Evidence for a language-independent conceptual representation of pronominal referents

Mora Maldonado (mora.maldonado@univ-nantes.fr)
Laboratoire de Linguistique de Nantes
Centre National de la Recherche Scientifique (CNRS)

Noga Zaslavsky (nogazs@mit.edu)
McGovern Institute for Brain Research
Massachusetts Institute of Technology

Jennifer Culbertson (jennifer.culbertson@ed.ac.uk)
Centre for Language Evolution
University of Edinburgh

Abstract

Across many semantic domains, cross-linguistic regularities in categorization systems (e.g., color or kinship terms) have been taken to reflect constraints on how humans perceive and conceptualize the world. Such conceptual representations are often assumed to be universal, and independent of an individual's experience with a particular language. However, in most cases, representational constraints have not been observed empirically on language-independent grounds. This study comes to fill in this gap. We use a card sorting task to provide the first empirical evidence for a common, language-independent representation of pronominal referents, shared by speakers of different languages.

Keywords: mental representations; categorization; personal pronouns; cross-linguistic variation; concepts; typology

Introduction

Categorization systems involving both content and functional vocabulary exhibit constrained variation across languages: while some are very frequent, others are very rare or do not occur at all. For example, while the world's languages exhibit different ways of categorizing kin concepts into kinship terms and color values into color words, there are apparent constraints on the types of attested categorization systems (Berlin & Kay, 1969; Murdock, 1970; and see also Kemp, Xu, & Regier, 2018, and references therein for other semantic domains). Cross-linguistic regularities of this sort have often been argued to reflect, at least in part, representational constraints—specific ways of representing the meaning space under consideration (see, e.g., Kay & McDaniel, 1978; Gardenfors, 2014). The idea is that the way humans represent the world imposes constraints on linguistic forms, such that more frequently attested systems are those that better reflect perceptual or conceptual aspects of the underlying mental representation of the domain.

In a number of domains, theories and models of the typology of categorization systems have assumed that these mental representations are universal, and independent of our experience with a particular language. For example, Zaslavsky, Maldonado, and Culbertson (2021) analyze systems of personal pronouns (e.g., 'I', 'you', 'they'), which categorize individuals by their role in the communicative context (e.g., speaker, addressee). Building on insights from linguistic theories of person (Harley & Ritter, 2002; Harbour, 2016; Zwicky, 1977), Zaslavsky et al. introduce a bias into the representation of possible pronominal referents, whereby referents that involve the speaker are more similar to each other than those that do not. This bias is assumed to be universal (shared by all speakers, regardless of their native language). Zaslavsky et al. adopt an information-theoretic framework and show that this biased representation of the domain leads to a better account of the typological distribution of systems of personal pronouns than an unbiased representation. However, they also speculate that at least some rarely attested systems may be explained by a weaker bias in the representation, raising the possibility of language-specific representations.

While it is compelling that cross-linguistic data are better captured by a model that assumes a universal representation of this space, in most cases it remains a theoretical assumption. In the case of color, this assumption has been justified by a language-independent perceptual space (Regier, Kay, & Khetarpal, 2007; Zaslavsky, Kemp, Regier, & Tishby, 2018), and in the case of containers a shared representation between two closely related languages (Dutch and French) has been reported (White, Malt, & Storms, 2017). However, most models of semantic categorization in other domains assume a universal representation space without grounding that assumption in language-independent evidence (e.g., Kemp & Regier, 2012; Xu, Liu, & Regier, 2020; Denić, Steinert-Threlkeld, & Szymanik, 2022; Steinert-Threlkeld, 2020). That is, whether or not the mental representation of a given semantic domain is independent from language is an empirical question that has so far remained largely unaddressed. Addressing this question would significantly strengthen our theories of the underpinnings of the conceptual representations relevant for categorization in language. For example, if there is evidence that representations are *not* language-independent, then models could potentially be improved by tuning the representational space for each language. On the other hand, evidence supporting universal representations would strengthen models that rely on that assumption.

