Karen Sullivan and Javier Valenzuela

12 Comparing word sense distinctions with bilingual comparable corpora: A pilot study of adjectives in English and Spanish

1 Introduction

Amidst the recent surge of interest in the applications of corpora in Cognitive Linguistics, and the wide range of methodologies now available (e.g. Gries and Stefanowitsch 2006), the types of corpora employed and their applications in translation and L2 instruction still has the potential for further expansion. The current chapter draws attention to one of the areas in which the field has the potential for growth, and suggests the gains it may have to offer translation and Second Language Acquisition (SLA). These points are then illustrated with examples from a pilot corpus study conducted on a set of adjectives in English and Spanish.

The pilot study is suggestive of the potential role of bilingual comparable corpora, that is, sets of non-parallel matched monolingual corpora, each in a different language, as an approach to comparative lexical semantics. This method is argued to hold several advantages over the more traditional lexical studies employing monolingual corpora and bilingual parallel corpora. The study sorts a set of adjective senses from an English corpus and a Spanish corpus on the basis of distributional variables (such as whether adjectives occur in predicative or attributive position), which allows a detailed analysis of the relatedness of the senses of each word in each language. The networks of related senses in the two languages can then be compared, and it can be seen which senses are similar or different in the two languages. This can serve as a guide for L2 students, translators and lexicographers interested in finding the best approximation for a given source-language meaning in a target language.

2 Types of corpora and their applications

Many types of corpora are currently available, several of which have been, or could be, employed in studies with implications for SLA or translator training.

Karen Sullivan, School of Languages and Cultures
Javier Valenzuela, University of Murcia

DOI 10.1515/9781501505492-013

To date, relatively few of the available types of corpora have been taken advantage of in translation studies. Granger (2003: 21) provides a summary of available corpus types and those that have been employed in contrastive linguistics and translation studies, a list updated to some extent in Marzo et al. (2010). The circumscribed range of corpus usage can be attributed in part to the relative newness of the field. For instance, the use of corpora in translation studies was not suggested until Baker (1995). The integration of corpora in SLA is slightly older, dating back at least to Johns and King (1991).

Types of corpora with past or potential applications within translation studies and SLA include (bilingual) parallel corpora, monolingual comparable corpora, and bilingual comparable corpora. We will briefly mention some of the relevant work done with these types of corpora, and the advantages and disadvantages of each type for SLA applications.

Parallel corpora are probably the most widely used corpora in translation studies. These are “corpora that contain a series of source texts aligned with their corresponding translations” (Malmkjaer 1998: 539, quoted in Granger 2003: 20). Translated texts without their source texts may also be used. These may be termed *translation* (or *translational*) *corpora* (Baker 1999). Parallel corpora may be employed either to study translation itself or as a basis for comparing the structure of two languages (Mason 2001). However, as Mason (2001) notes, parallel corpora may give deceptive results for research comparing linguistic structure, as their target language material will differ from non-translated data from that language. It may be influenced by the source language, or may be subject to artifacts stemming from the process of translation itself (see Olohan 2004: 26–28 for a discussion of examples). Some of these effects can be controlled for by employing two parallel corpora, one translated from language A to B and one translated from B to A (Johansson 1998; referred to as *bilingual parallel corpora* in Zanettin 1998).

In translation studies, artifacts arising from translation are an important focus of study in their own right. Largely for this reason, the use of monolingual comparable corpora is on the rise in translation studies (see Olohan 2004, Chapter 4). These corpora consist of translated and non-translated texts in a single language, examined “in order to explore how text produced in relative freedom from an individual script in another language differs from text produced under the normal conditions which pertain in translation, where a fully developed and coherent text exists in language A and requires recoding in language B” (Baker 1993: 233, quoted in Olohan 2004: 36). These corpora allow researchers to identify features of translated texts, some of which are outlined in Baker (1996). Although parallel corpora may be useful for understanding or conducting translation, they appear to be less immediately useful in SLA. For

SLA students, it may be more productive to be exposed to non-translated, rather than translated, data from the L2 (Johansson 2007).

Monolingual corpora of learner data are probably the most frequent type of corpus employed in SLA studies. The International Corpus of Learner English (initiated and directed by Sylviane Granger), for example, collects essays from 2nd- and 3rd-year university students studying English, representing sixteen L1s. Several studies have used this corpus to compare native speaker data with learner data from speakers of an L1, or a set of L1s, in order to draw attention to L1 transfer or interference effects, such as the frequency of the use of grammatical constructions (as in Valenzuela and Rojo 2008) or of particular lexical items (as in Ringbom 1998).[1] These results can be integrated into SLA instruction to help students avoid typical learner patterns of overuse or underuse.

