common European Framework of Reference for Languages (CEFR)[5]. We refer to Table 1 for the key descriptive statistics. Apart from Swiss-German dialects, Standard German, English and French, some participants reported knowledge of Italian (44), Spanish (38), Turkish (5), Portuguese, Russian (4), Greek (2), Albanian, Arabic, Bulgarian, Cantonese, Kurdish, Korean, Macedonian, Polish, Romansh and Czech (1).

4 From a larger sample of 107 participants, nine were excluded on the grounds that their English skills were not good enough to match any of the six CEFR descriptions ($n = 3$) or because they had not been tested under direct supervision ($n = 6$).

5 Available from http://europass.cedefop.europa.eu/de/resources/european-language-levels-cefr.

Table 1. Description of the two studies.

	Dataset 1	Dataset 2
n participants (of which women)	98 (65)	148 (85)
Age range	16–72	10–86
Mean age (SD)	34 (15)	42 (21)
Mean self-assessed English reading skills (SD)	4.0 (1.2)	3.5 (1.3)
Mean self-assessed French reading skills (SD)	3.3 (1.3)	3.2 (1.4)
n items	180	45
Lx	Danish, Dutch, Frisian, Swedish	Swedish

Note. Self-assessments were based on the CEFR grid (A1 = 1, C2 = 6).

Task. The cognate guessing task was a paper-and-pencil task and consisted of four lists featuring 50 words without context in a Germanic language (Danish, Frisian, Dutch and Swedish) each. 181 of these words had a German, English or French translation-equivalent cognate, which could in principle render them intelligible to readers without prior competences in the target language. These 'target words' were selected from a list with 384 frequent words compiled by the research group *Mutual intelligibility of closely related languages* (University of Groningen, PI: Charlotte Gooskens; see e.g. Gooskens et al., 2011; Kürschner et al., 2008). For details about the selection procedure, we refer to the studies cited. For the present analyses, one target word (*eftermiddag* 'afternoon') was left out of consideration as it is not a full cognate to either English (*afternoon*) or German (*Nachmittag*), but rather a mixture form of both. This left 180 target words for the analyses. The vast majority of these words had a German cognate (175), about half had an English cognate (84) and about a third had a French cognate (57).

Four to five words per list could not be translated without some pre-existing knowledge of the target language, e.g. Frisian *heit* 'father' and Danish *seng* 'bed'. Participants able to correctly translate more than two such words per language would have been assumed to have had too much pre-existing lexical knowledge of the language in question for our purposes and would have been excluded from the analyses (see Kürschner et al., 2008, for a similar filtering function of non-cognates). None of the participants was able to do so, however. For the full list of stimuli and their German, English and French translation-equivalent cognates, we refer to Appendix A.

We informed the participants that all 200 words were nouns and asked them to translate these into German. They were not forced to provide an answer in case none came to mind. The order of target languages was counter-balanced between participants (Danish–Dutch–Swedish–Frisian and the reverse) in order to prevent learning or fatigue effects from skewing the results. Each sheet clearly

indicated what language the stimuli were in.

Every translation was marked as correct or incorrect. Note that 'reasonable' but incorrect translations were scored as incorrect, too. Thus, a participant's attempt to translate the Danish word *kylling* ('chicken') as 'murder' (from 'killing') was rated as incorrect.

3.1.2 Dataset 2

The second dataset was collected for a subproject of the *Multilingualism through the lifespan* project (SNSF-130457). The goal of the subproject on cognate guessing was to track the lifespan trajectories of cognate guessing skills and to establish which cognitive and linguistic developments any such age trends could be attributed to (see Vanhove & Berthele, 2015; Vanhove, 2014). To this end, a total of 167 Swiss-Germans were recruited. Four of them experienced technical difficulties during the cognate guessing task; their data will not be analysed.

Participants. All participants self-rated their English, French and other language skills on the six-level CEFR scale. None of the participants reported any knowledge of Swedish nor of related North Germanic languages or were language experts. Beside Swiss-German dialects, Standard German, English and French, some participants reported knowledge of Italian (85), Spanish (57), Portuguese (7), Tagalog (5), Serbian (4), Hungarian, Romansh (3), Arabic, Cebuano, Dutch, Greek, Russian, Swahili (2), Bahasa, Catalan, Czech, Hebrew, Romanian, Tamil, Telegu, Thai and Turkish (1). Of the 163 participants that did not experience technical difficulties, we excluded a further fifteen participants because they did not provide self-ratings corresponding to any of the six CEFR levels for French or English, leaving a sample of 148 participants. We refer to Table 1 for the key descriptive statistics.

Task. We presented the participants 50 individual isolated Swedish words in random order on a computer screen. Forty-five of these words had German, English or French translation-equivalent cognates; five words had no cognate. Of the 45 target words, 40 had a German cognate, 27 an English cognate and 10 a French cognate. We refer to Appendix B for a list of the stimuli and their German, English and French cognates. As in Dataset 1, we had decided a priori to exclude participants able to translate more than two words without cognates correctly from the sample; none of the participants was able to do so, however. Participants were asked to indicate whether they thought that they might be able to translate the word presented into German. If so, a text box appeared in which they could enter their translation. Stimuli remained on-screen until the participants indicated whether they would attempt a translation. Every translation was marked as correct or incorrect.

3.2 Quantification of predictors

3.2.1 Formal distance

We computed the overall, consonantal and word-initial Levenshtein distances between each stimulus and its German, English and French counterparts as presented in Appendices A and B, i.e. nine distance measures per stimulus. Before the Levenshtein distance computations, the orthographic strings were

converted to lowercase. The German letter *ß*, which is not common in written standard German in Switzerland, was represented as *ss*.

We computed the overall Levenshtein distances using a binary algorithm that weighted all operations (insertions, deletions, substitutions) equally. Graphemes differing only in their diacritics, e.g. *ö* and *o* in *öppna–open*, were considered identical.[6] Only vowel–vowel and consonant–consonant mappings were allowed. For this purpose, we considered the graphemes *a*, *e*, *i*, *o*, *u* and *y* as well as their diacritical variants to be vowels and all other graphemes to be consonants. We length-normalised these raw overall distances by dividing them by the length of the longest lowest-cost alignment. For stimuli lacking a cognate in a given source language, we set the respective overall Levenshtein distance to the maximum, i.e. 1.

The consonantal Levenshtein distances were computed similarly to Gooskens et al.'s (2008) operationalisation. We first counted the number of consonant operations in the overall Levenshtein alignments and length-normalised these counts by dividing them by the length of the overall Levenshtein alignment. The consonantal Levenshtein distances for stimuli lacking a cognate in a given source language were set to 1.

Lastly, we computed the Levenshtein distances between the word beginnings of the stimuli and those of their German, French and English cognates. Word beginnings were operationally defined as up to and including the first consonant or consonant cluster. Thus, the word beginnings in the cognate pair *oorzaak–Ursache* (see Figure 1) are *oor* and *Ur*, which differ from each other in two slots. The word-initial distances were length-normalised by dividing them by the length of the overall Levenshtein alignment. In our example, this yields a word-initial Levenshtein distance of $2 \div 9 = 0.22$. The word-initial Levenshtein distance for stimuli lacking a cognate in a given source language was set to 1.[7]

3.2.2 Cognate frequency

As cognate frequency measures, we extracted the word frequencies per 1,000,000 words of the German and English translation-equivalent cognates of the L*x* stimuli from the SUBTLEX-DE (Brysbaert et al., 2011) and SUBTLEX-US (Brysbaert & New, 2009) databases, respectively. Word frequencies per 1,000,000 words for the French cognates were extracted from the Lexique 3 database (New, Brysbaert, Veronis & Pallier, 2007). These three databases, which are freely available from http://crr.ugent.be/programs-data/subtitle-frequencies and http://www.lexique.org, are highly comparable in that they are derived from similar corpora (film and television subtitles) that were selected according to similar principles. Unlike the SUBTLEX databases, however, Lexique 3 distinguishes homographs by part-of-speech. For full comparability with the SUBTLEX frequencies, we

6 Levenshtein distances for which aligned graphemes differing only in their diacritics received half the weight of an ordinary operation were correlated to the point of near-unity with the Levenshtein distances used for our analyses ($r > 0.99$).