Here, we address this open question in the domain of personal pronouns. Specifically, we test how speakers of languages with different pronominal systems conceptualize the space of possible pronominal referents. We do this using a well-known method for investigating non-linguistic representations: card sorting (Weller & Romney, 1988; Fincher & Tenenberg, 2005). To preview, we find that the structure of this meaning space—including the 'speaker' bias posited in Zaslavsky et al. (2021)—is indeed highly correlated across the populations we test.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

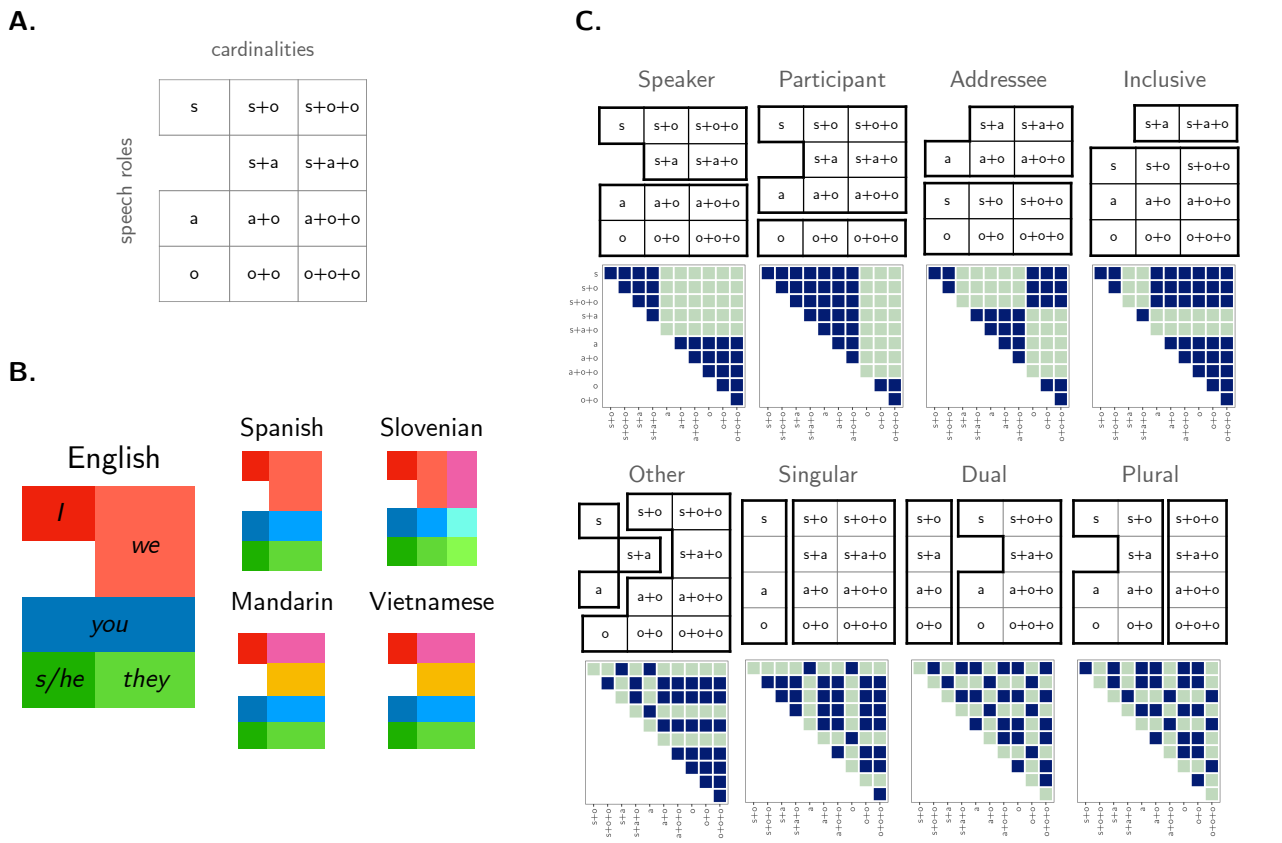


Figure 1: **A**. 11 Possible referents obtained by crossing three speech roles (abbreviated as *s*, *a* and *o*) with three possible cardinalities. **B**. Pronominal systems in the 5 languages tested in our study, plotted against the space shown in A. Colors correspond to distinct pronominal forms. **C**. Examples of possible conceptual representations of the 11 meanings illustrated in A. Each representation structures the domain around a given person/number feature and is labeled by that feature, whereby referents that share/lack this specific feature are encoded as more similar to each other. The upper panels illustrate the similarity structure schematically (i.e., more similar referents are bounded within a box); the lower panels show the corresponding similarity matrix between referents (dark blue indicates high similarity (i.e., 1), light green low (i.e., 0)).

The meaning of personal pronouns

Personal pronouns refer to individuals or groups of individuals by their role in the speech event. For example, in English, the pronoun 'I' is used to refer to the speaker, the pronoun 'you' to the addressee, and the pronoun 'they' to someone who is not actively participating in the conversation.

Here, we investigate how speakers represent a semantic space of 11 possible pronominal referents. This space accounts for three speech roles (**s**peaker, **a**ddressee, and non-participant **o**ther) interacting with three cardinalities (exactly one, exactly two, and more than two).[1] Fig. 1A illustrates this space. For simplicity, we assume a single speaker, a single addressee, and an undefined number of non-participant others (see Maldonado & Saldana, 2022, for a detailed discussion of these assumptions), which means that the three speech roles can be combined to form pronominal referents of varying cardinalities.

Languages vary in how they map these pronominal referents to forms: some languages express all referents using unique forms and others feature homophony, i.e. multiple meanings expressed by a single form. For example, English uses the pronominal form 'we' as long as the speaker is part of the relevant group, whereas Mandarin uses different forms depending on whether the group includes both speaker and addressee ('zánmen') or just the speaker ('wǒmen'). Fig. 1B illustrates the form-to-meaning mapping in the 5 different languages considered in our current study.