Despite the gains made with monolingual comparable corpora, relatively little research has so far been conducted using bilingual comparable corpora. When these corpora have been employed, they have focused on the study of specific genres, such as printed public notices in English and German (Schäffner 1998) or medical research articles (Williams 2010), or for the study of collocational frequency (Noël and Colleman 2010; Zanettin 1998). Bilingual comparable corpora have been proposed for use in monolingual Word Sense Disambiguation (WSD) (Kaji 2003) – that is, “translation equivalents” in one language can be used to define the various senses of a given word in a different language – but this procedure has little direct application for translator training and even less for SLA. It seems evident that comparable data from multiple languages are necessary if L2 learners and translators are to use corpus data to find the nearest equivalent, in a target language, for a lexical item in a source language. This suggests that bilingual comparable corpora as well as monolingual corpora are a potentially valuable resource for SLA and translation studies addressing lexical semantics.

3 Options in corpus analysis

Besides the choice of corpus type, several other decisions must be made by researchers interested in employing corpora in SLA studies. A study may focus on word senses, words, or phrases in the corpora, for example. It is also necessary to select the parameters that are taken into account, such as the words

[1] Numerous studies of this type are collected in an online Learner bibliography by the Centre for English Corpus Linguistics. See: http://sites-test.uclouvain.be/cecl/projects/learner_corpus_bibliography.html.

or syntax that co-occur with the items of interest. This section will address the choice of parameters, and the selection of words versus senses, selected in previous research and available for future investigations relevant to applications in SLA and translation. The discussion will focus on hierarchical cluster analyses (HCA), an exploratory data grouping method that has proven its usefulness in studies of monolingual polysemy (Gries and Divjak 2009; Sullivan 2012) and translational corpora (Jenset and Hareide 2013; Ke 2012). HCA also has the advantage that its results can be assessed with bootstrapping, a method by which data are shuffled and then re-clustered to statistically evaluate the validity of the clusters (Divjak 2010; Glynn 2010; Suzuki and Shimodaira 2011).

Perhaps the most readily available variables that can be employed in clustering consist of the items' collocations, that is, other items that tend to occur in proximity to the items in question (an approach adopted in Kaji 2003, and in monolingual corpus studies including Gibbs and Matlock 2001, and Kishner and Gibbs 1996). However, there are several reasons why the use of predominantly syntactic variables may allow for a more accurate impression of cross-linguistic equivalence than the more traditional reliance on collocations. Collocations tend to be highly language-specific, which is the primary reason that translators-in-training need to be exposed to the concordances of items in the target language, because they are likely to differ from the source language (see Hadley 2002 for discussion). This trait, which renders concordances a productive part of translator training, makes them less useful in comparisons between languages, since collocations are likely to often be too different between the two languages for meaningful comparisons to be made. Of course, languages have different syntactic structures as well as different concordance patterns, but we argue it is more revealing to compare syntactic structures between languages (i.e. adverbial modification) than to find analogous collocates (i.e. co-occurrence of *skin* with *soft* in English and that of *piel* 'skin' with *suave* 'soft/smooth' in Spanish). Additionally, individual collocates can distort a cluster analysis, particularly one comparing word senses (Gries and Divjak 2009), and reliance solely on collocates may do little to reveal ties between senses, since collocates can co-occur with only one sense.

The variables that are considered can be selected and manipulated in many different ways. Gries and Divjak (2009) employ morphosyntactic or semantic variables in their analysis which they call "ID tags". Though clustering analyses based primarily on syntactic data present several advantages, it must be acknowledged that tagging syntactic IDs is currently far more labor intensive than collecting collocations. Future advances in automatic corpus tagging could simplify the ID-tagging process, allowing even long-distance syntactic relations and large-scale syntactic structures to be identified and tagged automatically. Improved availability of comparable corpora in multiple languages with any degree of tagging beyond POS-tagging (for example, tagging of nouns and/or adjectives for plurality) would reduce the number of variable values that must be manually identified and assigned as ID tags in a given study.

In addition to the type of data chosen for annotation and consideration, corpus studies dealing with polysemy can choose whether to compare the various senses of individual words, thereby charting the structure of polysemy networks, or to ignore the different senses of each item and compare instead different words with each other, mapping the relatedness of "near synonyms" (Gries 2008; Divjak 2010) and discovering which words in a semantically related set are most similar based on their distribution and syntactic, semantic, and other properties.

Examining the concordances of one item at a time, in one language at a time, offers applications for translation and potentially for SLA (Hadley 2002). However, we suggest that there are also advantages to analyzing the connections between senses of various items, and in comparing these networks of senses in multiple languages. A speaker or translator will typically need to find the best approximation for one sense of a source language item in a target language. Examining concordances of a single item in the target language may give learners and translators a general feel for the usage range of a particular item, but may be less directly applicable to the everyday task of word translation than a corpus-based tool that compares multiple words and senses in both the target and source languages. This can be accomplished through the use of clustering of word senses in bilingual comparable corpora.