7 An alternative approach would be to length-normalise the word-initial Levenshtein distances by dividing them by the length of the word-initial alignment only. However, word-initial Levenshtein distances thus computed are highly correlated with the ones described in the main text. For internal consistency, we therefore normalised the word-initial Levenshtein distances analogously to the consonantal Levenshtein distances, i.e. by the length of the full alignments.

aggregated the Lexique 3 frequencies over homographs. For the English cognate frequencies, we summed over the US and British spellings as well (e.g. the frequencies of *dike* and *dyke* were combined). All frequencies were logarithmically transformed in order to prevent items with extremely high frequency counts from exerting undue influence on the analyses. We added 1 to every frequency count in order to deal with zero frequencies, the logarithm of which would otherwise be undefined: log-frequency = ln(frequency + 1).

3.3 Random forest-based variable importance

We identified four classes of predictors that could help to account for variance the intelligibility of the written cognates to our participants: (a) overall Levenshtein distance, (b) consonantal Levenshtein distance, (c) word-initial Levenshtein distance and (d) cognate frequency. In each class, we have three predictors each, one for each potential supplier language under consideration (German, English and French). Twelve predictors is obviously too many to consider jointly in a regression that models the intelligibility of merely 180 (Dataset 1) and 45 (Dataset 2) items, respectively. This problem is compounded by the fact that the predictors show a high degree of multicollinearity. Thus, we need to select from our set of potential predictors the ones that can best explain the intelligibility of the cognates in our data set whilst accounting for the collinearity between our predictors. A popular solution for selecting variables for a regression model is to enter the variables in a stepwise regression model. However, among other problems, stepwise regression models are sensitive to order in which the variables are entered and are particularly affected by correlated predictors (see Whittingham, Stephens, Bradbury & Freckleton, 2006, for references and discussion). Instead, we turned to a fairly recent development in statistics that is better able to handle such cases: *random forests* (Breiman, 2001). For reasons of space, we can only briefly describe this algorithm below; for more detailed yet accessible introductions, we refer to Strobl, Malley and Tuz (2009) and Tagliamonte and Baayen (2012).

3.3.1 The rationale of random forests

Random forests are ensembles of classification or regression trees. Classification and regression trees seek to explain the variance in an outcome variable by recursively partitioning the data by means of binary splits so as to reach ever purer (i.e. more uniform with respect to the outcome variable) nodes. Tree models are flexible quantitative tools in that they can easily cope with interacting predictors, non-linearities and a multitude of predictors relative to the number of cases. They do, however, suffer from two severe drawbacks in particular. First, they are highly unstable: small changes in the data can and often do result in dramatically different tree models. Second, they are piecewise constant: even continuous relationships between the outcome variable and the predictor variables are broken down into dichotomies.

A popular solution to these problems is to grow not one tree but several hundreds of trees. By randomly resampling from the original set of cases, ‘new’ data sets are created for which new, different trees can be grown. Due to the random fluctuations between the training datasets, the ensemble as a

whole is much more robust than a single tree and the hard-cut boundaries that are characteristic of single trees are smoothed. In order to grow even more diverse trees, the set of possible predictors can be reduced randomly as well. For instance, we can specify that at each stage, only four out of 12 predictors are to be taken into consideration. This approach, the defining feature of random forests (Breiman, 2001), "allows predictor variables that were otherwise outplayed by their competitors to enter the ensemble" (Strobl, Boulesteix, Knib, Augustin & Zeileis, 2008: 307) and may reveal subtle interaction effects that would have remained hidden if a handful of variables had otherwise dominated the tree growing process.

3.3.2 Estimating variable importance

Random forests often produce excellent prediction rates, but unlike single trees, they are difficult to visualise and interpret, effectively making them 'black boxes'. To gain some insight into which variables are important, we can compute variable importance measures, e.g. the so-called 'permutation importance'. This importance measure is derived by computing how much the overall prediction accuracy of the random forest decreases when the values of a given predictor are randomly permuted, thereby breaking the association between the predictor and the outcome variable. The more important a predictor is in a random forest ensemble, the more this permuting will affect the ensemble's overall prediction accuracy. Such an approach, however, may overstate the importance of correlated variables, which is why Strobl et al. (2008) proposed the 'conditional permutation importance'. For this measure, the intercorrelations between predictor variables are taken into account by means of a conditioning scheme, too, thereby reducing the effect of collinearity on the permutation scores (see Strobl et al., 2008, for technical details). The variable importance measures thus computed reflect more closely the partial effects of each variable.

3.3.3 Software and settings

Random forests were grown using the `cforest()` function in the `party` package (Hothorn, Hornik, Strobl & Zeileis, 2013) for `R` (R Core Team, 2013). The individual trees were grown on subsamples consisting of 63.2% of the cases of the original data set (see Strobl, Boulesteix, Zeileis & Hothorn, 2007). Each forest consisted of 1,000 trees (i.e. `ntree` = 1,000) and four randomly selected predictors were considered at each split (i.e. `mtry` = 4). We then computed the conditional permutation importance measures for the predictors using `party`'s `varimp()` function. If the significance (i.e. the *p*-value) of the association between a given predictor and another covariate was lower than 0.80 (the default), the relevant covariate was included in the predictor's conditioning scheme. Runs with different settings converged to the same results as those reported here; different runs with the same settings likewise yielded similar results. All data and computer code are available at

http://dx.doi.org/10.6084/m9.figshare.763246, allowing interested readers to reproduce the analyses or run alternative analyses.[8]

4 Results

The percentage of correct translations per stimulus ranged from 0.0% to 100% in Dataset 1 (mean ± SD: 49 ± 31%) and from 1% to 93% in Dataset 2 (mean ± SD: 43 ± 29%). Thus, neither outcome variable was characterised by floor or ceiling effects.

4.1 Variable importances

We grew two random forests (one for each dataset) modelling the proportion of correct translations in terms of the twelve variables discussed above and computed the variables' conditional permutation importance scores. These importance scores are presented in Figures 2 and 3 for Datasets 1 and 2, respectively. (Note, however, that the permutation scores cannot be compared between studies in absolute terms; Strobl et al., 2009.) The dotted vertical lines provide conservative indications of the extent to which permutation scores can differ from zero due to randomness alone. Variables with importance scores to the left of this line are effectively irrelevant predictors in the random forest.

Both Figures 2 and 3 reveal that the overall Levenshtein distance between the L*x* stimuli and their German cognates is by far the most important predictor of L*x* stimulus intelligibility. Additionally, the overall Levenshtein distance with respect to English appears to be a predictor of stimulus intelligibility in both datasets, too. The overall Levenshtein distance with respect to French, by contrast, only seems to play a marginal role in Dataset 1, whereas its importance in Dataset 2 is negligible. Additionally, both kinds of partial Levenshtein distances—consonantal and word-initial Levenshtein distances—play marginal roles at best in Dataset 1 and can be considered to be irrelevant in Dataset 2. Lastly, the corpus frequencies of the stimuli's German and English cognates emerge as potential predictors in both studies; the frequency of the stimuli's French cognates does not seem to affect L*x* stimulus intelligibility.

8 To the reader wishing to reproduce our analyses, we point out that the computation of permutation importances for variables in random forests involves random sampling at various stages. While we have verified that the results reported in this article are robust across runs and for different settings, the precise values do oscillate somewhat between runs.

