

Translation ambiguity in and out of context

ANAT PRIOR and SHULY WINTNER
University of Haifa

BRIAN MACWHINNEY and ALON LAVIE
Carnegie Mellon University

Received: January 29, 2009 Accepted for publication: October 13, 2009

ADDRESS FOR CORRESPONDENCE

Anat Prior, Edmond J. Safra Brain Research Center for the Study of Learning Disabilities,
University of Haifa, Room 276, Faculty of Education, Mount Carmel, Haifa, Israel.
E-mail: aprior@construct.haifa.ac.il

ABSTRACT

We compare translations of single words, made by bilingual speakers in a laboratory setting, with contextualized translation choices of the same items, made by professional translators and extracted from parallel language corpora. The translation choices in both cases show moderate convergence, demonstrating that decontextualized translation probabilities partially reflect bilinguals' life experience regarding the conditional distributions of alternative translations. Lexical attributes of the target word differ in their ability to predict translation probability: form similarity is a stronger predictor in decontextualized translation choice, whereas word frequency and semantic salience are stronger predictors for context-embedded translation choice. These findings establish the utility of parallel language corpora as important tools in psycholinguistic investigations of bilingual language processing.

Bilinguals are often faced with the task of translating words from one language to the other. In this task, there are frequently two or even more plausible translations for a given word. For example, when translating from Spanish to English, one has to decide whether the verb *decir* should be translated as *say* or as *tell*. When translating from English to Spanish, one similarly has to decide whether *glass* should be translated as *vidrio* (the substance) or *vaso* (the utensil). For common words like *take* or *for*, translation ambiguity may lead to there being as many as five, six, or even more possible translations. In a study of Dutch and English, Tokowicz, Kroll, De Groot, and van Hell (2002) found that, even for a sample of words that were selected to ostensibly have only one translation, 24% were found to have more than one possible translation. In a study of English and Spanish that included a freer sampling of words, Prior, MacWhinney, and Kroll (2007) found that 60% of the items displayed translation ambiguities. The case of English and Spanish might be especially prone to multiple translations, because both languages

are widely spoken in different parts of the world, and local variations in word preference might arise over time.

There is good reason to believe that translation ambiguity plays a major role in real-life bilingual language processing. Bilingual speakers have more difficulty producing less probable as opposed to more probable translations (Prior, Kroll, & MacWhinney, 2006; Tokowicz & Kroll, 2007; Tokowicz, Prior, & Kroll, 2009). However, professional translators and simultaneous interpreters must nonetheless be able to access a less probable translation in real time, presumably by relying on cues from the discourse context.

To the best of our knowledge, the issue of translation ambiguity has only been examined for words presented in isolation in controlled experiments. However, it is unclear to what degree measurements of this type are related to the processing of translation ambiguity in real-life contexts. To better evaluate the general nature of this problem, we can supplement experimental approaches with methods from corpus analysis. In the present work, we examine translation ambiguities in large electronic corpora created by professional language translators, focusing on translation between English and Spanish. We then compare the single-word translation choices made by bilingual speakers of English and Spanish with those made by professional translators. This comparison illuminates three important issues. First, through study of these corpora, we can assess the general level of translation ambiguity in real-life, contextualized, corpus materials. Second, we can compare the specific choices made in the contextualized situation with those made in the decontextualized laboratory situation. Can common influences on translation choice (e.g., word frequency) be identified in the two cases? Third, from a methodological point of view, we can examine the extent to which the study of parallel language corpora can support, refine, and extend psycholinguistic studies of bilingual language processing.

In the following sections we outline previous behavioral findings regarding translation ambiguity, sketch the use of large language corpora in psycholinguistic research, and explain the origin and composition of parallel language corpora. We conclude by laying out the hypotheses guiding the present study.

TRANSLATION AMBIGUITY

Natural languages are notoriously ambiguous on various levels. Semantically, a single word can have more than one meaning, with the two readings belonging either to the same (i.e., *bank*) or different (i.e., *back*) grammatical categories or parts of speech (Lyons, 1995). Syntactic ambiguity arises when an entire sentence can imply more than one underlying structure, as in *flying planes can be dangerous* (for perspectives on ambiguity resolution, see Clifton, Frazier, & Rayner, 1994; Gorfien, 2002). Thus, the intended meaning of a single word can vary greatly depending on the linguistic context in which it appears. Psycholinguistic studies of monolingual language processing have demonstrated that, in most cases, both meanings of ambiguous words are accessed, and that the cognitive system overcomes this obstacle mostly by relying on the linguistic context (Gernsbacher, Robertson & Werner, 2001; Kambe, Rayner & Duffy, 2001; Kellas, Ferraro & Simpson, 1988; for a review of cross-linguistic lexical ambiguity resolution, see

Altarriba & Gianico, 2003). It is somewhat surprising that the cross-linguistic implications of word sense ambiguity, as well as additional sources of translation ambiguity, have only recently come under experimental scrutiny within cognitive psycholinguistics (Degani, Prior, & Tokowicz, 2009; Jiang, 2002; Prior et al., 2007, Tokowicz & Kroll, 2007; Tokowicz et al., 2009).

Translation equivalents may have a one-to-many mapping for different reasons.

