Semantic Typology and Parallel Corpora: Something about Indefinite Pronouns

Barend Beekhuizen
Department of Computer Science
University of Toronto
barend@cs.toronto.edu

Julia Watson
Department of Computer Science
University of Toronto
j.watson@mail.utoronto.ca

Suzanne Stevenson
Department of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Abstract

Patterns of crosslinguistic variation in the expression of word meaning are informative about semantic organization, but most methods to study this are labor intensive and obscure the gradient nature of concepts. We propose an automatic method for extracting crosslinguistic co-categorization patterns from parallel texts, and explore the properties of the data as a potential source for automatically creating semantic representations for cognitive modeling. We focus on indefinite pronouns, comparing our findings against a study based on secondary sources (Haspelmath 1997). We show that using automatic methods on parallel texts contributes to more cognitively-plausible semantic representations for a domain.

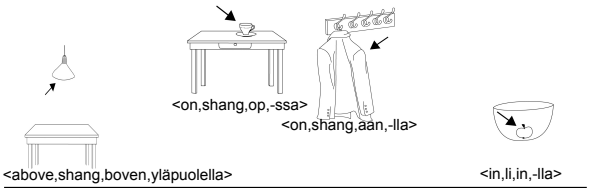
Keywords: semantic typology; semantic representation; parallel corpora

Introduction

An important goal of cognitive science is to determine valid semantic representations, e.g., for use in computational cognitive models of language acquisition and processing. Semantic typology – which studies the patterns of crosslinguistic variation in what words and other linguistic elements mean – reveals universal tendencies in how languages carve up the space of a semantic domain (Haspelmath, 2003; Regier, Kemp, & Kay, 2015). In particular, Bowerman (1993) argues that (all else being equal) the greater the number of languages that label a pair of situations (objects, events, ...) with the same word (called *co-categorization*), the more conceptually similar these situations are. For instance, many languages co-categorize situations of 'stable support' (see Fig. 1) with those of 'tenuous support', but use a different term for 'containment', reflecting that the first two situations are more semantically similar than the last.

More generally, such crosslinguistic co-categorization patterns can define a geometric semantic similarity space (Levinson et al. 2003). To obtain such a space, we first represent a situation as a vector of terms used to express that situation across languages (cf. the row of terms in Fig. 1). These vectors are then projected into a lower-dimensional space (cf. the distances in the two-dimensional space of Fig. 1). This insight has informed cognitive modeling work on spatial relations (Beekhuizen, Fazly, & Stevenson, 2014) and color (Beekhuizen & Stevenson, 2016), where descriptions of situations, elicited from speakers of a number of languages, were used to create vector-based geometric semantic representations. A computational learning model trained on those representations successfully simulated developmental error patterns in word meaning acquisition.

In order to deploy such approaches to additional semantic domains, we need practical and robust methods for semantic typological analysis. Elicitation data (e.g., Berlin & Kay, 1969; Bowerman & Pederson, 1992) – terms describing non-linguistic stimuli, obtained from informants – allows control in defining the set of stimuli for a domain, but is resource intensive and limited to concrete domains. Expert (Haspelmath, 1997) or automatic (Youn et al., 2016) analyses of secondary sources (such as dictionaries and grammars) don't rely on access to informants across many languages, but are focused on coarser-grained semantic distinctions than are found in elicitation stimuli. Both of these methods lack the frequencies and patterns of actual usages in natural communication.



		English	Mandarin	Dutch	Finnish	
Horiz., no contact	Lamp above table	< above	shang	boven	yläpuolella	>
Stable support	Cup on table	< on	shang	op	-ssa	>
Tenuous support	Coat on hook	< on	shang	aan	-lla	>
Containment	Apple in bowl	< in	li	in	-lla	>

Figure 1: Representing the conceptual distance between situations as the number of languages co-categorizing them.