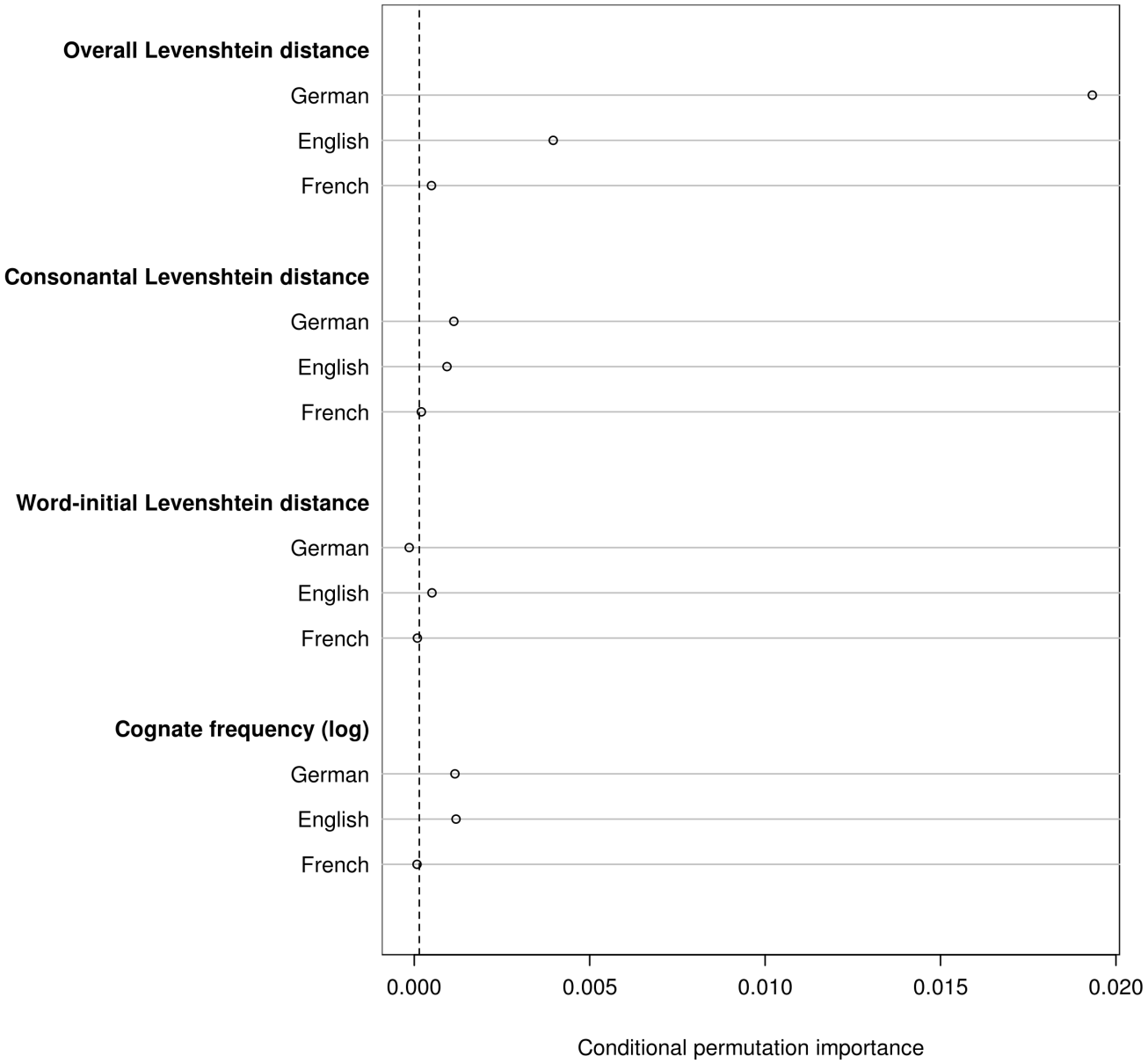

Figure 2. Conditional permutation importances for a random forest (mtry = 4, ntree = 1000) grown on Dataset 1 (180 items). The dotted vertical line provides a conservative indication of the extent to which the importance scores differ from zero due to random fluctuations; variables with permutation scores to the left of this line can be considered irrelevant in the random forest.

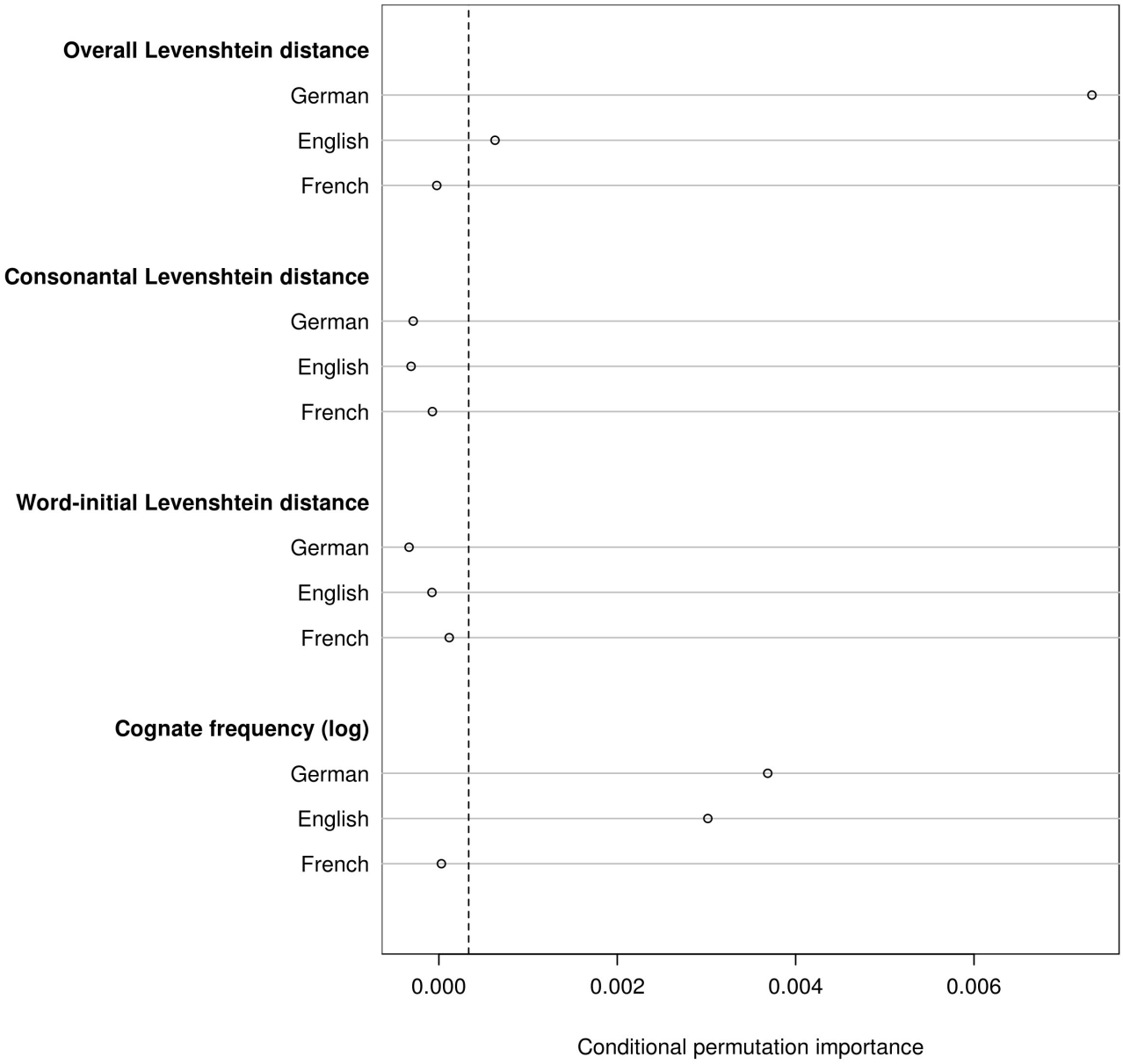

Figure 3. Conditional permutation importances for a random forest (mtry = 4, ntree = 1000) grown on Dataset 2 (45 items). The dotted vertical line provides a conservative indication of the extent to which the importance scores differ from zero due to random fluctuations; variables with permutation scores to the left of this line can be considered irrelevant in the random forest.