As seen in the previous section, most clustering studies – and all of those employing the methodological choices described above – have been employed with monolingual corpora (Divjak and Gries 2008; Gries 2006; Sullivan 2012). Which of the choices explored above are most compatible with the use of bilingual corpora? It has already been suggested that syntactic, as opposed to collocational, data, are more appropriate to cross-linguistic studies. In terms of the choice between "near synonyms" and word senses, it seems that the latter may prove more useful. No L2 speaker or translator would want to always equate one specific lexical item in the source language, such as English *soft*, with one specific item in the target language, such as Spanish *suave*. It seems more realistic that one specific sense of *soft* might, indeed, always be best translated as *suave*. There may in turn be a specific sense of *suave* that can always be felicitously translated into English as *soft*. It may be most useful in SLA and translation, therefore, to look for similarity between word senses rather than between words. For this, studies of bilingual comparable corpora with clustering of word senses may prove the most productive choice.

4 Pilot study: method

As a preliminary assessment of the effectiveness of word sense clustering using bilingual comparable corpora, our sample study collected 300 examples of each of four adjectives: English *soft* and *smooth* and Spanish *suave* and *blando*. English examples were randomly selected from all instances of the lemmas *soft* and *smooth* tagged as adjectives in the British National Corpus, and Spanish examples were randomly selected from all instances of the lemmas *suave* and *blando* tagged as adjectives in the Corpus del Español. These examples were assigned ID tags and analyzed in context to identify the sense instantiated by each corpus example. Identification of senses and annotation of ID tags for *blando* and *suave* was assisted by a team of undergraduate native speakers of Spanish. Senses in both English and Spanish were chosen by consensus among the authors and the undergraduate team, and the choice of which senses should be considered as separate was continually reassessed as data were analyzed. The senses of each word were clustered based on the ID tags.

As discussed, the ID tags in our study were primarily syntactic. In addition to the reasons discussed above for using syntactic versus collocational tags, we chose syntactic over semantic tags because we aimed to make the annotation as objective and unbiased as possible. We found evaluations of syntactic features to be more consistent across annotators than semantic judgments.

Given the preliminary nature of the study, only nine ID tags were included for each language. Eight were the same for both languages and one tag was used for each language that was not applicable for the other. For both English and Spanish, ID tags were assigned for the type of construction in which the adjective appeared (attributive, predicative or resultative); modification of the adjective by one or more adverbs; presence of other adjectives modifying the same noun; presence of a PP complement on the modified NP; presence of the NP within a PP; whether the modifiee was expressed anaphorically; whether the modified noun was a mass or count noun, and its number (singular or plural). English ID tags included tough-movement, which does not exist in Spanish, and Spanish ID tags included pre- or post-nominal position of the adjective, which is far more variable in Spanish than in English. Adjective gender was not included as an ID tag in Spanish because it is largely semantically arbitrary, an observation confirmed by the apparently randomizing effect its inclusion had on the resultant cluster analysis. We intend to expand the number of ID tags in subsequent studies on texture adjectives in English and Spanish, though we will continue to emphasize syntactic variables.

In all clustering studies of sense relatedness, no matter how objective the ID tags, sense labeling itself is subjective to some degree. The application of criteria such as those of the principled-polysemy approach (Evans 2005: 41; Tyler and Evans 2001; discussed in Gries and Divjak 2009) can make the process of distinguishing senses less arbitrary, but total objectivity or agreement between all researchers is almost impossible. The main problem for distinguishing senses is granularity (i.e. at which level similar senses should be distinguished). Granularity was resolved partly based on frequency: senses with three or fewer examples were preferentially grouped with others rather than put in the "other" category; and also on classification accuracy. An overly high granularity is unproblematic when there are an adequate number of examples of each sense, because similar senses cluster together. High granularity only becomes truly problematic when there are few examples of each sense – as was occasionally the case in our small-scale study – because a small set of examples cannot be expected to be representative of the contexts in which a given sense occurs, leading to inaccurate clustering.

In the procedure used here, ID tags were annotated in columns in an Excel file in the format shown in Table 1. Note that the "sense" label is purely for convenience, and that these one-word labels are not taken in any way to be descriptions or definitions of the senses, but merely as labels for senses which are treated as distinct from other senses. We argue that it is neither necessary nor desirable for SLA or translator training to define word senses using a one-word "synonym" in either the same language or in a different language (see Kaji 2003), as is common practice in WSD. These "synonyms" are a necessity in machine translation, but for human corpus users they are less useful than more exact definitions. It is convenient to have a short label for word senses, especially as a shorthand in annotating and as inputs to analysis software, but for human audiences these labels can be accompanied by in-depth explanations of the nuances of each particular sense, the semantic range of the sense, and its boundaries with other senses. These explanations should not be *a priori*, but should be based on observations and examples from the corpus itself. The semantics of any word sense are likely to be complex, and we see no advantage to artificially constraining or simplifying the descriptions of senses.