A complementary source of data recently used in semantic typology is parallel text (e.g., Cysouw & Wälchli, 2007) – i.e., the same text translated into different languages. While parallel text has its own potential disadvantages, such as a risk of "translationese" or mistranslations (Levshina, 2017), it can be applied to abstract domains that are hard to obtain elicitation data for, and has actual usage tokens that can reveal nuances of meaning not captured in dictionaries. This latter point is especially relevant for creating semantic spaces for cognitive modeling, as semantic categories display prototype structures with more and less central members (Rosch, 1973). Deriving semantic representations from actual usages can yield a continuous semantic similarity space which potentially reflects such structures; training computational cognitive models on such representations thus has the potential to better match behavioural data. To exploit this potential of parallel text, we need automatic methods for extracting the co-categorization patterns – the terms used across multiple languages for the same situation (cf. Fig. 1) – that can form the basis for such vector-based representations.

In this paper we first propose an automatic method for extracting crosslinguistic co-categorization patterns from parallel texts, to complement elicitation data and secondary sources. Next, we explore the properties of the resulting data as a potential source for automatically creating semantic spaces for cognitive modeling. We focus on indefinite

Acronym	Semantic function	Example
SP-K	specific, known	I want to tell you something.
SP-U	specific, unknown	Someone broke into our apartment.
NS	irrealis non-specific	I need someone strong for the job.
CD	conditional	Let me know if anybody shows up.
QU	question	Is anything bothering you?
IN	indirect negation	I don't think anything matters.
DN	direct negation	Nobody came.
CP	comparison	She can run faster than anybody.
FC	free choice	You can pick anything!

Table 1: Haspelmath's 9 functions with examples.

pronouns as an abstract semantic domain for which elicitation would be a difficult method, but for which we have a good understanding of the typology from expert judgments and secondary sources (Haspelmath, 1997). By using parallel texts, we are able to get a fuller picture of the semantic structure of this domain, in particular seeing evidence for gradience in multiple ways: finer-grained semantic functions that show gradient patterns across languages, and gradient relationships (distances) among the semantic functions. We thus show that using automatic methods on this complementary data source can contribute to more cognitively-plausible semantic representations, by fleshing out expert analysis of secondary sources with usage data that reflects the discourse use and frequency of the semantic functions.

Indefinite pronouns

Indefinite pronouns, such as *somebody*, *anything*, and *nowhere*, are used to express indefinite reference – i.e., introducing a discourse referent which the speaker typically does not intend the hearer to uniquely identify. Reference may be to an entity from any of the major ontological categories such as PEOPLE, THINGS, and PLACES. Haspelmath (1997) outlines 9 semantic functions that indefinite pronouns can express; see Table 1. To identify the set of functions, he draws on semantic motivation – whether a coherent functional definition can be established for each. Importantly, linguistic evidence is considered for deciding whether two related functions should be merged or split: specifically, if at least one language has a term that can be used for only one of the functions – i.e., if there is a language with a term that does not co-categorize the two – then the two functions are considered distinct.

The identified semantic functions are analogous to stimuli in an elicitation task, although at a coarser grain: each function represents *a set of situations* that are co-categorized. Like elicitation data, terms in each language are associated with each of the semantic functions they can express, and patterns of crosslinguistic co-categorization can be revealed. These patterns can be visualized in a graphical *semantic map*: functions (nodes) are connected by edges such that connected subgraphs correspond to sets of functions that can be co-categorized. The semantic map of Haspelmath (1997), in Fig. 2, shows that, in both example languages, the terms carve out different, but in both cases connected, partitionings of the graph.

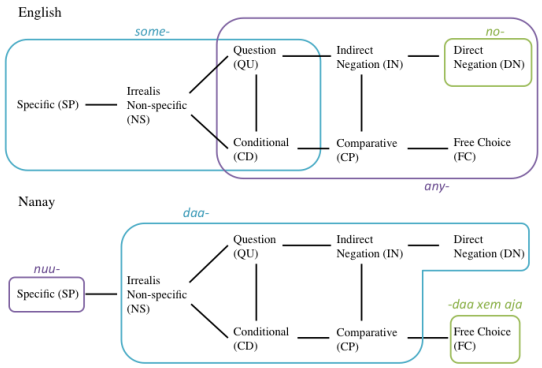


Figure 2: Semantic map from Haspelmath (1997) with English and Nanay terms.