In order to take these possibilities into account, we computed 'Germanic' Levenshtein and frequency variables. For each stimulus, the overall Germanic Levenshtein distance equals whichever of the German or English Levenshtein distance is the lower one. The consonantal and word-initial Germanic Levenshtein distances are derived from the alignments associated with this overall Germanic Levenshtein distance. The Germanic cognate frequency is the mean of the stimulus's German and English cognate frequencies, which was then logarithmically transformed (ln(mean frequency + 1)). These four Germanic variables were added to the original 12 in order to grow another pair of random forests (`mtry` = 4, `ntree` = 1,000), on the basis of which new sets of conditional permutation importances were computed. These permutation scores are presented in Figures 4 and 5 for Datasets 1 and 2, respectively. In both cases, the overall Germanic Levenshtein distance clearly emerges as more important than its language-specific counterparts and Germanic cognate frequency pips the language-specific frequencies.

4.2 Regression modelling

The only two kinds of predictors that were found to have non-zero variable importance scores across the two datasets are the overall Levenshtein distances and the log-transformed corpus frequencies of the stimuli's cognates. In both cases, the 'Germanic' variables, in which the information with respect to German and English was collapsed, outperform the language-specific variables in terms of importance score. In order to render their effects more interpretable, we fitted cognate guessing success in terms of these two variables using regression techniques. As before, the data from the two studies were fitted separately to gauge the robustness of our results across similar but different, independent datasets.

Translation accuracy (correct–incorrect) was modelled in logistic linear mixed-effect models with random intercepts for both participants and stimuli and with overall Germanic Levenshtein distance and log-transformed Germanic cognate frequency as fixed-effect predictors (for introductions to logistic linear mixed-effect models geared towards language researchers, see Baayen, 2008, Chapter 7; Jaeger, 2008).[9] In order to account for individual-level differences in the effects of these variables (some participants may be more sensitive to Levenshtein distance or corpus frequency than others), we also modelled by-participant random slopes for these effects. The random-effect terms of the models are not discussed in this paper but can straightforwardly be recomputed using the supplementary materials that are available from http://dx.doi.org/10.6084/m9.figshare.763246.

9 Preliminary analyses (not reported here) indicated that the degree of non-linearity between the predictors and cognate guessing success was negligible. Furthermore, the Germanic Levenshtein distance and cognate frequency variables were not collinear with each other and could therefore safely be entered jointly into the regression model.

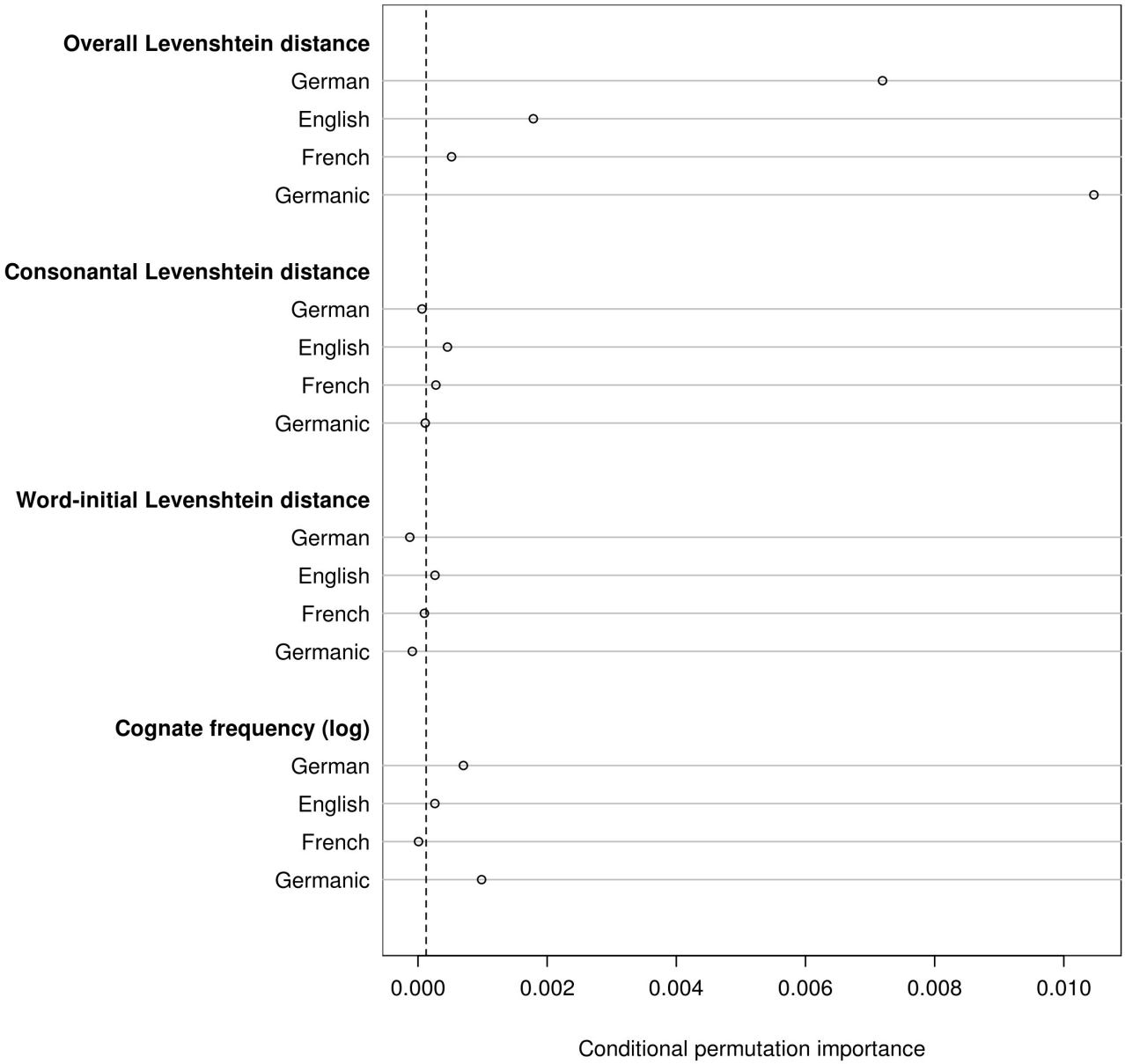

Figure 4. Conditional permutation importances for a random forest (mtry = 4, ntree = 1000) grown on Dataset 1 (180 items). The variables with respect to German and English were collapsed into 'Germanic' variables. The dotted vertical line provides a conservative indication of the extent to which the importance scores differ from zero due to random fluctuations; variables with permutation scores to the left of this line can be considered irrelevant in the random forest.