1. *Synonymy*: Words that are very close in meaning in a first language (L1) may have a single translation in a second language (L2). For example, English *close* and *shut* both translate to Spanish *cerrar*; Spanish *serpiente* and *culebra* both translate to English *snake*.
2. *Polysemy*: One word in the L1 may have several related meanings, each expressed by a different word in the L2. For example, Spanish *sombra* can be translated to English as either *shade* (of a building or a tree) or *shadow* (cast by a person).
3. *Homography, homophony, and homonymy*: Linguistic “accidents” can cause two unrelated words to be written in the same way. Such forms are called *homographs*. For example, English *bark* can be mapped to Spanish *corteza* (outer layer of a tree) or *ladrido* (sound made by a dog). Homographs may have the same pronunciation, in which case they are also *homophones*, but not necessarily. The English word *bark* has two homographs that are also homophones. In contrast, the English word *row* has two homographs that are not homophones. One homograph of *row* matches Spanish *pelea* (fight) and the other matches Spanish *hiler*a (straight line). When homographs are also homophones, they are called *homonyms*. Homographs may belong to the same grammatical class (i.e., nouns, verbs) and share the same part of speech, as in the examples above, or belong to different parts of speech, as in *bow*-noun (*arco*, belonging to a musical instrument) and *bow*-verb (*inclin*ar).
4. *Morphological ambiguity*: In languages with limited morphology, like English, inflectional and derivational variants of some lexeme may have identical forms. When translated to a morphologically rich language, like Spanish, such variants may be mapped to different forms. For example, the English word *walk* can be translated into several Spanish nouns such as *paseo*, *caminata*, or *vuelta*. In addition, it can be translated into over 20 possible verb forms based on the stem *camina*- (*caminas*, *caminaste*, *caminar*, *caminaba*, *caminaron*, etc.), depending on the tense, aspect, person, and number of the verb.
5. *Semantic discrepancy*: There are cases where multiple translations are a result of the differences in the conceptual–lexical mappings of the two languages. The meaning of the English verb *know*, which covers both knowing facts and knowing people, is carried by two distinct verbs in Spanish, *saber* for the former and *conocer* for the latter. Conversely, the Spanish noun *reloj* covers the concepts denoted by both *clock* and *watch* in English, each of which is a correct translation.

What happens when a bilingual needs to choose among possible translations of a given ambiguous word? In an experimental laboratory setting, when words are presented without supporting context, any one of the translations constitutes a satisfactory response. Bilinguals are also able to provide second and third translations for some words, when requested to do so (Degani et al., 2009). Under these circumstances, translation choice seems to be sensitive to the lexical properties of

the optional translations in the target language: bilinguals are more likely to choose translations that are rated as being more imageable (Prior et al., 2007; Tokowicz et al., 2002). Bilinguals are also sensitive to the degree of form overlap between the translation equivalents in the two languages, and show a preference toward producing a cognate translation, if one exists. Finally, the probability of selecting a specific translation, namely, its conditional probability given the ambiguous word in the source language, is related to the overall lexical frequency of the word in the target language. Higher frequency items tend to be higher probability translations. These lexical variables are also known to influence the speed and accuracy of translating unambiguous words (De Groot, 1992).

Linguistic context can act to reduce lexical ambiguity. As a result, fewer translations may remain appropriate in a given context. This might be the case when multiple translations are a result of homonymy or polysemy, that is, when the word in the source language has more than one meaning. So, in the context of the sentence “John finished drinking and placed his *glass* on the table,” it is clear that the appropriate translation in Spanish is *vaso*, which denotes the drinking vessel and not *vidrio*, which denotes the substance. Similarly, morphological ambiguity can be resolved by the syntactic context. However, when translation ambiguity is a result of synonymy in one of the languages, context might not necessarily act to determine a single correct translation. Thus, semantic context cannot normally act to determine that either “autumn” or “fall” are correct in a specific sentence, because they are mostly interchangeable.

PARALLEL LANGUAGE CORPORA

In recent years, large language corpora have become increasingly available in electronic formats, and have been used for examining various psycholinguistic issues. These corpora may include either spoken or written language, and have different origins, from the transcription of parent–child verbal interactions to Internet news group text. Corpora have been used to develop measures of meaning (Burgess, 1998; Rohde, Gonnerman, & Plaut, 2004), to track vocabulary learning (Landauer & Dumais, 1997), and to develop and test complex theories of child language acquisition (MacWhinney, 2004). Bilingual corpora of child speech (Yip & Matthews, 2007) have been used to follow the simultaneous acquisition of two languages.

In the present study we make use of *parallel corpora* to investigate processes of translation and cross-linguistic mapping. A parallel corpus consists of two (or more) single-language corpora that carry the same meaning, one being a translation of the other. To date, parallel corpora have been used almost exclusively in the domain of human language technology, primarily in statistical machine translation (Brown et al., 1990). Because they provide large samples of meaning equivalents across languages, parallel corpora can be used for training machine translation implementations. Further, the performance of machine translation systems can then be evaluated against a gold standard set by human translators (Lavie, Sagae & Jayaraman, 2004; Papineni, Roukos, Ward, & Zhu, 2001).

However, the knowledge that two large samples of text in two languages carry the same overall meaning is not sufficient to allow further processing. Meaningful

analysis mostly depends on a finer grained alignment between the two languages, by identifying parallels at the sentence or even the word level (Gale & Church, 1993; Och & Ney, 2000). One common solution to the problem of sentence level alignment makes use of the notion that parallel sentences should be of similar lengths in the two languages. This assumption is justified because, in practice, short sentences are normally translated by similar short sentences, and the same holds for long sentences. Thus, sentence length, defined either in words or in characters, can be used to identify parallels in the two languages, without relying on any linguistic or lexical knowledge (Gale & Church, 1993).

The present study makes use of two sentence-aligned parallel corpora in English and Spanish. Matched sentences were searched for translation pairs, and the incidence of different possible translations was tracked.