Speakers may structure their representation of these 11 referents around different primitive concepts or relations between primitive concepts. Evidence from artificial language learning experiments shows that speakers dissociate pronominal referents that involve different speech roles and/or cardinalities, even when these are mapped into the same form in their native language (Maldonado & Culbertson, 2020; Lee, 2020). This suggests that some of these different speech roles (e.g., speaker, addressee) and cardinalities (e.g., exactly two, more than two) are treated as distinct primitives.

Importantly, speakers might also represent some referents as more similar to others, leading to further structure, or non-

[1]Larger spaces are possible, and indeed likely given the different number systems found across the world's languages.

uniformity, in their representation of the space. Fig. 1C shows eight hypothetical representations of the pronominal domain in terms of similarity relations between the 11 referents in the space of Fig. 1A. Each of these reflects some representational structure based on similarity between referents, depending on which feature of the referent drives conceptual similarity. For example, the 'speaker' representation is one where referent similarity is driven by whether the referent includes the speaker or not (similar to the 'speaker' bias proposed by Zaslavsky et al., 2021); the 'singular' representation encodes a difference between referents depending on their cardinality (i.e., exactly one individual vs. more than one individual); and so on.

Cross-language study: Conceptual similarities between pronominal referents

Our goal in this work is to understand the structure of speakers' mental representation of the domain of pronominal referents. Specifically, we are interested in understanding (a) whether representations of the domain are shared by all speakers, regardless of their native language(s); and (b) how speakers structure this semantic space.

To this end, we conducted a cross-language study based on a successive card sorting experiment (Boster et al., 1994). In this task, participants are given a collection of cards depicting the referents of interest and are instructed to sort them successively in two piles (see Fig. 2B). Participants are thus forced to rely on some strategy that allows them to pair referents. The degree of similarity between any two referents in our domain is then estimated by the likelihood of them being sorted in the same pile.