The ID tags for each item form a behavioral profile vector (the set of variables the values of which are represented by ID tags), which can be inputted into a hierarchical agglomerative cluster (HAC) analysis. This can be done in a number of ways. Here, we are following the procedure described in Gries and Divjak (2009), using the Behavioral Profiles (BP) program for R written by Gries (2008). Among other functions, this script performs a HAC that sorts the examples on the basis of their behavioral profiles. This results in a tree-like clustering diagram, called a "dendrogram", in which similar senses are clustered. The current

Table 1: Sample senses and ID tags of *soft*.*

Sense	Syn.	PP comp	In PP	Adv.	Other adjs.	Count N?	Number of N	
Consistency	a	yes	no	yes	no	yes	s	for the table, continues on to an arugula salad with dates and a meltingly **soft** pork shank with rye gnocchi and sauerkraut.
Flexible	a	no	no	no	no	yes	pl	are very uncommon in snowboarding. And at the same time, you're wearing **soft** boots that you can run around in,
Force	a	no	yes	no	no	yes	s	his eyes brushing my neck, my jaw, and my mouth with a **soft** force, and then resting deep inside my eyes.
Gentle	p	no	no	no	yes	yes	s	guy is about 5' 5", 130 pounds, sweetheart, intelligent, **soft** and gentle. Not someone who's prone to be a tough guy.
Humanities	a	no	yes	no	no	yes	pl	business, engineering, and the like – has clearly decided to write off the **soft** disciplines, namely the humanities and the arts.
Indirect	a	no	no	no	no	no	n	Well put. . . . we need the most severe changes to restrict the **soft** money which, as I say a couple of times, is a blight on. . .
Noise	p	no	no	yes	no	yes	s	it as best as he can. The sound, however, is still understandably **soft**. # STARKS waits and then reaches for the knob on. . .

* a = attributive; p = predicative; s = singular; pl = plural; n = not applicable

study utilized the Canberra similarity metric to make the best use of the relatively small data set, and used the Ward amalgamation strategy, in order to encourage clusters of an easily interpretable size.

The BP script for R also incorporates the pvclust script (Suzuki and Shimodaira 2011) that assesses the reliability of the HAC analysis with bootstrap resampling. That is, the instances of each word or sense are repeatedly shuffled and then re-clustered. In the BP script, data are re-clustered 10,000 times. The results of this resampling are reported as Approximately Unbiased (AU) *p*-values, which are assigned to each cluster and which report how often the cluster emerged in the resamplings. For example, an AU *p*-value of 70% would mean that a particular cluster occurred in 70% of the resamplings. The apparent cluster is less likely to be due to chance than a cluster with a lower AU *p*-value, and more likely to be a chance occurrence than a cluster with a higher AU *p*-value.

5 Pilot study: results and analysis

The outcome of the HAC analysis can be represented in dendrograms such as Figures 1–4. Distance (“height”) between points of amalgamation represents the difference between the clusters. As this is a preliminary study, which involved the use of relatively few ID tags and small corpus samples, “height” is fairly low and height distinctions are small, meaning that clusters could be subject to change with the addition of more data. Nevertheless, even these preliminary results give some indications of the potential applications of clustering of senses found in bilingual comparable corpora.

The AU *p*-values are given above each cluster and to the right. When these are low, the apparent cluster does not replicate well and is probably due to chance. Higher AU *p*-values indicate clusters that are more strongly supported by the data.

As might be expected, our trial study confirmed that polysemy networks across languages demonstrate frequent mismatches. Of course, similar-seeming items, such as English *smooth* and Spanish *suave*, have some senses that they share and others that they do not (see Figures 1 and 2). For example, both items can refer to texture, as in *textura suave* or *smooth texture* (labeled as “textura” and “slick”, respectively). On the other hand, *smooth* has an “efficient” sense that *suave* lacks, as in *smooth efficiency*, and *suave* has a “gentle” sense not expressed by *smooth*, as in *soplo suave* ‘gentle breeze’.

Not only does each item each have senses not shared by their near equivalents in another language, but items typically have some senses that are better expressed with one word in an L2 and other senses that are better expressed with a different L2 word. Some senses of English *soft* are close equivalents of senses of Spanish *suave*, and some can be more closely equated with Spanish *blando* (compare Figure 3 with Figures 2 and 4).

For example, senses shared by *soft* and *suave* refer to the texture of skin (the sense labeled “skin” in *soft skin* and the “piel” sense in *suave piel;* but #*blanda piel*) and to silkiness of hair or fur (*soft curls;* these senses are labeled “silky” and “sedoso” in Figures 3 and 2). On the other hand, *soft* and *blando* share a set of senses referring to gentle forces (*soft push;* senses labeled “force” and “fuerza” in Figures 3 and 4), squishy surfaces (*soft mud;* labeled “squishy” and “malleable”), yielding springy surfaces (*soft cushions;* “yielding” and “mullido”), and internal consistency (*soft butter;* “consistency” and “consistencia”). The “yielding”, “force”, “squishy” and “silky” senses of *soft* cluster together in Figure 3, and therefore behave similarly in English, even though some senses resemble *blando* and some