THE PRESENT STUDY

The work reported here was undertaken with the specific aim of answering three questions. The first goal was to explore the degree of similarity between single word translation choices performed in a laboratory, without supporting linguistic context, and choices made in the translation of large language samples. In a previous study (Prior et al., 2007) we compiled normative data regarding the distribution of possible translations for a sample of words in English and Spanish. In that study, single translations for a list of words in English and Spanish were collected from 80 Bilingual speakers. The participants were native speakers of either English or Spanish, residing in Pittsburgh, PA, who had reached a high level of proficiency in the other language through study and immersion. However, these participants were not professional translators.

Each word was translated by 20 bilinguals, half of them Spanish dominant and half English dominant. Cue words were then classified as being unambiguous if they received an identical translation from all participants or ambiguous, if they received two or more different correct translations. For the ambiguous items, the probability of each possible translation was calculated by dividing the number of people who generated that translation by the total number of correct responses. Thus, for unambiguous words the translation probability was always 1, whereas for ambiguous words the translation probability ranged between .05 and .95.

In the present study, a second measure of the translation probabilities of the responses generated in the norming study was derived from the parallel corpora. This measure reflects the choices made by translators when translating words embedded in a rich linguistic context. As in the norming study, probabilities were calculated separately for translation choices from English to Spanish, and vice versa. Thus, for each occurrence in the parallel corpora of an English cue word, the matching sentence in the Spanish corpus was searched for each of the possible translations generated in the norming study. The number of occurrences of the cue and each of the translation equivalents was tallied, and the probability of each translation was calculated as its number of occurrences divided by the total number of occurrences of the cue word. This process was then completed in reverse for the Spanish to English translation pairs.

We then proceeded to examine the correlation between the laboratory and the corpus-derived measures of translation probability, the former being based on decontextualized translation and the latter on context-embedded translation. Our guiding hypothesis was that the probabilities manifested in the laboratory task reflect bilingual participants' experience of the distribution of the appropriateness of various translation options in the general use of their two languages. Because parallel corpora can be taken as a representative sample of the naturalistic language exposure of bilinguals (albeit a biased representation), we predict that the laboratory measure would correlate with the corpus-derived measures of translation probability.

The theoretical perspective driving this prediction was first articulated by Brunswik (1956) in his theory of ecological validity and extended most recently in approaches such as ecological psychology for perception (Gibson, 1986; Turvey & Shaw, 1979) and the competition model for language processing (MacWhinney, 2008). These ecological perspectives all assume that the organism comes to respond to environmental cues in terms of their relative cue validity or reliability. In the case of translation ambiguities, this means that forms that are most frequently appropriate in contextual real life translation should also be the strongest in translation out of context. This is essentially a process of frequency matching. To the degree that a corpus accurately samples the complete universe of conditional probabilities of usage for a given form, a correlation between preferred translations in and out of context would be expected.

The second issue addressed is the predictive role of several psycholinguistic variables in determining translation choice. Previous work by Prior, MacWhinney, and Kroll (2007) demonstrated that word frequency, imageability, and form overlap all contribute significantly to the probability of a specific translation option being chosen, when no linguistic context was available. In principle, these lexical factors operate quite independently of the first principle of cue validity discussed above. For example, the preferred translation for the Spanish *incidente* to English is *incident* (conditional probability = .95) and not the alternative translation *event*. This disregards that *event* matches the meaning of *incidente* and also has higher lexical frequency in English than *incident* (81 vs. 49 occurrences per million, respectively). This can be attributed to the former translation being a cognate and form overlap being known to exert its influence on translation choice. Thus, different lexical factors may compete in determining translation choice.

We wished to examine whether the various lexical variables play similar roles when translation is embedded within a rich linguistic context. If we found contextual and decontextualized translation probabilities to be correlated, it might be expected that lexical factors have similar influences in both cases. In contrast, when translation is embedded in linguistic context, some of the optional translations are eliminated and freedom of choice is reduced, possibly leading to smaller influences of lexical variables. We therefore treat this issue as more exploratory in nature and as a means of gaining a better understanding of the nature of contextual and decontextualized translation processes.

Finally, depending on the outcomes regarding the first hypothesis and the degree of correlation found between experimental translation probabilities collected in the laboratory and corpus-based translation probabilities derived from parallel

corpora, the door could be open to using corpus-derived measures as research tools for psycholinguistic investigations. Therefore, the present findings will further our understanding of the types of information that can be derived from parallel language corpora and their possible utility as tools for psycholinguistic investigations of bilingual language processing.

METHOD

Materials

Decontextualized translation equivalents. We used the cues and translations generated in a previous norming study (Prior et al., 2007). The norming study included 670 English cue words and 760 Spanish cue words. These induced 1,400 unique English–Spanish translation pairs, and 1,342 unique Spanish–English translation pairs that were included in the present study (the ambiguous cue words generated more than a single translation). The following lexical variables were analyzed as predictors of translation probability: length in letters, part of speech, written frequency in English (Kucera & Francis, 1967) and in Spanish (Pérez, Alameda, & Cuetos, 2003), rated imageability, concreteness, age of acquisition in English (Bird, Franklin, & Howard, 2001; Coltheart, 1981; Wilson, 1988) and in Spanish (taken from LEXESP, Sebastián-Gallés, Martí, Cuetos & Carreiras, 2000; using B-pal, Davis & Perea, 2005), and a rating of the degree of form overlap (cognate rating) of the translation pairs (Prior et al., 2007). As described in the introductory section, the data collected in the norming study allowed us to calculate for each cue word in the source language the probability of each of its possible translations. Because each cue or source word was translated by 20 bilingual speakers of English and Spanish, the probability of each translation response ranges from 1 for translations given by all participants to 0.05 for a translation given by a single participant in the sample.