Despite the insight they provide, semantic maps do not capture certain properties of the underlying semantic space that are important to semantic representation. Two related issues in particular motivate our work here. First, there is no indication of the distance in semantic space that an edge in the map represents, although it is likely that some functions connected to a node may be closer or further semantically than others. For example, although IN connects to both DN and CP by a single edge in Fig. 2, it is likely more similar to DN. Second, the use of a single node for a function assumes (instrumentally) that functions are internally homogeneous. However, functions may display a gradient internal structure – e.g., some cases of DN may be 'better' instances than others. Both of these factors may contribute to the cognitive plausibility of a semantic space for use in computational modeling.

As discussed above, parallel usage data has the potential to address these issues by providing a more continuous representation than secondary data. Actual usage data may reveal how related Haspelmath's various functions are, and how homogeneous they are internally. Such insights are crucial for the use of semantic-typological analyses in cognitive science, e.g., in modeling the acquisition of such terms.

Method: Translations from Parallel Text

Our goal is to construct geometric semantic spaces through the use of parallel (translated) usage data. We draw on the patterns of how terms are translated across many languages to find co-categorization patterns, which can then be used to derive a semantic space. We propose an automatic method that extracts the translations of each occurrence of a seed word (here, English indefinite pronouns) in every other language in our corpus. These extracted arrays of translations form a vector of terms across languages analogous to those obtained through elicitation data (cf. Fig. 1), and can be used to construct a geometric space.

Corpus and language sample. We extracted our data from a sentence-aligned parallel corpus of subtitles of films and TV series (Lison & Tiedemann, 2016; `www.opensubtitles.org`). We selected the 30 (out of 65) languages across 9 language families for which the most

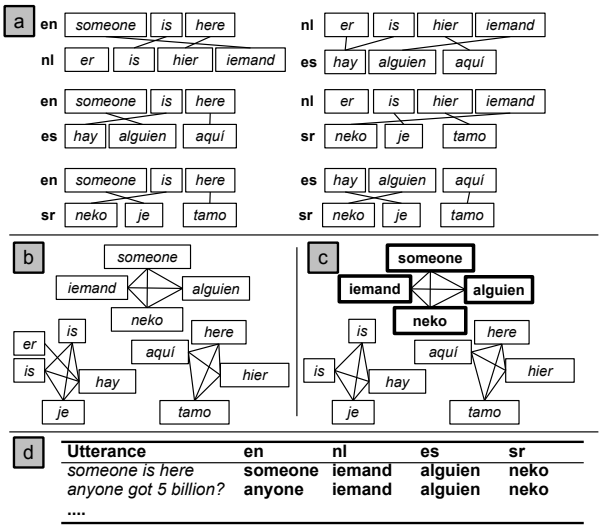


Figure 3: Extraction of situation vectors; see text.

parallel data was available,[1] and extracted all utterances for which we found a translation into all languages.

Identifying translations across languages. We first obtained automatic alignments of translated words for each pair of languages in our corpus, using the HMM implementation of Liang, Taskar, and Klein (2006) with the default settings; see Fig. 3(a) for an example with four languages. From the pairwise alignments, we created a graph, per utterance, with edges between all words that are aligned with each other, (Fig. 3(b)). From this graph, we extracted the subgraphs that were densely connected (i.e., for which the words are often mutually aligned),[2] and select those subgraphs that contain one of the indefinite pronouns in English (Fig. 3(c)). Each such subgraph is then linearized to form the vector representation of a situation (Fig. 3(d)). The Table in Fig. 3(d) illustrates the correspondence to semantic typology: Every row contains a 'stimulus', for which the various languages present elicited terms (cf. the table in Fig. 1). Note that sometimes responses are missing, or multiple words form the response.