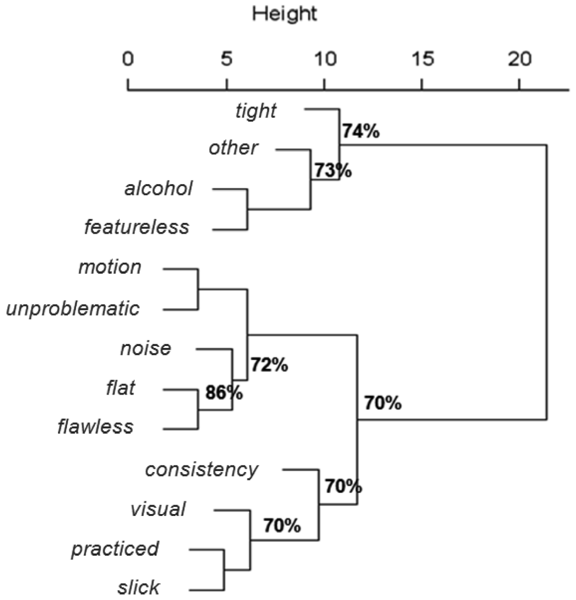

Figure 1: Dendrogram for *smooth* with AU values (values below 70% not shown).

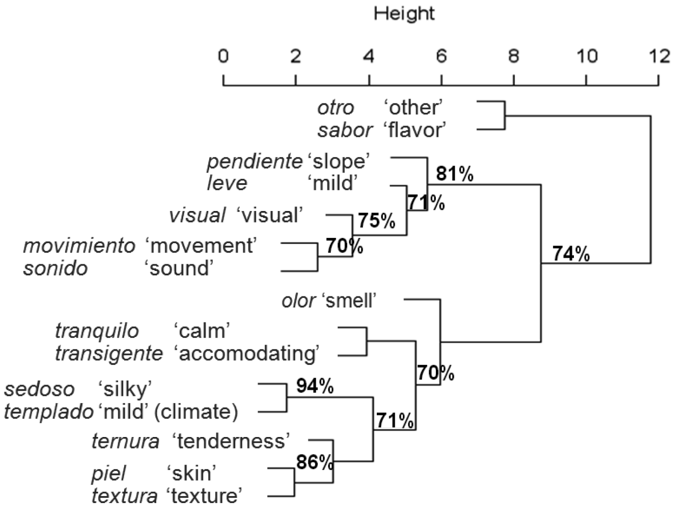

Figure 2: Dendrogram for *suave* with AU values (values below 70% not shown).

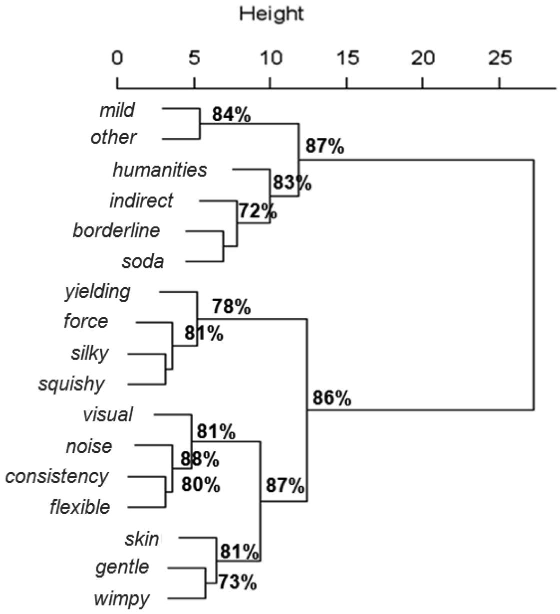

Figure 3: Dendrogram for *soft* with AU values (values below 70% not shown).

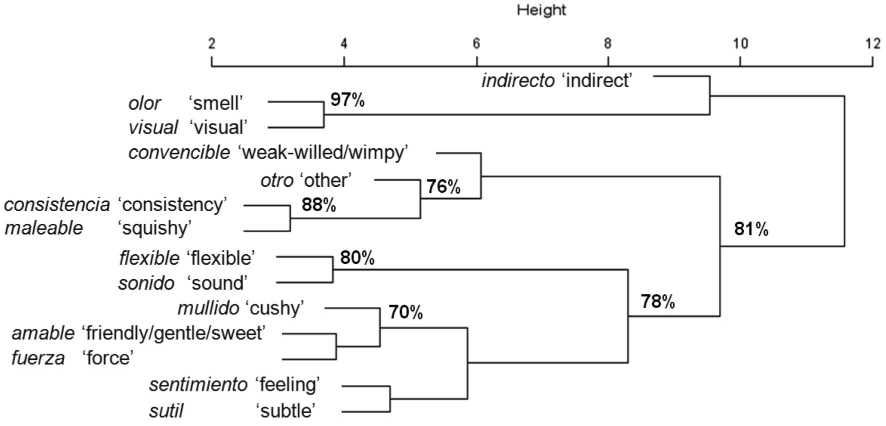

Figure 4: Dendrogram for *blando* with AU values (values below 70% not shown).

suave in Spanish. The "yielding" / "force" / "squishy" / "silky" cluster had an AU *p*-value of 78%, meaning that the cluster recurred in 78% of the bootstrap resamplings (see Figure 3). It therefore occurs with the majority of possible initial orders, and does not depend to a great extent on the initial order of the data.