Parallel corpora. Two parallel corpora in English and Spanish were used in the present study. The first parallel corpus included protocols of European Parliament (EP) sessions from 1996 to 2001 (Koehn, 2005). The corpus consists of 746,274 aligned sentences and a total of approximately 21 million words. Sentence alignment was performed using an implementation of the Gale and Church (1993) algorithm. The original text was in any one of the official languages of the European Union at the time, and was translated by professional translators, so that the protocols exist in their entirety in all 11 official languages of the European Union. Within this corpus, approximately 25% of the text was originally produced in English and 10% in Spanish.¹ The protocols are based on transcripts of the sessions' simultaneous interpretation.

The second parallel corpus we used consisted of United Nations (UN) documentation from 1988 to 1993, which was provided by the Linguistic Data Consortium in English and in Spanish (Graff, 1994). This corpus includes 1,334,502 aligned sentences and a total of 45.4 million words. Original documents were again composed in various languages, with 78% being originally in English or Spanish. Documents were then translated by official UN translators. The sentence

alignment process resulted in the identification of matching sentences in the two languages. It is worth noting that sentence alignment is a heuristic process: often, a sentence in the source language is translated to two or more sentences in the target language, in which case the alignment can go astray. Gale and Church (1993) report a 4% error rate, with better results on 1:1 translations.

That some of the texts were not originally produced in either English or Spanish does not constitute a major obstacle to using these materials to study translation ambiguities between English and Spanish. Despite variability in the original source, the materials we used all allow us to measure the basic alignment between English and Spanish. Further, the quality of corpus analyses is directly linked to the size of the language samples, and we wished to maximize this factor. As will be apparent in the later analyses, a comparison of the two corpora that differ markedly in the percentage of English/Spanish original utterances (35% vs. 78%) reveals very similar patterns, further contributing to our decision to include all of the available data in the current study. Finally, because additional languages served as the original source, it would most likely have the effect of increased variance and “noise” in the data, thereby making it *harder* to find a meaningful relation between the laboratory and the corpus-derived measures of condition translation probabilities.² Thus, if the source materials were consistently derived from English or Spanish originals only, we would expect to find somewhat higher correlations, or at the very least equivalent, between corpus measures and experimental results.

Analyses

The parallel corpora were morphologically analyzed and disambiguated with computerized language analysis (MacWhinney, 2000) using English and Spanish lexicons and grammars as appropriate. This analysis resulted in an identification of the base forms of each lexical item in the corpus, as well as the part of speech (noun, verb). Each corpus was then searched for translation pairs that were identified in the norming studies in both directions. This search was limited by part of speech so that when calculating the translation probabilities for an English word such as *anger*, we distinguished between cases when it was used as a noun and cases when it was used as a verb in the English corpus. Accordingly, in the first case only Spanish noun translations (such as *ira* and *rabia*) were searched and tabulated and in the second case only Spanish verb translations (such as *enojar* and *molestar*) were searched and tabulated. For English to Spanish translation pairs, the English sentences were first searched for the cue words. When an English sentence including the cue word was identified, the aligned Spanish sentence was then searched for each of the possible translation equivalents. The number of occurrences of the cue and each of the translation equivalents was tallied, and the conditional probability of each translation equivalent was calculated. This process was then repeated in reverse for the Spanish to English translation pairs.

Conditional probabilities were calculated separately for the UN and the EP corpora to allow for a possible comparison of written and spoken language and to gauge the importance of the percentage of English/Spanish source language in each corpus. However, to foreshadow the results, the language used in the plenary sessions of the EP is quite formal and thus the data patterns that emerged from the

two corpora were quite similar overall. No striking differences were found between the two corpora, hinting at the conclusion that the percentage of English/Spanish source language material did not have a major influence on the results.

RESULTS

Laboratory and corpus-derived translation probabilities

We will first describe the basic overlap in the items between the laboratory and corpora generated translations. Of the 1,490 English and Spanish cue words presented in the norming study, we identified 1,373 in one or both of the parallel corpora. In addition, of the 2,570 unique translation pairs generated for these items in the norming study, 419 (16.3%) did not appear in either corpus. Further, for the 1,005 noun cue words in both English and Spanish at least one of the possible translations generated in the norming study was identified in the corpus 53.3% of the time. Similarly, for the 491 verb cue words in English and Spanish at least one of the possibilities from the norming study was identified as being present in the corpus 31.2% of the time.

We report several correlation and regression analyses. Because the probabilities of the different translations to a given cue word are not statistically independent, we could not include all of the 2,570 unique translation pairs in the regression analyses. For the ambiguous cue words we instead chose to include only the translation that received the highest probability in the norming study, for a total of 1,315 unique cue–translation pairs (termed dominant translations). In additional nonparametric analyses we also examined the similarities between the lab-generated and corpus-generated translation probabilities of all of the translation pairs.

As a first step, we calculated the raw correlations between all three measures of translation probability. Into this analysis we entered 1,315 unambiguous and dominant translation pairs. The *laboratory probability* is the off-line decontextualized translation choice by participants translating single words in a laboratory setting. These probabilities were found to be highly and significantly correlated with both *corpus-derived measures of translation probability*, the probability derived from the UN ($r = .402$) corpus and that derived from the EP corpus ($r = .414$). When the analysis was limited to the ambiguous items in the sample (i.e., pairs for which the laboratory translation probability was less than 1), we remained with 690 cue–translation pairs, and the correlations were slightly lower, but still significant (see Table 1). This is a first indication that translation choice performed in a laboratory without supporting linguistic context reflects at least to some degree the distribution of appropriate translations in naturalistic language use.