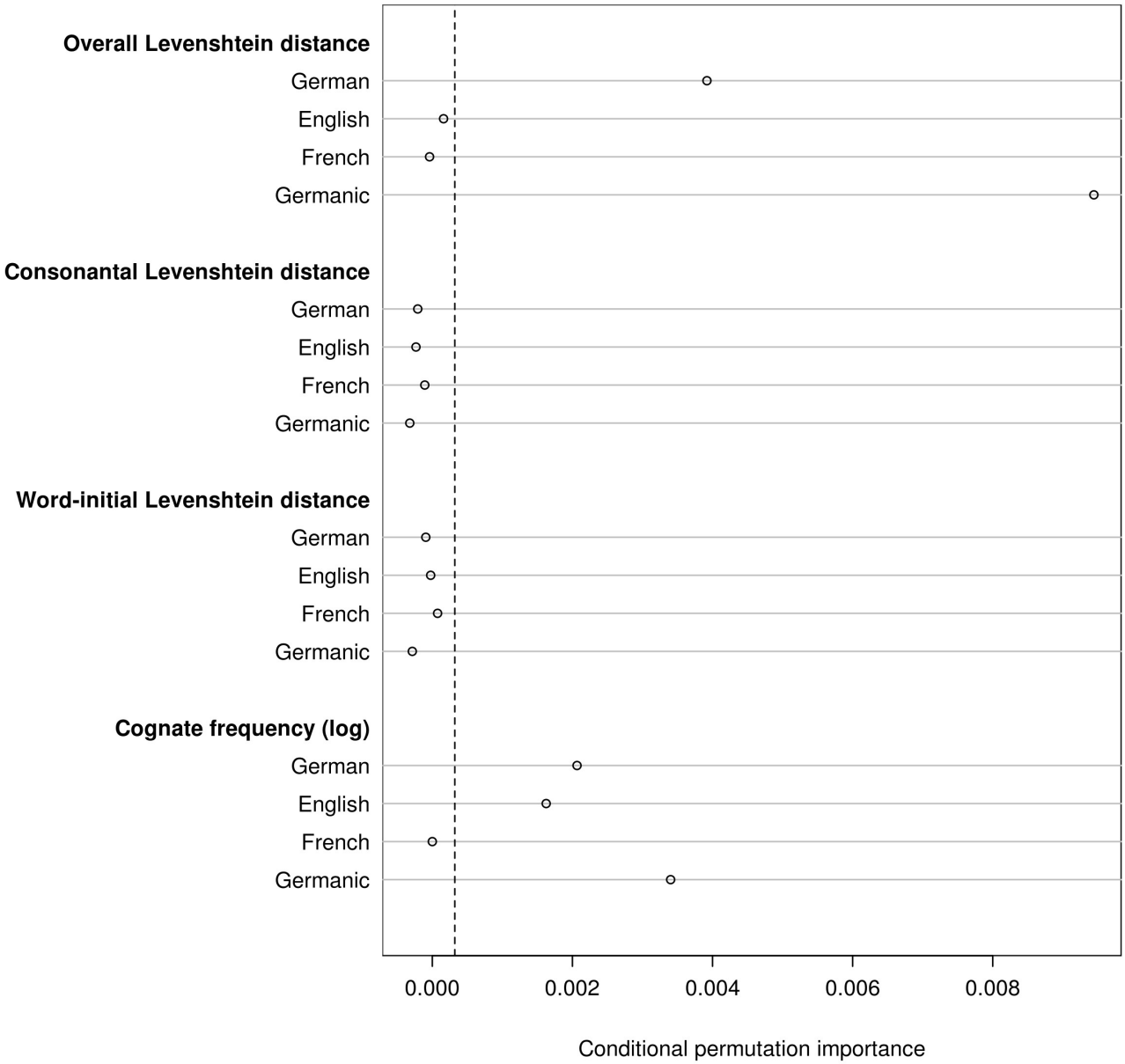

Figure 5. Conditional permutation importances for a random forest (mtry = 4, ntree = 1000) grown on Dataset 2 (45 items). The variables with respect to German and English were collapsed into 'Germanic' variables. The dotted vertical line provides a conservative indication of the extent to which the importance scores differ from zero due to random fluctuations; variables with permutation scores to the left of this line can be considered irrelevant in the random forest.

The fixed-effect parts of the models for Datasets 1 and 2 are presented in Tables 2 and 3, respectively. As expected, overall Germanic Levenshtein distance is negatively associated with cognate guessing success and Germanic cognate frequency shows a positive effect on translation accuracy. Furthermore, the parameter estimates (as well as the effect sizes) for both effects are highly similar between both studies: the estimated parameters for the overall Levenshtein distance effect (−5.4 ± 0.8 and −5.2 ± 1.1) do not appreciably differ from one another and the estimated parameters for the frequency effects (0.34 ± 0.09 and 0.41 ± 0.12) are likewise within less than one standard error from one another. This suggests that both effects are robust across similar yet different datasets in which the predictors were operationalised identically.

Table 2. Fixed-effect estimates of the logistic linear mixed-effect model for Dataset 1.

Variable	Estimate ± SE	p (z-test)
(Intercept)	−0.067 ± 0.157	0.671
Overall Germanic Levenshtein distance	−5.4 ± 0.8	< 0.001
Germanic cognate frequency (log)	0.34 ± 0.09	< 0.001

Note. Apart from the fixed effects, the model includes by-participant and by-item random intercepts and by-participant slopes for the Levenshtein and frequency effects. Parameters are expressed in log-odds. Predictor variables were centred at their means.

Table 3. Fixed-effect estimates of the logistic linear mixed-effect model for Dataset 2.

Variable	Estimate ± SE	p (z-test)
(Intercept)	−0.59 ± 0.23	0.011
Overall Germanic Levenshtein distance	−5.2 ± 1.1	< 0.001
Germanic cognate frequency (log)	0.41 ± 0.12	< 0.001

Note. See Table 2 for details.

An analysis of the models' residuals and by-item random intercepts did not reveal any robust patterns with respect to any of the variables considered in this study. This suggests that the fit of our models is satisfactory and that no variables not present in the model would help to explain between-item differences in cognate guessing success.

5 Discussion and conclusions

In this article, we investigated the impact of a selection of item-related characteristics on the probability with which multilinguals can guess the meaning of visually presented words in an unknown language (L*x*) with known cognates. Importantly, we did so by taking into account the possibility that the participants relied on transfer bases not just from their L1s (esp. Standard German) but also from two

of their L2+s (viz. French and English). We first gauged the impact of four kinds of variables (each operationalised with respect to three potential supplier languages) on cognate guessing accuracy. Then, we used those insights to fit regression models not tainted by collinearity between the predictor variables. The danger of overfitting that is associated with exploratory analyses was reduced by analysing two different but similar studies independently from one another, thereby ensuring that the results are robust with respect to both the stimuli and the participants that were sampled.

The results indicated that the overall degree of orthographic discrepancy between an L*x* stimulus in a Germanic language and a known cognate in German or English (computed by means of the Levenshtein algorithm) is the most important item-related predictor of cognate guessing accuracy. This finding replicates the results of similar studies that found orthographic distance to be a major factor in cognate guessing (see Section 2.1). At the same time, it provides additional evidence that multilingual participants draw not merely on their L1 when guessing the meaning of cognates in a related language but also on their knowledge of a related foreign language: orthographic distances computed with respect to the closest German or English cognate outperformed those with respect to the German cognate alone. The orthographic distance between the L*x* stimuli and their French cognates turned out not to be a robust predictor of cognate guessing accuracy and its effect was marginal at best. We will discuss this point below.

Previous studies had suggested that participants particularly rely on the consonantal frames and word-initial elements when taking part in cognate guessing tasks or receptive multilingualism. If consonants and word beginnings play a privileged role in cognate guessing, we would expect consonantal and word-initial orthographic discrepancies between the stimuli and their known cognates to predict task performance over and beyond what can be accounted for by the overall, full-word orthographic discrepancies. This was not the case: consonantal and word-initial Levenshtein distances were not robust predictors of cognate guessing accuracy across the two datasets when overall Levenshtein distances were considered, too. In contrast to findings by Berthele (2011), Gooskens et al. (2008), Möller (2011), Möller and Zeevaert (2010) and Müller-Lancé (2003), our data therefore suggests that consonants and word beginnings do not impact cognate guessing accuracy more than vowels and word middles and endings. Rather, the correlations between consonantal and word-initial Levenshtein distances and cognate guessing accuracy seem to be by-products of their intercorrelations with the more important predictor, i.e. overall Levenshtein distance. This is not to say that performance would not have been better *if* the participants had focused more strongly on consonantal and word-initial similarities as opposed to vocalic and word-medial or -final similarities—an altogether different question. But our results do carry the pedagogical implication that multilingual language learners and participants in receptive multilingualism cannot necessarily be counted on to do so without prior sensitisation or experience.

A predictor that did account for between-item differences in cognate guessing accuracy, however, was the corpus frequencies of the stimuli's German and English cognates. Stimuli with high-frequency

German and English cognates were translated correctly more often than stimuli with low-frequency cognates. As with the overall Levenshtein distances, a measure comprising both German and English performed better than its single-language counterparts. This further suggests that multilinguals make use of related languages beside the L1 when guessing the meaning of words in a related language. The corpus frequencies of the stimuli's French cognates, however, did not emerge as a relevant predictor of cognate guessing accuracy.