Extraction of indefinite pronouns. We focus on the two ontological categories PEOPLE and THINGS; other categories (e.g., TIME and PLACE) were too infrequent. To identify indefinite pronoun usages in our corpus, we extracted utterances for which the English expression consists of any of the 9 words combining *some-, any-, no-* with *-thing, -body, -one* (cf. rows in bold in Fig. 3(d)). From among these situations, we selected only those that included an expression from each of at least 25 languages, to ensure sufficient linguistic variation for each situation.[3] The resulting data consisted of 698 situations – i.e., exemplars of indefinite pronoun usage represented by vectors of terms in 25-30 languages.

Annotation. In order to compare the patterns in our usage data to Haspelmath's (1997) analysis, it was necessary to identify the semantic function (see Table 1) of the indefinite pronoun usages in each of our situations. To do so, three annotators (the authors) labelled the English indefinite pronoun in each situation with its Haspelmath function. Annotators were provided the sentence containing the pronoun, as well as some context before and after. We merged Specific Known and Specific Unknown into one function called Specific (SP), given the uncertainty in this task of judging whether something is known to the speaker.[4] 152 cases consisted of negative English indefinite pronouns like *nothing* and *no one*, which we automatically marked as DN. On the remaining 546 exemplars, inter-annotator agreement was satisfactory for a task of this difficulty (pairwise Cohen's $\kappa = [.84, .80, .79]$), and the majority annotation was used for each situation.

Further experimental set-up. Although the extracted situations are generally of a high quality,[5] sometimes mistranslations are extracted. To reduce noise, we only use those terms that are statistically significantly associated with at least one of the annotated functions (using a Fisher Exact test). This way, low-frequency translations that are dispersed over functions are filtered out. To avoid the risk of overinterpreting patterns or overtuning models on the basis of a single sample, we split the data set into a development (dev) and test set. The examples and patterns reported below come from both the dev and test set, but quantitative results are provided for the test set only. We conduct all analyses on PEOPLE and THINGS separately, because we found in exploratory data analysis that PEOPLE and THINGS showed differences in their patterns which have potentially interesting cognitive implications. The full data set, including stemming dictionary, annotation schemas, and all software used for the analyses, can be found at `https://github.com/dnrb/indefinite-pronouns`

Results

With the extracted situation vectors, we can now study the semantic space derived from parallel usage data, and see how similar it is to Haspelmath's (1997) semantic typology based (primarily) on secondary sources. In particular, we are interested to see where the parallel usage data reveals characteristics of the semantic space not observed in Haspelmath's map.

Are all semantic functions equally important?

Table 2 presents the frequency of the semantic functions. We see that most functions in the center of Haspelmath's

[1] The set of languages is (per language family, in ISO 639-2): (Semitic) ar, he; (Indo-European) bg, bs, cs, da, de, el, en, es, fr, hr, it, nl, no, pl, pt, ro, ru, sl, sr, sv; (Finno-Ugric) et, fi, hu; (Austronesian) id; (isolate) ja; (Turkic) tr; (Vietic) vi; (Sino-Tibetan) zh.

[2] We used k-clique percolation (Palla et al., 2005) with $k = 9$.

[3] We used a manually compiled stemming dictionary to lemmatize the words and correct spelling and alphabetic variation.

[4] Annotation was done for English only. It is possible that Haspelmath's functions are not always translated: a conditional may be translated as an declarative. Being relatively infrequent, we consider these cases noise.

[5] Evaluating the method on a parallel Bible corpus against a gold standard of Strong number annotations gives a cluster purity of .89 and a cluster recall of .90.

	SP	NS	CD	QU	IN	DN	CP	FC
PEOPLE	.16	.20	.07	.16	.05	.28	.01	.08
THINGS	.28	.15	.05	.09	.02	.36	.00	.06
Overall	.24	.17	.06	.11	.03	.33	.00	.06

Table 2: Distribution of functions given ontological category.

$k=$	2	3	**4**	5	6	7	8	9	10
PEOPLE	.20	.25	**.41**	.35	.34	.34	.32	.30	.32
THINGS	.30	.38	**.47**	.36	.35	.35	.33	.39	.33

Table 3: Adjusted Rand index score for PEOPLE and THINGS with k-means clustering, given various values of k.