To further examine the basic concordance between the laboratory- and corpus-derived probability distributions of the possible translations, we used two nonparametric measures. First, we determined that the order of the translations by probability was identical across the two measures of probability for 66% of the ambiguous cue words (452 out of 685, in both English and Spanish), whereas by chance this would only be expected to occur for 33% of the sample.

Second, we computed the distance between the probability distributions of the possible translations for each of the 685 ambiguous cue words. To allow for a

Table 1. *Correlation matrix of laboratory, European Parliament (EP)-derived, and United Nations (UN)-derived measures of conditional translation probability for all dominant translations and for ambiguous items only*

		1	2
All items	1. UN translation probability	1.000	
	2. EP translation probability	.777*	1.000
	3. Laboratory translation probability	.402*	.414*
Ambiguous items	1. UN translation probability	1.000	
	2. EP translation probability	.733*	1.000
	3. Laboratory translation probability	.304*	.301*

* $p < .001$.

Table 2. *First language (L1) distance averages for laboratory and corpus-derived translation probability distributions by number of translations*

No. of Trans.	No. of Cue Words	L1 Distance
2	391	0.471*
3	155	0.703*
4	74	0.793*
5	37	1.03*

* $p < .001$.

meaningful comparison between the laboratory and the corpus-derived probabilities we normalized the latter, by transforming them into probability vectors, namely, expressed the probability of each translation as the proportion of the times it was encountered divided by the total number of times that any of the translations was encountered (eliminating the cases when the cue word was detected but none of the translations were identified). We then computed for each cue word the “L1 distance” between the probability distribution of the translations generated in the lab and that derived from the corpora. L1 distance was computed as follows: given two probability vectors \mathbf{V} and \mathbf{U} of the same length n , the L1 distance is the following sum:

$$\sum_{i=1}^n |\mathbf{V}_i - \mathbf{U}_i|.$$

For probability vectors, this sum is between 0 (identical vectors) and 2 (maximally different vectors). The average L1 distances for the ambiguous cue words, separated by number of translations, are presented in Table 2. There were an additional 14 cue words with six different translations and 2 cue words with seven different translations, but these numbers were too small to allow for a meaningful analysis of the results.

The statistical significance of these values was assessed by running a permutation test, because probability distributions violate the assumptions of traditional statistic tests. Thus, 1,000 random permutations of the corpus-derived probabilities were created separately for the cue words of each number of translations, and their L1 distance from the lab-generated probabilities were computed. We then determined the percentile of the original distance in this list of 1,000 random distances to arrive at the level of significance of our result (p value).³ As can be seen in Table 2, all of the L1 distances between the laboratory- and the corpus-derived probabilities were highly significant, because they were much closer than would be expected by chance. This finding again points to a basic alignment and similarity between the translations produced without context in a laboratory setting, and those derived from parallel language corpora.

Our previous study of decontextualized behavioral translation choice (Prior et al., 2007) identified several lexical variables that are significant predictors of this measure of translation probability. Therefore, the correlations reported above between the behavioral and corpus-derived translation probabilities might simply be a result of all three measures being driven by the influence of lexical variables. Thus, in the following analysis we examine the relation between the behavioral and corpus-derived probability measures, after statistically removing the influence of other variables.

We performed a hierarchical regression analysis to probe the role of lexical variables in order to avoid concerns associated with standard stepwise regression (e.g., Juhasz, 2005). In addition, we examined the semipartial correlation coefficients to address problems associated with multicollinearity (Cohen, Cohen, West, & Aiken, 2003).

The dependent variable was the laboratory, single-word translation probability. Each translation pair had a single value, which reflected the percentage of participants that gave the specific translation to the specific cue word or the conditional probability of a specific target translation given a specific cue. Once again, only unambiguous cue words and the dominant translations of the ambiguous cue words were included in this analysis to avoid problems associated with dependencies within the data set. The predictor variables were properties of the translation word, including frequency and measures of semantic salience (familiarity, imageability, and concreteness). These three measures were highly correlated, and different combinations were available for various stimuli. We therefore chose to enter all three measures as a single step in the regression analysis. The form overlap (cognate rating) between the translation and cue words was also used as a predictor. A previous analysis of the laboratory norms (Prior et al., 2007) indicated that these factors significantly predict translation choice, so all were included in the model before examining the relation of corpus-derived probability to the laboratory probability. Finally, we entered the corpus-derived probabilities as predictors in the model to assess the degree to which they correlated with the laboratory probability after the variance accounted for by the previous lexical variables had been partialled out of the model.

We entered 732 translation pairs in both directions of translation (English to Spanish and Spanish to English) into the analysis. The pairs included in the analysis were those for which lexical properties were available (specifically, semantic

variables and/or cognate ratings were not available for the entire set of translation pairs). Table 3 provides the beta and semipartial correlation coefficient values for the significant steps of the analysis.

Translation word frequency and length were entered in the first step of the model and were found to significantly predict translation probability, $R^2 = .01$, $F(2, 729) = 3.5$, $p < .05$, such that words with a higher frequency in the language were more likely to be selected as translations and had higher conditional probabilities. The semantic salience of the translation word was entered in the second step of the model and predicted translation probability, $\Delta R^2 = .012$, $F(3, 726) = 2.8$, $p < .05$. Specifically, words high on the measures of semantic salience were more likely to be given as translations. In the third step the cognate rating of the translation pair contributed significantly to the model, $\Delta R^2 = .061$, $F(1, 725) = 48.3$, $p < .001$, because translation options were chosen more often when their form was similar to that of the cue word. We then entered the UN corpus-derived probability, which significantly predicted translation probability, $\Delta R^2 = .063$, $F(1, 724) = 53.8$, $p < .001$. In the fourth and last step, we entered the EP corpus-derived probability, which accounted for additional variance in translation probability, even after the effects of all previous variables had been removed, $\Delta R^2 = .014$, $F(1, 723) = 12.5$, $p < .001$. A combination of all predictor variables explained 16% of the variance in the probability of translations based on laboratory translations.