So, for both the orthographic distances and corpus frequency, the variables computed with respect to French did not turn out to be robust predictors of cognate guessing accuracy. Speculatively, this lack of importance of French-based variables may have multiple reasons. First, the self-assessments indicate that the participants' French skills were less developed than their English skills (see Table 1), even though French was the first foreign language for most of them. Participants may have been less likely to bring to bear their knowledge of French as a result (Meißner & Burk, 2001; Williams & Hammarberg, 1998). Second, the number of Germanic target words with a French cognate was limited, and most target words with a French cognate also had German or English cognates. A higher proportion of target words with French cognates but without German or English cognates could have yielded different results. However, the fact of the matter is that most words with French origins occurring in Danish, Dutch, Frisian and Swedish also occur in German and English. A study with a higher proportion of French-only cognates would thus have been less relevant to receptive multilingualism in the Germanic language family. A third, related point is that the psychotypological distance (Kellerman, 1977, 1983) from the Germanic L*x*s to German and English is likely to have been smaller than that to French, enhancing the relative likelihood of German and English serving as the supplier languages. It seems conceivable, however, that this psychotypological distance can be influenced by including a higher proportion of target words related only to a French cognate. Thus, our results should not be interpreted as suggesting that German-speaking Swiss participants *will not* under any circumstance draw on their knowledge of French when guessing the meaning of related words in an unknown language, but rather that a task more conducive to French–L*x* transfer is likely needed to detect such an effect.

Lastly, we found random forests, a relatively new advance in statistics, to provide us with useful, orderly insights into the structure underlying our datasets. While random forests are not the only way in which these data could be analysed, they are part of an increasingly growing toolbox of statistical methods that have come to the fore in recent years with which multidimensional datasets can be analysed and that could fruitfully be applied to multilingualism data.

Acknowledgements

The first author received private financial support from Ambros Boner, M.D. Dataset 2 was supported by a *Sinergia* grant from the Swiss National Science Foundation (*Multilingualism through the lifespan*, project number 130457, PI Raphael Berthele). Thanks are due to Charlotte Gooskens (University of

Groningen) for making available the multilingual list of frequent nouns used in Dataset 1. We also thank our students Carlos Pestana, Christof Chesini, Katharina Karges, Thomas Michel, Magalie Desgrippes, Barbara Grüter, Fabienne Strässle, Barbara Rampolla, Marylin Krieg, Nathalie Bagnoud, Annina Keller and Melahat Yapici (in no particular order) for collecting the data for Dataset 1 and our colleagues Irmtraud Kaiser, Lenny Bugayong and Nuria Ristin (University of Fribourg) for their assistance in collecting the data for Dataset 2.

References

Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Berthele, R. 2008. Dialekt-Standard Situationen als embryonale Mehrsprachigkeit. Erkenntnisse zum interlingualen Potenzial des Provinzlerdaseins. *Sociolinguistica* 22, 87–107.

Berthele, R. 2011. On abduction in receptive multilingualism: Evidence from cognate guessing tasks. *Applied Linguistics Review* 2, 191–219.

Berthele, R. & A. Lambelet. 2009. Approche empirique de l'intercompréhension: répertoires, processus et résultats. *Lidil* 39, 151–162.

Braunmüller, K. & L. Zeevaert. 2001. Semikommunikation, rezeptive Mehrsprachigkeit und verwandte Phänomene: Eine bibiographische Bestandsaufnahme. *Arbeiten zur Mehrsprachigkeit*, B-19. Hamburg: Sonderforschungsbereich 538 *Mehrsprachigkeit*, University of Hamburg.

Breiman, L. 2001. Random forests. *Machine Learning* 45, 5–32.

Broerse, A. C. & E. J. Zwaan. 1966. The information value of initial letters in the identification of words. *Journal of Verbal Learning and Verbal Behavior* 5, 441–446.

Brysbaert, M., M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte & A. Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology* 58, 412–424.

Brysbaert, M. & B. New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41, 977–990.

Carton, A. S. 1971. Inferencing: A process in using and learning language. In P. Pimsleur & T. Quinn (eds), *The psychology of second language learning: Papers from the Second International Congress of Applied Linguistics, Cambridge, 8–12 September 1969*, 45–58. London: Cambridge University Press.

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

De Groot, A. M. B. & R. Keijzer. 2000. What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning* 50, 1–56.

Dijkstra, T. & W. J. B. van Heuven. 2002. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition* 5, 175–197.

Fox, J. & S. Weisberg. 2011. *An R companion to applied regression*, 2nd edn. Thousand Oaks, CA: Sage.

Gibson, M. & B. Hufeisen. 2003. Investigating the role of prior foreign language knowledge: Translating from an unknown into a known foreign language. In J. Cenoz, B. Hufeisen & U. Jessner (eds), *The multilingual lexicon*, 87–102. Dordrecht: Kluwer.

Gooskens, C., W. J. Heeringa & K. Beijering. 2008. Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing* 2, 63–81.

Gooskens, C., S. Kürschner & R. van Bezooijen. 2011. Intelligibility of Standard German and Low German to speakers of Dutch. *Dialectologia* Special issue, II, 35–63.

Heeringa, W. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. PhD thesis, University of Groningen.

Hothorn, T., K. Hornik, C. Strobl & A. Zeileis. 2013. party: A laboratory for recursive partitioning. R package, version 1.0-6. Available from http://cran.r-project.org/package=party

Hufeisen, B. & N. Marx (eds). 2007. *EuroComGerm — Die sieben Siebe: germanische Sprachen lesen lernen*. Aachen: Shaker.

Jaeger, T. F. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59, 434–446.

Johnson, R. L. & M. E. Eisler. 2012. The importance of the first and last letter in words during sentence reading. *Acta Psychologica* 141, 336–351.

Kellerman, E. 1977. Toward a characterisation of the strategy of transfer in second language learning. *Interlanguage Studies Bulletin* 2, 58–145.

Kellerman, E. 1983. Now you see it, now you don't. In S. M. Gass & L. Selinker (eds), *Language transfer in language learning*, 140–156. Rowley (MA): Newbury House.

Klein, H. G. & T. D. Stegmann (eds). 2000. *EuroComRom — Die sieben Siebe: romanische Sprachen sofort lesen können*. Aachen: Shaker.

Kürschner, S., C. Gooskens & R. van Bezooijen. 2008. Linguistic determinants of the intelligibility of Swedish words among Danes. *International Journal of Humanities and Arts Computing* 2, 83–100.

Lemhöfer, K., T. Dijkstra, H. Schriefers, R. H. Baayen, J. Grainger & P. Zwitserlood. 2008. Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34, 12–31.

Lotto, L. & A. M. B. de Groot. 1998. Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning* 48, 31–69.

Meißner, F.-J. & H. Burk. 2001. Hörverstehen in einer unbekannten romanischen Fremdsprache und methodische Implikationen für den Tertiärsprachenerwerb. *Zeitschrift für Fremdsprachenforschung* 12, 63–110.

Moates, D. R. & E. A. Marks. 2012. Vowel mutability in print in English and Spanish. *The Mental Lexicon* 7, 326–350.

Möller, R. 2011. Wann sind Kognaten erkennbar? Ähnlichkeit und synchrone Transparenz von Kognatenbeziehungen in der germanischen Interkomprehension. *Linguistik online* 46, 79–101.

Möller, R. & L. Zeevaert. 2010. “Da denke ich spontan an Tafel”: Zur Worterkennung in verwandten germanischen Sprachen. *Zeitschrift für Fremdsprachenforschung* 21, 217–248.

Müller, M., L. Wertenschlag, S. Gerhartl, A. Halibasic, I. Kaiser, E. Peyer, N. Shafer & R. Berthele. 2009. *Chunsch druus? Schweizerdeutsch verstehen – die Deutschschweiz verstehen.* Bern: Schulverlag.

Müller-Lancé, J. 2003. A strategy model of multilingual learning. In J. Cenoz, B. Hufeisen & U. Jessner (eds), *The multilingual lexicon*, 117–132. Dordrecht: Kluwer.