(1997) semantic map are rather infrequent (CD, IN, CP). This may explain Haspelmath's observation that, across languages, there are no terms that solely apply to two functions in the middle of the map: Languages typically co-categorize infrequent functions with one of the more frequent neighboring functions (e.g., NS or DN). It also explains aspects of the graphical structure of the map: low-frequency functions are in the middle of the map because sometimes they share a term with the left side of the map, and sometimes with the right.

A notable exception is FC, located at the edge of the map despite its low frequency. This suggests FC is conceptually different from the other functions (except CP). Many languages co-categorize FC and universal quantification – unlike English, which generally uses *any-* vs. *every-* respectively. The use in many languages of a universal quantification term for the semantic function FC may account for its distinctive position in the map despite its low frequency.

Are the functions at the right level of granularity?

A second issue worth investigating is whether Haspelmath's proposed functions constitute the best way of grouping the usage data into sets with related semantics: actual usage data may reveal that the functions are not well discriminable or have further coherent subdivisions. We explore this through automatic clustering of the parallel usage data. Each of our extracted situations is a vector of mutually-translated indefinite pronouns (see Fig. 3(d)); together they form a vector space within which we can measure situation (dis)similarity. Thus we can determine the optimal partitioning of the data into clusters and see how well those clusters correspond to the gold annotation. Here, we use k-means clustering (MacQueen, 1967), an unsupervised technique that partitions the data into k clusters. The input for k-means is a distance matrix between all pairs of situations belonging to either PEOPLE or THINGS. The distance d between a pair of situations s, s' is given by taking the Jaccard index over the sets of terms[6] $T_l(s)$ and $T_l(s')$ used to express each of s, s' in each of the languages $l \in L$, and summing over all languages l:

[6]We use *sets* of translated terms, because an indefinite pronoun in English may be translated to multiple terms in other languages.

	Function								Evaluation		
Cluster	SP	NS	CD	QU	IN	DN	CP	FC	P	R	F_1
1	**18**	**24**	**6**	3	0	2	0	0	.91	.92	.91
2	1	0	2	**15**	1	4	0	2	.60	.83	.70
3	0	0	1	0	**5**	**27**	0	0	.97	.82	.89
4	0	0	0	0	0	0	**1**	**7**	.80	1.00	.89

Table 4: Correspondence table for the 8 functions with $k = 4$ clusters for PEOPLE; rightmost columns present cluster precision (P), recall (R) and F_1 score for every cluster against function tuples (SP,NS,CD); (QU); (IN,DN); (CP,FC).

$$d(s,s') = \sum_{l \in L} \frac{|T_l(s) \cap T_l(s')|}{|T_l(s) \cup T_l(s')|}. \quad (1)$$

We assess the relative quality of different numbers of clusters by comparing their fit to the annotations using the adjusted Rand index (Rand, 1971). Table 3 presents the results.

If Haspelmath's set of functions is the best way of describing the data, k-means clustering with $k = 8$ should be the k with the highest correspondence to the annotated functions, partitioning the data into 8 clusters corresponding to the 8 functions. However, with $k = 8$, a relatively poor Rand index score is achieved, and rather than aligning with the semantic functions, the inferred clusters mostly cross-cut them (e.g., there are 2 clusters containing many DN). The fact that the optimal partitioning cross-cuts functions suggests that there are finer semantic distinctions within the functions that play out in the way languages label these.

Instead, we find that $k = 4$ gives the highest correspondence with the manually annotated clusters. The 4 clusters correspond to 4 sets of related functions: (SP,NS,CD), (QU), (IN,DN), and (FC,CP); see Table 4. There is some leakage between the clusters (see the non-boldface numbers in Table 4), but the precision, recall, and F scores using these sets of functions as the target labels for the 4 clusters are very high, showing these sets of related functions have a clear similarity structure.[7]

These results yield two distinct views of the data. On one hand, the typological usage data points to more fine-grained semantic distinctions *within* some of the 8 functions. On the other hand, we find semantic similarity *between* the functions that reveals a coarser grouping of the functions than is apparent from the semantic map structure of Haspelmath (1997). These findings point to a key role of gradience in understanding the semantic space of indefinite pronoun usage.