Although this type of observation may be facilitated by the use of corpora, it could be achieved by careful study alone. A unique contribution of clustering based on word senses is that this analysis can reveal how clusters of senses, as well as individual senses, are related across languages. Our results suggest that many of the senses which are shared between *blando* and *soft*, but not by similar adjectives such as Spanish *suave* and English *smooth*, cluster closely together in English and also to some extent in Spanish. If several of these senses (such as those related to internal consistency and squishy surfaces) are closely related in both languages, it may be no accident that all senses in the cluster are expressed by the same lexical item in each language. Instead, the senses are probably related by underlying semantic commonalities, and may even be predicted, based on these commonalities, to be expressed by a common lexical item in languages other than English and Spanish. Learning the clusters of senses that are near-equivalents in two languages may be an efficient way for L2 learners to understand how to best express a given meaning in their L2, by looking at the clusters of meanings expressed by each L2 item and comparing these with the meanings in their L1. Learning corresponding clusters of senses in the two languages is more efficient than memorizing all the corresponding senses individually.

Cluster analyses based on syntactic features take advantage of more shared variables between senses than analyses based solely on collocations (such as Kaji 2003), in that slightly different senses are far more likely to share syntactic structures than to share individual collocates. Examples (1)–(3) are from the texts taken from the British National Corpus and Corpus del Español and used in the current study (as are all subsequent examples). Each excerpt in (1)–(3) represents a different sense of *soft*, and few collocates are shared by the different senses. However, the shared syntactic structures are immediately evident. The cluster of senses of *soft* that can be characterized as referring to skin texture – “skin” in Figure 3, as in example (1), as well as gentleness of personality “gentle”, as in (2), and weakness of will “wimpy”, as in (3) – demonstrates syntactic attributes typical of the cluster, such as the use of the copula, the predicative position of the adjective, and the presence of other adjectives coordinated with *soft*. The “skin” / “gentle” / “wimpy” cluster has an AU *p*-value of 81%.

(1) *You want the skin there to be as smooth and* ***soft*** *as possible*

(2) *the guy is about 5’ 5”, 130 pounds, sweetheart, intelligent,* ***soft*** *and gentle*

(3) *she’s very well spoken, but she’s pretty* ***soft***

ID tags typical of this cluster include the use of multiple adjectives (as in *smooth and soft*), predicative use of *soft* (*to be… soft, is… soft*) and adverbial modification (*pretty soft*). These uses of *soft* almost always modify singular nouns, typically count nouns (such as *the guy*). For the latter two senses exemplified above, these nouns are also typically animate and human, but this type of semantic information was not taken into consideration in this analysis – a choice which allowed clustering such as with the sense in (1) to come through more strongly. The ID tags listed above are of course not shared by every instance of a sense in the cluster, but did help to typify the cluster relative to senses outside the cluster.

To give a Spanish example, a cluster is formed by the senses of *blando* referring to (4) springy cushiness ("mullido"), (5), weak or gentle force ("fuerza") and (6) mildness or sweetness of personality or behavior ("amable"). This cluster has an AU value of 70% (see Figure 4).

(4) *me acomodaba en los* ***blandos*** *almohadones de un coche del ferrocarril*
'I got settled on the **soft** cushions of a train car'

(5) *el rostro sonrosado por los* ***blandos*** *golpes de la espuma...*
'the face rosy from the **soft** splashes of the spray...'

(6) *¿Y si los* ***blandos*** *halagos de esta niña pudiesen cicatrizar las úlceras de mi corazón?*
'and if the **soft** praises of this girl could heal the wounds in my heart?'

These uses were often in the plural (e.g. *blandos golpes* 'soft splashes/blows'). The adjective *blando* was typically attributive and pre-nominal (*blandos almohadones* 'soft cushions' vs. *almohadones blandos*) and often occurred in a noun phrase with a PP modifier (*... de un coche, ... de la espuma, ... de esta niña* 'of a car, of the spray, of this girl').

Clustering based on syntax can prove useful where both collocates and intuitions are misleading. For example, both Spanish *suave* and English *smooth* modify nouns denoting motion, as in *suave movimiento* and *smooth motion*. However, the sense of *suave* referring to motion ("movimiento" in Figure 2) and the sense of *smooth* describing motion ("motion" in Figure 1) are not comparable. A "smooth" motion is a graceful or practiced motion, whereas *suave movimiento* refers to a weak or feeble movement. This sense of *suave* should probably never be translated as, or equated with, *smooth*, and vice versa, despite the superficial similarity of the expressions that might prompt the senses to be viewed as near-equivalents. The difference in meaning is, however, apparent in the clustering of the senses of *suave* and *smooth* in each language. In Spanish, this sense of *suave* appears to cluster with senses referring to dim visual stimuli ("visual") and weak audio stimuli ("sonido"), and (less closely) with mildness of a condition, such as a disease ("leve") (see Figure 2; this cluster has an AU value of 71%). This suggests that the sense refers to a low position on a scale of intensity – in this case, intensity of the motion described. In English, on the other hand, *smooth* referring to motion ("motion" in Figure 1; as in example [7]) clusters with the sense of *smooth* referring to the unproblematic accomplishment of a goal ("unproblematic"; as in [8]). This cluster has an AU value of only 56%, but is nevertheless worth mentioning due to its incontrovertible difference from the Spanish pattern.