New, B., M. Brysbaert, J. Veronis & C. Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28, 661–677.

Odlin, T. 1989. *Language transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.

Otwinowska-Kasztelanic, A. 2011. Awareness and affordances: Multilingual versus bilinguals and their perception of cognates. In G. De Angelis & J.-M. Dewaele (eds), *New trends in crosslinguistic influence and multilingualism research*, 1–18. Bristol: Multilingual Matters.

R Core Team. 2013. *R: A language and environment for statistical computing*. Software, version 2.15.3. Available from http://www.r-project.org

Ringbom, H. 2007. *Cross-linguistic similarity in foreign language learning*. Clevedon: Multilingual Matters.

Scaltritti, M. & D. A. Balota. 2013. Are all letters really processed equally and in parallel? Further evidence of a robust first letter advantage. *Acta Psychologica* 144: 397–410.

Singleton, D. & D. Little. 1984. A first encounter with Dutch: Perceived language distance and language transfer as factors in comprehension. In L. Mac Mathúna & D. Singleton (eds), *Language across cultures: Proceedings of a symposium held at St Patrick’s College, Dublin 8–9 July 1983*, 259–270. Dublin: Irish Association for Applied Linguistics.

Strobl, C., A-L. Boulesteix, T. Kneib, T. Augustin & A. Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9, 307.

Strobl, C., A.-L. Boulesteix, A. Zeileis & T. Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8, 25.

Strobl, C., J. Malley & G. Tuz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 323–348.

Swarte, F., A. Schüppert & C. Gooskens. 2013. Do speakers of Dutch use their knowledge of German while processing written Danish words? *Linguistics in the Netherlands* 30, 146–159.

Tagliamonte, S. A. & R. H. Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24, 135–178.

Van Bezooijen, R. & C. Gooskens. 2005. How easy is it for speakers of Dutch to understand Frisian and Afrikaans, and why? *Linguistics in the Netherlands* 22, 13–24.

Van Bezooijen, R., C. Gooskens & S. Kürschner. 2012. Deens is makkelijker voor Friezen dan voor Nederlanders — feit of fabel? In P. Boersma, G. T. Jensma & R. Salverda (eds), *Philologia Frisica anno 2008. Lêzings fan it achttjinde Frysk Filologekongres fan de Fryske Akademy op 10, 11 en 12 desimber 2008*, 286–298. Leeuwarden: Afûk/Fryske Akademy.

Van Heuven, V. J. 2008. Making sense of strange sounds: (Mutual) intelligibility of related language varieties. A review. *International Journal of Humanities and Arts Computing* 2, 39–62.

Vanhove, J. 2014. *Receptive multilingualism across the lifespan. Cognitive and linguistic factors in cognate guessing*. PhD thesis, University of Fribourg.

Vanhove, J. & R. Berthele. 2013. Factoren bij het herkennen van cognaten in onbekende talen: algemeen of taalspecifiek? *Taal en Tongval* 65, 171–210.

Vanhove, J. & R. Berthele. 2015. The lifespan development of cognate guessing skills in an unknown related language. *International Review of Applied Linguistics in Language Teaching* 53, 1–38.

Weinreich, U. 1953. *Languages in contact. Findings and problems*. New York: Linguistic Circle of New York.

Wenzel, V. 2007. Rezeptive Mehrsprachigkeit und Sprachdistanz deutsch-niederländisch. *Zeitschrift für Fremdsprachenforschung* 18, 183–208.

Whittingham, M. J., P. A. Stephens, R. B. Bradbury & R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology* 75, 1182–1189.

Williams, S. & B. Hammarberg. 1998. Language switches in L3 production: Implications for a polyglot speaking model. *Applied Linguistics* 19, 295–333.

Appendices

Stimuli for Dataset 1

L*x*	Stimulus	German translation	German cognate	English cognate	French cognate
DK	antal	Anzahl	Anzahl		
	befolkning	Bevölkerung	Bevölkerung		
	besøg	Besuch	Besuch		
	bidrag	Beitrag	Beitrag		
	bjerg	Berg	Berg		
	brød	Brot	Brot	bread	
	butik	Boutique	Boutique	boutique	boutique
	dag	Tag	Tag	day	
	detalje	Detail	Detail	detail	détail
	enhed	Einheit	Einheit		
	farve	Farbe	Farbe		
	flaske	Flasche	Flasche		
	forbrænding	Verbrennung	Verbrennung	burning	
	forening	Verein; Vereinigung	Vereinigung		
	forsikring	Versicherung	Versicherung		
	frugt	Frucht	Frucht	fruit	fruit
	hjælp	Hilfe	Hilfe	help	
	indsigt	Einsicht	Einsicht	insight	
	klokke	Glocke; Uhr	Glocke	clock	cloche
	konkurrence	Konkurrenz	Konkurrenz		concurrence
	kursus	Kurs	Kurs	course	cours
	kyst	Küste	Küste	coast	côte
	lyst	Lust	Lust		
	mængde	Menge	Menge		
	majoritet	Mehrheit	Majorität	majority	majorité
	måned	Monat	Monat	month	mois
	marts	März	März	March	mars
	mening	Meinung; Bedeutung; Sinn	Meinung	meaning	
	opmærksomhed	Aufmerksamkeit	Aufmerksamkeit		
	plads	Platz	Platz	place	place
	pligt	Pflicht	Pflicht		
	pris	Preis	Preis	price	prix

L*x*	Stimulus	German translation	German cognate	English cognate	French cognate
	præst	Priester	Priester	priest	prêtre
	regn	Regen	Regen	rain	
	ræsonnement	Darlegung	Räsonnement	reasoning	raisonnement
	sammenkomst	Zusammenkunft	Zusammenkunft		
	skole	Schule	Schule	school	école
	skridt	Schritt	Schritt		
	skræk	Schreck	Schreck		
	smerte	Schmerz	Schmerz		
	sne	Schnee	Schnee	snow	
	stol	Stuhl	Stuhl		
	stykke	Stück	Stück		
	synd	Sünde	Sünde	sin	
	tryk	Druck	Druck		
	kylling	Huhn	(profile word)		
	penge	Geld	(profile word)		
	seng	Bett	(profile word)		
	skov	Wald	(profile word)		
	æske	Dose	(profile word)		
NL	aarde	Erde	Erde	earth	
	armoede	Armut	Armut		
	bedrog	Betrug	Betrug		
	boom	Baum	Baum		
	coördinatie	Koordination	Koordination	coordination	coordination
	ervaring	Erfahrung	Erfahrung		
	geest	Geist	Geist	ghost	
	gevaar	Gefahr	Gefahr		
	gevoel	Gefühl	Gefühl	feeling	
	gezondheid	Gesundheit	Gesundheit		
	gordijn	Gardine	Gardine		
	hemel	Himmel	Himmel		
	hond	Hund	Hund		
	huid	Haut	Haut		
	kaas	Käse	Käse	cheese	
	kennis	Kenntnis	Kenntnis		
	lucht	Luft	Luft		
	mond	Mund	Mund	mouth	
	nadeel	Nachteil	Nachteil		
	oorzaak	Ursache	Ursache		