Thus, both corpus-derived measures of translation probability remain significant and the strongest predictors of the probability of decontextualized translation performance in the laboratory, even after removing the variance accounted for by lexical factors. This finding leads to the conclusion that there is moderate similarity between the contextualized and decontextualized translation tasks, and that they tap at least some of the same mechanisms. However, we should note that overall the weights of the predictor variables entered into the regression model were not very high, and there remains a large percent of unaccounted variance.

Predictors of translation choice in context

The final question we sought to address is the relative importance of the different lexical variables in influencing translation choice with and without linguistic context. We conducted three separate hierarchical regressions, one for each of the three translation probability measures, again limited to unambiguous cue words and dominant translations of ambiguous cued words, for a total of 732 pairs. The predictor variables entered into the model were identical across the three analyses: translation word length and frequency were entered on the first step, measures of semantic salience (concreteness, familiarity, and imageability) were entered on the second step, and form overlap (cognate rating) was entered on the third and final step of these analyses. The predictor variables all contributed significantly to the explained variance in all three analyses (all $ps < .01$), but there was some variability in the magnitude of overall explained variance ($R^2 = .08$ for the laboratory translation probabilities, $R^2 = .128$ for the UN corpus-derived translation probabilities, and $R^2 = .142$ for the EP corpus-derived translation probabilities). Figure 1 illustrates the relative weight of each variable across the three models.

Table 3. Summary of the hierarchical regression analysis for laboratory translation probability

Variable	Model 1		Model 2		Model 3		Model 4		Model 5	
	β	r_{sp}	β	r_{sp}	β	r_{sp}	β	r_{sp}	β	r_{sp}
Frequency	0.064*	.063	0.092*	.086	0.085*	.080	0.039	.037	0.014	.013
Length	-0.062	-.061	-0.031	-.029	-0.120**	-.107	-0.159**	-.140	-0.173**	-.151
Concreteness			-0.025	-.014	-0.027	-.015	-0.079	-.043	-0.098	-.053
Familiarity			-0.01	-.01	0.013	.012	0.006	.006	0.006	.006
Imageability			0.135*	.072	0.138*	.074	0.134*	.072	0.137*	.073
Cognate rating					0.264**	.247	0.199**	.180	0.198**	.180
UN probability							0.270**	.252	0.125*	.078
EP probability									0.195**	.120

Note: r_{sp} , semipartial correlation; UN, United Nations; EP, European Parliament.

* $p < .05$. ** $p < .01$.

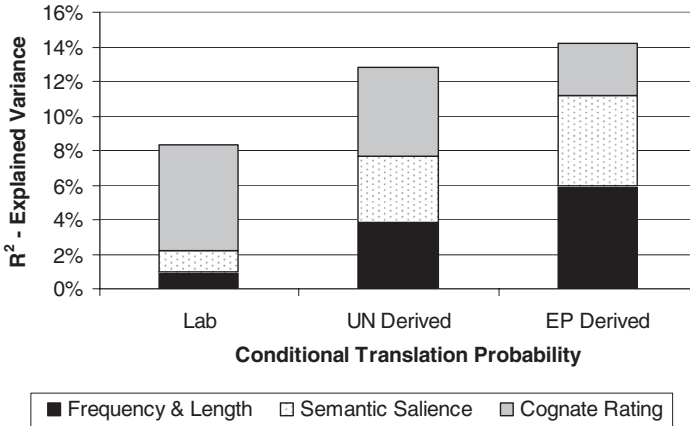


Figure 1. The explained variance in laboratory- and corpus-derived conditional translation probabilities attributed to predictor variables. UN, United Nations; EP, European Parliament.

The most striking difference is the greater influence of form overlap on the laboratory translation probabilities than on the corpus-derived probabilities, both relatively as the percentage of explained variance and absolutely in the magnitude itself. As mentioned earlier, this may be because, when translation occurs within a meaningful context, some of the translations are eliminated as viable possibilities. Thus, under these conditions there is less of an actual choice and therefore possibly a weaker influence of lexical properties of the various translations, especially at the form level, on selection. This issue will be taken up in the Discussion Section.

DISCUSSION

To the best of our knowledge, the work reported here is the first attempt to compare the performance of bilingual participants translating single words, presented without linguistic context in a laboratory setting, with the conditional probabilities of alternative translations as they are derived from large parallel language corpora. The goals of this work were to provide a preliminary illustration of the similarities and differences between these two cases, and to evaluate the influence of meaningful context on translation choice. The present investigation was also intended to probe the utility of large parallel language corpora as a tool in conducting psycholinguistic research. We will discuss each of these issues in turn and conclude with several suggestions for future work in this promising new avenue of research.

The results show that laboratory, decontextualized translation probabilities and corpus-derived translation probabilities demonstrate moderate convergence. This is evidenced by the direct correlations among the measures, the significant findings of the nonparametric comparison of the probability vectors, and the results of the hierarchical regression showing the residual concordance between corpus-derived and laboratory translation probabilities after partialing out the variance

assigned to lexical properties of the translation word. These findings lend support to the notion that bilinguals' single-word translation choices seem to reflect the distribution of the appropriateness of various translation options, if we consider parallel corpora as providing a good approximation of the naturalistic language exposure of bilinguals, specifically the distribution of the different translation options. This finding confirms the predictions of theories of ecological validity and supports a process of frequency matching across different contexts.