The perspective of a similarity space

The clustering over the parallel usage data suggests more gradience in the semantic space underlying indefinite pronoun semantics, both within and between functions, than the semantic map of Haspelmath (1997) suggests. We take a more

[7]These clusters do not completely coincide with the further analysis of Haspelmath (1997, par. 5.6), but we leave that comparison for future research.

(a) PEOPLE (b) THINGS (c) Terms for DN THINGS in (left-to-right) Estonian, Croatian, English, Slovene.

Figure 4: OC plots for indefinite pronouns (best viewed on screen).

direct way of obtaining insight into this space by representing the similarity between all situations in a low-dimensional space. Visualizing this space can also be informative about the use of such a space in a computational cognitive model.

We apply 2-dimensional Optimal Classification (OC), a dimensionality reduction technique useful for typological data (Croft & Poole, 2008). The input for this algorithm is the list of individual situations, each represented as a set of terms (across all languages) used for that situation. For each term w across all languages, OC creates a cutting line in the 2-dimensional plane, which divides the situations into those expressed with w and those not expressed with it. This way, pairs of situations expressed with similar sets of terms will typically be located close together in the OC space. (In our data, we have $n = 303$ cutting lines for PEOPLE and $n = 435$ cutting lines for THINGS.) Our data yields very high accuracy (proportion of situations being on the correct side of the cutting line, averaged over all cutting lines) of .94 (PEOPLE) and .95 (THINGS). Because each situation is represented by a set of terms across all languages, this result shows high agreement among the languages in how they carve up the situations into functions – although, as we see next, they exhibit gradience in the gradual shifting of terms for related situations.

The topology of the function annotations in the two-dimensional space generally follows Haspelmath's map, despite working from different data and with different methods (see Figures 4a and 4b): in the top-left corner, we find NS and SP followed by CD and QU towards the center; the top-right cluster contains DN and IN, whereas the bottom cluster consists of CP and FC. [8] However, there are several aspects of the functions that are observable in this continuous space that are not apparent in the graphical semantic map.

First, we observe that not all functions that neighbor each other in Haspelmath's map are equidistant in the OC solution: QU has an edge to each of NS and IN in the map, but is closer to the former in the OC projection. The projection furthermore displays gradience among the functions: some QU-labeled situations are closer to NS, whereas others are closer to IN, DN, or FC, suggesting that the functions are more continuous than the graphical map suggests.

Language						
bs	hr	en	sl	pt	da	Functions
išta	išta	anything	kaj	alguma coisa	noget	QU
	što					QU, CD
nešto	nešto	something				QU, CD
			nekaj			QU, CD, NS
				algo		NS, SP

Table 5: A gradient for the (SP,NS,CD,QU) region.

Second, the functions display internal gradience. Fig. 4c shows terms in four languages for THINGS annotated as DN. The gradient comes about because of languages whose terms form supersets of each other: Estonian *keegi* is a superset of Croatian *nitko*, which is a superset of English *nobody*, which is a superset of Slovene *nihče*. Across languages there thus seems to be agreement about a scale of subtypes of DN, but languages vary on the placement of the lexical boundaries.

Finally, we find gradients that cross-cut the function boundaries. Table 5 illustrates a gradient of terms standing in a superset-subset relation to each other that cross-cuts the functions SP, NS, CD, and QU. This gradient was obtained by running a one-dimensional OC on the situations in the (SP,NS,CD,QU) region, which lays out all situations on one line so as to obtain a maximal accuracy in placing cutting points for terms. This analysis yields an accuracy of .96, which suggests that languages strongly agree on having a single dimension roughly cross-cutting the functions SP, NS, CD, and QU on which they locate their term boundaries.