L*x*	Stimulus	German translation	German cognate	English cognate	French cognate
	paal	Pfahl	Pfahl	pole	
	paard	Pferd	Pferd		
	politie	Polizei	Polizei	police	police
	rechter	Richter	Richter		
	relatie	Verhältnis	Relation	relation	relation
	rook	Rauch	Rauch		
	samenhang	Zusammenhang	Zusammenhang		
	schip	Schiff	Schiff	ship	
	straat	Strasse	Strasse	street	
	tante	Tante	Tante	aunt	tante
	toegang	Zugang	Zugang		
	touw	Seil, Tau	Tau		
	trein	Zug		train	train
	uitgave	Ausgabe	Ausgabe		
	vak	Fach	Fach		
	verhoging	Erhöhung	Erhöhung		
	versie	Version	Version	version	version
	vlag	Flagge	Flagge	flag	
	vleugel	Flügel	Flügel		
	vluchteling	Flüchtling	Flüchtling		
	voorbereiding	Vorbereitung	Vorbereitung		
	vrouw	Frau	Frau		
	water	Wasser	Wasser	water	
	werk	Arbeit	Werk	work	
	zaak	Sache	Sache		
	zaal	Saal	Saal		salle
	lichaam	Körper	(profile word)		
	meisje	Mädchen	(profile word)		
	voorwerp	Gegenstand	(profile word)		
	wet	Gesetz	(profile word)		
FR	aai	Ei	Ei	egg	
	behandeling	Behandlung	Behandlung		
	bibleteek	Bibliothek	Bibliothek		bibliothèque
	budzjet	Budget	Budget	budget	budget
	doarp	Dorf	Dorf		
	dyk	Deich	Deich	dyke	digue
	eilân	Insel; Eiland	Eiland	island	île
	feest	Fest	Fest	feast	fête

L*x*	Stimulus	German translation	German cognate	English cognate	French cognate
	ferdigening	Verteidigung	Verteidigung		
	ferplichting	Verpflichtung	Verpflichtung		
	fertrouwe	Vertrauen	Vertrauen		
	ferwizing	Verweisung	Verweisung		
	frede	Frieden	Frieden		
	freondin	Freundin	Freundin	friend	
	garaazje	Garage	Garage	garage	garage
	geslacht	Geschlecht	Geschlecht		
	groep	Gruppe	Gruppe	group	groupe
	haven	Hafen	Hafen		
	klas	Klasse	Klasse	class	classe
	koar	Chor	Chor	choir	chœur
	masine	Maschine	Maschine	machine	machine
	materiaal	Material	Material	material	matériel
	namme	Name	Name	name	nom
	oanfiering	Anführung	Anführung		
	oplossing	Auflösung; Lösung	Auflösung		
	organisaasje	Organisation	Organisation	organisation	organisation
	parlemint	Parlament	Parlament	parliament	parlement
	posysje	Position	Position	position	position
	punt	Punkt	Punkt	point	point
	regearing	Regierung	Regierung		
	rivier	Fluss		river	rivière
	sintimeter	Zentimeter	Zentimeter	centimeter	centimètre
	sintrum	Zentrum	Zentrum	centre	centre
	situaasje	Situation	Situation	situation	situation
	slang	Schlange	Schlange		
	stim	Stimme	Stimme		
	stoarm	Sturm	Sturm	storm	
	sûkelade	Schokolade	Schokolade	chocolate	chocolat
	tiisdei	Dienstag	Dienstag	Tuesday	

L*x*	Stimulus	German translation	German cognate	English cognate	French cognate
	ungelok	Unglück	Unglück		
	untwikkeling	Entwicklung	Entwicklung		
	útspraak	Aussprache	Aussprache		
	woartel	Wurzel	Wurzel		
	wyn	Wein	Wein	wine	vin
	ynfloed	Einfluss	Einfluss	influence	influence
	belied	Politik	(profile word)		
	buert	Nachbarschaft	(profile word)		
	gat	Loch	(profile word)		
	heit	Vater	(profile word)		
	lawaai	Lärm	(profile word)		
SE	anslutning	Anschluss	Anschluss		
	applåd	Applaus	Applaus	applause	applaudissement
	balans	Balance	Balance	balance	balance
	blad	Blatt	Blatt		
	blod	Blut	Blut	blood	
	bröst	Brust	Brust	breast	
	chans	Chance	Chance	chance	chance
	cyckel*	Fahrrad; Zyklus	Zyklus	cycle	cycle
	eftermiddag	Nachmittag	Nachmittag	afternoon	
	elev	Schüler			élève
	fjäder	Feder	Feder	feather	
	frihet	Freiheit	Freiheit	freedom	
	förbättring	Verbesserung	Verbesserung		
	förhandling	Verhandlung	Verhandlung		
	förmåga	Vermögen	Vermögen		
	gemenskap	Gemeinschaft	Gemeinschaft		
	hår	Haar	Haar	hair	
	intryck	Eindruck	Eindruck		
	invånare	Einwohner	Einwohner		
	konst	Kunst	Kunst		

* The correct spelling is *cykel*.

L*x*	Stimulus	German translation	German cognate	English cognate	French cognate
	krig	Krieg	Krieg		
	makt	Macht	Macht		
	morgon	Morgen	Morgen	morning	
	mur	Mauer	Mauer		mur
	nivå	Niveau	Niveau		niveau
	näsa	Nase	Nase	nose	nez
	omgivning	Umgebung	Umgebung		
	ordning	Ordnung	Ordnung	order	ordre
	press	Presse	Presse	press	presse
	resa	Reise	Reise		
	riktning	Richtung	Richtung		
	skada	Schaden	Schaden		
	stjärna	Stern	Stern	star	star
	strävan	Streben	Streben		
	succé	Erfolg		success	succès
	sång	Lied		song	
	tunga	Zunge	Zunge	tongue	
	tält	Zelt	Zelt		
	undersökning	Untersuchung	Untersuchung		
	universitet	Universität	Universitat	university	université
	utbildning	Ausbildung	Ausbildung		
	utland	Ausland	Ausland		
	väg	Weg	Weg	way	
	öga	Auge	Auge	eye	
	ögonblick	Augenblick	Augenblick		
	bonde	Bauer	(profile word)		
	hav	Meer	(profile word)		
	helg	Wochenende	(profile word)		
	önskan	Wunsch	(profile word)		
	skäl	Grund	(profile word)		

Stimuli for Dataset 2

Stimulus	German translation	German cognate	English cognate	French cognate
alltid	allzeit, immer	allzeit		
avskaffa	abschaffen	abschaffen		
bakgrund	hintergrund		background	
behärska	beherrschen	beherrschen		
borgmästare	Bürgermeister	Bürgermeister		
byrå	Büro	Büro	bureau	bureau
bäbis†	Baby	Baby	baby	bébé
cyckel‡	Zyklus; Fahrrad	Zyklus	cycle	cycle
fiende	Feind	Feind	fiend	
fåtölj	Fauteuil	Fauteuil	fauteuil	fauteuil
försiktig	vorsichtig	vorsichtig		
förutsättning	Voraussetzung	Voraussetzung		
full	voll	voll	full	
hård	hard	hart	hard	
kanel	Zimt			cannelle
kejsar	Kaiser	Kaiser		
kniv	Messer		knife	
kung	König	König	king	
kyrka	Kirche	Kirche	church	
kyssa	küssen	küssen	kiss	
löpa	laufen	laufen		
mjölk	Milch	Milch	milk	
möjlig	möglich	möglich		
rytmisk	rhythmisch	rhythmisch	rhythmic	rythmique
rådhus	Rathaus	Rathaus		
saliv	Speichel		saliva	salive
skola	Schule	Schule	school	
skrubba	schrubben	schrubben	scrub	

† *Bäbis* is a common misspelling for *bebis*.
‡ The correct spelling is *cykel*.

Stimulus	German translation	German cognate	English cognate	French cognate
skyskrapa	Hochhaus, Wolkenkratzer		skyscraper	
sitta	sitzen	sitzen	sit	
skön	schön	schön		
spegel	Spiegel	Spiegel		
språk	Sprache	Sprache		
stjärn	Stern	Stern	star	(star)
söka	suchen	suchen	seek	
torsdag	Donnerstag	Donnerstag	Thursday	
tunga	Zunge	Zunge	tongue	
tvivla	zweifeln	zweifeln		
tårta	Torte	Torte	tart	tarte
varm	warm	warm	warm	
viktig	wichtig	wichtig		
värld	Welt	Welt	world	
ytterst	äusserst	äusserst		
öppna	öffnen	öffnen	open	
översätta	übersetzen	übersetzen		
barn	Kind	(profile word)		
häst	Pferd	(profile word)		
leka	spielen	(profile word)		
mycket	viel; sehr	(profile word)		
städa	putzen	(profile word)		