Note, however, that the correlations that we report, although statistically significant because of the large number of observations, are not very high. This finding might be attributed to several factors: first, contextualized and decontextualized translations differ in several respects, specifically, certain translation choices might be eliminated by linguistic context, thus limiting the similarity between the two cases. Second, the bilinguals providing the translations in the parallel corpora were professional translators, whereas the laboratory data were collected from mostly unbalanced nonprofessional bilinguals. Again, these differences in population might have acted to reduce the similarity between the two data sources. Third, unsupervised automated processing of large language corpora inevitably leads to noisy and imperfect results, which might limit our ability to find higher correlations between the corpus-derived measures and the more precise laboratory-derived probabilities.

The correlation between the two corpus-derived measures ($r = .777$) was higher than that between each of the corpus derived measures and the laboratory probability. This might reflect the inherent difference between decontextualized and contextualized translation choice. Alternatively, the high correlation between the corpora may reflect that they both rely on the performance of professional translators or that they are fairly similar to one another (UN and EP) in terms of genre and context. Of course, a fuller sampling of corpus types would provide a better test of the match of cues in and out of context.

The examination of the roles played by the lexical variables in predicting translation probabilities sheds light on the difference between contextualized and decontextualized translation choice. As expounded in the introductory section, supporting context may at times eliminate one or more possible translations, narrowing the selection available to the translator. It is interesting that the relative importance of the lexical variables that we examined as predictors of translation probability varied between the two cases. Thus, it seems that form similarity (cognate rating) is more important in decontextualized translation, a finding that probably reflects the faster and more robust activation of form-related translations (e.g., de Groot, 1992). In the case of translation in context, either the cognate translation was rendered incorrect by the context, or on occasion the activation of alternative noncognate translation options received enhanced support from other words in the context and was thus robust enough to successfully compete for selection (Schwartz & Kroll, 2006, demonstrate how context can reduced cognate effects though in a different setting).

The conditional translation probabilities with and without supporting linguistic context were found to be fairly similar overall. This finding leads to the conclusion that the distribution of translations to single words produced in a laboratory setting reflects the life experience of bilinguals regarding the characteristic distributions

of the various alternative translations to some degree, but it by no means matches it perfectly.

FUTURE RESEARCH AND CONCLUSION

In light of this conclusion, we see the present study as a favorable demonstration of the utility of parallel language corpora for psycholinguistic investigations. Specifically, the current use of parallel corpora allowed us a preliminary comparison of contextualized and decontextualized translation choice and revealed important differences between the two cases. A similar comparison conducted with experimental methods would most likely have covered a smaller scope of items and at the same time would have required greater investments of time and resources.

A more general question, however, centers on the wider utility of such corpus-derived measures for studying bilingual language processing. The current study used an existing laboratory-generated list of translation pairs to search for distributional probabilities. Because we found a moderate correlation between the corpus-derived and behavioral measures of conditional translation probability, it seems premature to conclude at this point that the two methods provide interchangeable estimates of translation probability. The question remains as to which method might be preferred as an accurate reflection of the psychologically valid hierarchy of translation options. However, this issue must await further research comparing the utility of laboratory- and corpus-derived probability as predictors of bilingual performance in a variety of linguistic tasks (e.g., translation production or recognition; Prior et al., 2006).

Future research might also examine the degree of accordance between laboratory and corpus-derived translation probabilities in the different types of translation ambiguity described in the introduction. Thus, it might be that the similarity between the two types of measures would be higher for translation ambiguity resulting from synonymy in one of the languages, because a bilingual speaker must choose a translation in a similar manner whether linguistic context is available or not. However, corpus- and laboratory-derived measures of probability might exhibit a lower degree of convergence in cases of translation ambiguity resulting from part of speech ambiguity in one of the languages, because linguistic context eliminates any choice of translation under these circumstances.

Finally, the study presented in this paper is only unidirectional, in the sense that it makes use of parallel language corpora to validate behavioral measures, thereby contributing to research in psycholinguistics; but it does not use insights from psycholinguistics to advance computational linguistic applications. One direction for such a contribution might be the area of machine translation evaluation: researchers in computational linguistics are searching for automatic measures that are indications of quality (machine) translation, and psycholinguistic research in the area of bilingualism may provide insights for such investigations. Thus, future work might examine the possible translations generated automatically by machine translation algorithms, to achieve a better understanding of their alignment with the performance of human bilingual translators. A second avenue of research centers on finer distinctions of the parallel-corpora being used, for example, the original language of utterance and the direction of translation and possible influences of

these variables on the quality of the translation probabilities generated. These are but two demonstrations of the wide range of issues that might benefit from work combining parallel language corpora and methods from machine translation with psycholinguistic investigations of bilingualism.

ACKNOWLEDGMENTS

The authors thank Leonid Spektor and Giora Unger for programming assistance, Dr. Yuval Nov for statistical assistance, and Wouter Duyck and two anonymous reviewers for helpful comments on a previous version of the manuscript. The first author was provided funding by postdoctoral support (NRSA F32HD049255). This work was partially supported by NSF Grant SBE-0354420 to the Pittsburgh Science of Learning Center and by Grant 2007241 from the United States–Israel Binational Science Foundation.