Visualizing crosslinguistic usages in a continuous space gives further insight into the structure of the underlying semantic domain. The observed gradients call for further analysis and provide predictions for behavioural experiments. In

[8] Here and elsewhere, we observe evidence of a finer-grained semantic space for PEOPLE than for THINGS, in line with typological observations such as Silverstein (1976): The distribution of functions given each ontological category is more spread out for PEOPLE than for THINGS (Tab. 2), and decreasing the number of clusters to 3 or 2 deteriorates the Rand score less for THINGS than for PEOPLE (Tab. 3). We note that usage data reveals distinctions that remain obscured when glossing over ontological categories.

particular, if the patterns of crosslinguistic variation are indicative of cognitive distinctions in semantic space, we expect to see evidence in both adult behaviour and developmental patterns in children.

Conclusions

Crosslinguistic patterns of co-categorization yield insight into the semantic space underlying linguistic usages. We deploy parallel usage data in the form of movie subtitles to study the patterns of crosslinguistic variation in the categorization of indefinite pronouns. We find the cross-linguistic usages display a more fine-grained pattern than suggested by a study on the basis of (primarily) secondary data (Haspelmath, 1997). In particular, the frequencies of the identified semantic functions vary, the distances between the functions are not uniform, and within functions, coherent subgroupings could be established. Our findings suggest the parallel usage data captures something about the semantic space that is not represented in the more static secondary sources.

The current method can easily be applied to other domains, but also involves several restrictions. Using pairwise alignments on parallel texts makes the approach computationally intractable beyond 30–50 languages, as a set of alignments has to be extracted for every language pair. We are looking into methods to circumvent this aspect of the method. The inability of the model to go 'below' the word level is also limiting, as many well-established patterns of cross-linguistic semantic variation involve morphology (e.g., case marking, nominalization patterns).

Furthermore, it is crucial to establish the cognitive plausibility of the semantic similarity space independently by seeing if it can predict behavioral experiments such as word usage similarity judgments, or developmental patterns. For example, we must explore whether, as for space and color, the semantic space for indefinite pronouns predicts aspects of the acquisitional pattern of these words: Is English *any-*, for instance, hard to acquire because it covers a large, rather disjunct region of the semantic space? Are indefinite pronoun systems in languages that follow the typologically more common patterns easier to acquire for first and/or second language learners? We hope these automatic methods for using parallel text in semantic typology can help us further understand patterns of learning and usage in abstract domains of meaning.

Acknowledgments

We would like to thank Martin Haspelmath for sharing his indefinite pronoun data, and support from NSERC of Canada.

References

Beekhuizen, B., Fazly, A., & Stevenson, S. (2014). Learning meaning without primitives: Typology predicts developmental patterns. In *Proceedings CogSci.*

Beekhuizen, B., & Stevenson, S. (2016). Modeling developmental and linguistic relativity effects in color term acquisition. In *Proceedings CogSci.*

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: UC Press.

Bowerman, M. (1993). Typological perspectives on language acquisition: Do crosslinguistic patterns predict development? In E. Clark (Ed.), *Proceedings of the 25th annual Child Language Research forum* (pp. 7–15).

Bowerman, M., & Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. In *Annual report of the MPI for Psycholinguistics* (pp. 53–56).

Croft, W., & Poole, K. (2008). Inferring universals from grammatical variation: multidimensional scaling for typological analysis. *Theoretical Linguistics*, 1–37.

Cysouw, M., & Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *Language Typology and Universals*, *60*, 95–99.

Haspelmath, M. (1997). *Indefinite pronouns*. Oxford: OUP.

Haspelmath, M. (2003). The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In M. Tomasello (Ed.), *The new psychology of language* (pp. 211–242).

Levinson, S. C., Meira, S., & The Language and Cognition Group. (2003). 'Natural concepts' in the spatial topological domain – Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, *79*(3), 485–516.

Levshina, N. (2017). Subtitles as a corpus: An n-gram approach. *Corpora*.

Liang, P., Taskar, B., & Klein, D. (2006). Alignment by agreement. In *Proceedings NAACL* (pp. 104–111).

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings LREC.*

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297).

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*, 814–818.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263).

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Silverstein, M. (1976). Hierarchy of features and ergativity. In R. Dixon (Ed.), *Grammatical categories in Australian languages* (pp. 112–171).

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., Croft, W., & Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *PNAS*, *113*, 1766–1771.