(7) *His gait was* ***smooth****, as if his hip sockets had been oiled…*

(8) *As moose rescues go, this was a* ***smooth*** *one, says Sinnott…*

This clustering suggests that the sense of *smooth* referring to swift unimpeded motion is metaphorically related to the sense referring to swift unhindered accomplishment of a goal. This is a different type of association to that suggested by the clustering of the "movimiento" sense in Spanish. Awareness of this type of clustering can draw attention to the difference in meaning between the two superficially similar senses of *smooth*.

In general, our analysis suggests that metaphoric senses such as *smooth* "unproblematic" do not cluster exclusively with other metaphoric senses in either Spanish or English, but instead cluster with specific non-metaphoric senses. For example, in Spanish, the "amable" sense of *blando* referring to 'kindness' or 'sweetness' clusters with "mullido" ('yielding surface'; AU value 70%), whereas the "convencible" ('weak-willed') sense clusters more closely with "consistencia" ('liquid consistency'; though with an AU value of only 60%); that is, a friendly human being is *blando* in the manner of a comfortable chair, whereas a weak-willed human being is "malleable" like a semi-liquid jelly. English *soft* lacks an "amable" sense referring to sweet behavior or character (expressions such as *soft-hearted* have suggestions of this sense, though these were not well-represented in the corpus). On the other hand, the "wimpy" sense of *soft* in English is connected to "consistency" (as part of a larger cluster with AU value 87%), as in Spanish. The patterns of semantic extension in the languages therefore appear similar, in that specific metaphoric senses are tied to specific non-metaphoric senses, but the resultant networks differ in their details. Awareness of these distinctions is a key to the correct usage of these senses with the appropriate connotations. For example, a Spanish speaker learning English might be unaware that English *soft* lacks some of the positive connotations of the Spanish "amable", but that the negative sense "convencible" translates well as English *soft* "wimpy". The other members of the clusters of these senses in each language make it clear which senses are closer in meaning between the languages. This can be especially useful in understanding metaphoric senses, for which the connections to other senses may not be apparent to an L2 learner.

6 Conclusion

Recent advances in corpus applications have contributed much to Cognitive Linguistics, and increasingly to translation studies and SLA. However, the types of corpora that have been adopted for SLA applications remain largely limited to monolingual corpora. These corpora have proven their utility in translation studies and SLA: monolingual untranslated corpora can give SLA students a feel for the native usage of lexical items and constructions in their L2, and monolingual learner corpora allow SLA students to avoid common mistakes in their L2. We suggest here that bilingual comparable corpora may prove equally well-suited for SLA studies of vocabulary and lexicon. In particular, a comparison of sense clustering in an L1 and L2 can allow students to recognize which types of senses of an item in their L1 correspond most closely to particular items in the L2. For example, this type of analysis demonstrates graphically which groups of senses of English *soft* resemble senses of Spanish *blando*, and which senses of *soft* more closely resemble senses of Spanish *suave*. At the same type, these analyses can draw attention to mismatches between deceptively similar L1 and L2 items, such as English *smooth* and Spanish *suave*, both of which can modify nouns denoting types of motion, but which have very different meanings and hence different positions in the dendrograms of these English and Spanish adjectives. The clusters can also help students choose lexical items with the intended connotations, by drawing attention to the relatedness of these senses with other clearly positively or negatively connotated senses, as in the above example comparing English *soft* and Spanish *blando*. Finally, clusters can aid students in the appropriate use of metaphoric senses, by illustrating how these senses are connected to less metaphoric senses, the meaning and use of which may help clarify the items' metaphoric meanings.

Results from bilingual comparable corpus studies are a long way from being integrated in the SLA classroom. We argue that this lack of progress can be attributed at least in part to the relative paucity of corpus studies aimed at SLA applications, and the lack of diversity in the studies that do exist. Our results, though tentative, suggest that additional types of corpus studies may be productive for SLA. We have also suggested certain methodological choices that may be pursued in order to generate benefits for SLA. It is hoped that recognition of the varied types of corpora and methodologies available for SLA research will lead to the expansion of corpus studies aimed at SLA application, and ultimately the productive integration of these studies and their results in the SLA classroom.

References

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7 (2): 223–243.

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, Harold Somers (ed.), 175–186. Amsterdam: John Benjamins. 1999. The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics* 4 (2): 281–298.

Divjak, Dagmar. 2010. *Structuring the Lexicon: A Clustered Model for Near-synonymy.* Berlin: Mouton de Gruyter.

Divjak, Dagmar, and Stefan T. Gries. 2008. Clusters in the mind? Converging evidence from near-synonymy in Russian. *The Mental Lexicon* 3 (2): 188–213.