NOTES

1. Unfortunately, source-language tags in this corpus are not complete, rendering it difficult to establish with certainty the original language of any specific utterance. Thus, the above percentages are only estimates.
2. It is also possible that, on occasion, translations to both English and Spanish from a third language may yield higher seeming similarity with the behavioral data. For example, when the word “leger” appears in Dutch in the corpus and is translated into English as “army” and into Spanish as “armada,” it gives rise to form overlap in the translations, although in each translation there was no form similarity to the Dutch original word. (We thank Wouter Duyck for bringing this possibility to our attention.) However, it is not likely that such occasions would outnumber the cases where translation from a third source language would result in *less* alignment between the corpus and the laboratory data. Thus, in the worst case these two types of occurrences would cancel each other out.
3. The permutation test was also performed with 10,000 random repetitions. However, the results were virtually identical, so we do not report them in detail.

REFERENCES

- Altarriba, J., & Gianico, J. L. (2003). Lexical ambiguity resolution across languages: A theoretical and empirical review. *Experimental Psychology*, *50*, 159–170.
- Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, and Computers*, *33*, 73–79.
- Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., et al. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*, 79–85.
- Brunswik, E. (1956). *Perception and representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.
- Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, and Computers*, *30*, 188–198.
- Clifton, C., Frazier, L., & Rayner, K. (Eds.). (1994). *Perspectives on sentence processing*. Hillsdale, NJ: Erlbaum.

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.
- Coltheart, M. (1981). *MRC psycholinguistic database user manual: Version 1*. London: Birkbeck College.
- Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, 37, 665–671.
- Degani, T., Prior, A., & Tokowicz, N. (2009). *Bidirectional semantic transfer: The effect of sharing a translation*. Unpublished manuscript.
- de Groot, A. M. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1001–1018.
- Gale, W., & Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19, 75–102.
- Gernsbacher, M. A., Robertson, R. R. W., & Werner, N. K. (2001). The costs and benefits of meaning. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp 119–137). Washington, DC: American Psychological Association.
- Gibson, J. J. (1986) *The ecological approach to visual perception*. Hillsdale, NJ: Erlbaum. (Original work published 1979)
- Gorfien, D. S. (Ed.). (2002) *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity*. Washington, DC: American Psychological Association.
- Graff, D. (1994). *UN parallel text (complete)*. Philadelphia, PA: Linguistic Data Consortium.
- Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, 24, 617–637.
- Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, 131, 684–712.
- Kambe, G., Rayner, K., & Duffy, S. A. (2001). Global context effects on processing lexically ambiguous words: Evidence from eye fixations. *Memory & Cognition*, 29, 363–372.
- Kellas, G., Ferraro, F. R., & Simpson, G. B. (1988). Lexical ambiguity and the timecourse of attentional allocation in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 601–609.
- Koehn, P. (2005). *Europarl: A parallel corpus for statistical machine translation*. Paper presented at the Machine Translation Summit, Phuket, Thailand.
- Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.
- Lavie, A., Sagae, K., & Jayaraman, S. (2004). *The significance of recall in automatic metrics for MT evaluation*. Paper presented at the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004), Washington, DC.
- Lyons, J. (1995). *Linguistic semantics*. New York: Cambridge University Press.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.) Mahwah, NJ: Erlbaum.
- MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language*, 31, 883–914.
- MacWhinney, B. (2008). A unified model. In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition*. Mahwah, NJ: Erlbaum.
- Och, F. J., & Ney, H. (2000). *Improved statistical alignment models*. Paper presented at the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2001). *BLEU: A method for automatic evaluation of machine translation*. Paper presented at the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA.
- Pérez, M. A., Alameda, J. R., & Cuetos, F. (2003). Frecuencia, longitud y vecindad ortográfica de las palabras de 3 a 16 letras del diccionario de la lengua Española (RAE, 1992). *Revista Electrónica de Metodología Aplicada*, 8, 1–10.
- Prior, A., Kroll, J. F. & MacWhinney, B. (2006). *The role of translation probability and word class in two translation tasks*. Poster presented at the 47th Annual Meeting of The Psychonomic Society, Houston, TX.

- Prior, A., MacWhinney, B., & Kroll, J. F. (2007). Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods*, 39, 1029–1038.
- Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2004). *An improved method for deriving word meaning from lexical co-occurrence*. Retrieved January 30, 2008, from <http://dlt4.mit.edu/~dr/COALS/Coals.pdf>
- Schwarz, A., & Kroll, J. F. (2006). Bilingual lexical activation in sentence context. *Journal of Memory and Language*, 55, 197–212.
- Sebastián-Gallés, N., Martí, M. A., Cuetos, F., & Carreiras, M. (2000). *LEXESP: Léxico informatizado del español*. Barcelona: Edicions de la Universitat de Barcelona.
- Tokowicz, N., & Kroll, J. F. (2007). Number of meanings and concreteness: Consequences of ambiguity within and across languages. *Language and Cognitive Processes*, 22, 727–779.
- 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, and Computers*, 34, 435–451.
- Tokowicz, N., Prior, A., & Kroll, J. F. (2009). *Bilingual speech production depends on translation ambiguity*. Unpublished manuscript.
- Turvey, M. T., & Shaw, R. E. (1979). The primacy of perceiving: An ecological reformulation of perception for understanding memory. In L. G. Nilsson (Ed.), *Perspectives on memory research*. Hillsdale, NJ: Erlbaum.
- Wilson, M. D. (1988). The MRC Psycholinguistic Database: Machine readable dictionary, version 2. *Behavior Research Methods, Instruments, and Computers*, 20, 6–11.
- Yip, V., & Matthews, S. (2007). *The bilingual child: Early development and language contact*. Cambridge: Cambridge University Press.