Gibbs, Raymond W., and Teenie Matlock. 2001. Psycholinguistic perspectives on polysemy. In *Polysemy in Cognitive Linguistics*, Hubert Cuyckens, and Britta Zawada (eds.), 213–239. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle, and Stefan T. Gries. 2009. Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5 (1): 1–26.

Glynn, Dylan. 2010. Synonymy, lexical fields, and grammatical constructions. A study in usage-based Cognitive Semantics. In *Cognitive Foundations of Linguistic Usage-Patterns*, Hans-Jörg Schmid, and Susanne Handl (eds.), 89–118. Berlin/New York: Mouton de Gruyter.

Granger, Sylviane. 2003. The corpus approach: a common way forward for CL and TS. In *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Sylviane Granger, Jacques Lerot, and Stephanie Petch-Tyson (eds.), 17–30. Amsterdam/New York: Rodopi B.V.

Gries, Stefan T. 2006. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, Stefan T. Gries, and Anatol Stefanowitsch (eds.), 57–99. Berlin/New York: Mouton de Gruyter. 2008. Behavioral Profiles 1.0. A program for R 2.7.1 and higher. Available from the author.

Gries, Stefan T., and Dagmar Divjak. 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In *New Directions in Cognitive Linguistics*, Vyvyan Evans, and Stephanie Pourcel (eds.), 57–75. Amsterdam: John Benjamins.

Gries, Stefan T., and Anatol Stefanowitsch (eds.). 2006. *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis.* Berlin/New York: Mouton de Gruyter.

Jenset, Gard B., and Lidun Hareide. 2013. A multidimensional approach to aligned sentences in translated text. *Bergen Language and Linguistics Studies* 3 (1): 195–210.

Johansson, Stig. 1998. On the role of corpora in cross-linguistic research. In *Corpora and Cross-Linguistic Research: Theory, Method and Case Studies*, Stig Johansson, and Signe Oksefjell (eds.), 3–24. Amsterdam: Rodopi. 2007. *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam/Philadelphia: John Benjamins.

Johns, Tim F., and Philip King (eds.). 1991. *Classroom Concordancing.* Birmingham: Centre for English Language Studies.

Kaji, Hiroyuki. 2003. Word sense acquisition from bilingual comparable corpora. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – Volume 1*, 32–39. Edmonton: Association for Computational Linguistics.

Ke, Shih-Wen. 2012. Clustering a translational corpus. In *Quantitative Methods in Corpus-Based Translation Studies: a Practical Guide to Descriptive Translation Research*, Michael P. Oakes, and Meng Ji (eds.), 149–174. Amsterdam: John Benjamins.

Kishner, Jeffrey M., and Raymond W. Gibbs. 1996. How *just* gets its meanings: Polysemy and context in psychological semantics. *Language and Speech* 39 (1): 19–36.

Malmkjaer, Kirsten. 1998. Love thy neighbour: Will parallel corpora endear linguists to translators? *Meta: Journal des Traducteurs* 43 (4): 534–541.

Marzo, Stefania, Kris Heylen, and Gert de Sutter (eds.). 2010. *Corpus Studies in Contrastive Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Mason, Ian. 2001. Translator behaviour and language usage: Some constraints on contrastive studies. *Hermes* 26: 65–80.

Noël, Dirk, and Timothy Colleman. 2010. Believe-type raising-to-object and raising-to-subject verbs in English and Dutch: A contrastive investigation in diachronic construction grammar. In *Corpus Studies in Contrastive Linguistics*, Stefania Marzo, Kris Heylen, and Gert de Sutter (eds.), 7–31. Amsterdam/Philadelphia: John Benjamins.

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies.* New York: Routledge.

Ringbom, Håkan. 1998. Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In *Learner English on Computer*, Sylviane Granger (ed.), 41–52. London/New York: Addison Wesley Longman.

Schäffner, Christina. 1998. Parallel texts in translation. In *Unity in Diversity? Current Trends in Translation Studies*, Lynne Bowker, Michael Cronin, Dorothy Kenny, and Jennifer Pearson (eds.), 83–90. Manchester: St Jerome.

Sullivan, Karen. 2012. It's hard being soft: Antonymous senses vs. antonymous words. *The Mental Lexicon* 7 (3): 307–326.

Suzuki, Ryota, and Hidetoshi Shimodaira. 2011. Pvclust v1.2-2: Hierarchical clustering with p-values via multiscale bootstrap resampling. [Software] Osaka: Ef-prime. Available from http://cran.r-project.org/web/packages/pvclust/index.html.

Valenzuela, Javier, and Ana M. Rojo. 2008. What can language learners tell us about constructions? In *Cognitive Approaches to Pedagogical Grammar*, Sabine de Knop, and Teun de Rycker (eds.), 197–230. Berlin/New York: Mouton de Gruyter.

Williams, Ian A. 2010. Cultural differences in academic discourse: Evidence from first-person verb use in the methods sections of medical research articles. *International Journal of Corpus Linguistics* 15 (2): 214–239.

Zanettin, Federico. 1998. Bilingual comparable corpora and the training of translators. *Meta: Journal des Traducteurs* 43 (4): 616–630.