The sorting task does not directly involve language usage or knowledge, thus participants' sorting strategies can be taken as a proxy for non-linguistic conceptual representations of the domain. In order to explore the possible impact of native language on this representation, we included in our study speakers of five different languages: English, Spanish, Slovenian, Mandarin, and Vietnamese. These languages differ in their person pronominal systems, as shown in Fig. 2B. For example, despite their differences, English, Spanish, and Slovenian use the same pronominal form to refer to groups that include the speaker, regardless of whether the addressee is also included or not. This can be thought of as a specific instantiation of a 'speaker' representational bias, whereby all referents including the speaker are perceived as more similar to each other than those that do not. In contrast, Mandarin and Vietnamese make a clusivity distinction in their pronominal system: they use different pronouns to refer to groups that include both speaker and addressee (inclusive) or only the speaker (exclusive).

We expect that two pronominal referents will be perceived as more similar to each other (and thus sorted in the same pile) as a function of: (a) overlap of primitives, and (b) relative importance/weight of the primitives. By (a), referents that share primitives, either conversational roles (speaker, addressee, other) or cardinalities (one, two, more than two), should be more likely to be sorted into the same pile than those that do not. For example, a card that depicts the speaker and a non-participant other might be perceived as more similar to a card that depicts only the speaker than to a card that depicts only the addressee. By (b), some primitives might have more weight in the representation than others, such that the similarity between referents that share those primitives is stronger. For example, referents that include the speaker might be perceived as more similar to each other than referents that include the addressee (see Fig. 1C for alternatives).

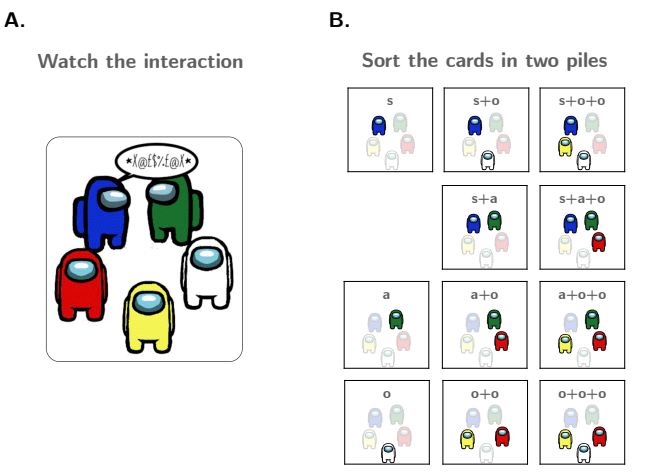


Figure 2: Illustration of experimental procedure. **A**. In the first phase, participants watch an interaction between two characters. **B**. In the second phase, participants are given cards depicting the 11 referents of interest (see Fig. 1A) and are asked to successively sort them in two piles.

Methods

The preregistered design and analysis plan is accessible here. The study was approved by the Ethics Committee of [redacted].

Participants

A total of 812 participants, native speakers of English, Spanish, Slovenian, Mandarin and Vietnamese, participated in the successive pile sorting task. Speakers of English (131), Spanish (131) and Slovenian (122) were recruited through Prolific. Mandarin (106) and Vietnamese (356) speakers were recruited by word of mouth with the only requirement that they were currently living in China and Vietnam respectively.

Per our pre-registration, we excluded from the analysis participants who: (a) in the first sorting round (when all sorting criteria were available) repeatedly sorted two instances of the same meaning (category) in two different piles; and (b) who sorted all the cards in the same pile in the first two rounds. Following previous work using the same methodology (Ameel, Storms, Malt, & Sloman, 2005), we gathered participants in each language group until we reached a reliability (i.e., consistency of sorting strategies within a lan-

guage) of 0.90. Data from 370 participants were used for the analysis (ENG=76; SPA=79; MAN=67: SLV=51: VIE=97).

Materials and Procedure

Participants were first shown four GIFs, one after the other, depicting 5 cartoon characters of different colors (i.e., green, red, blue, yellow and white). All GIFs displayed an interaction between two of these characters: one character above whom a speech bubble appeared (acting as the *speaker*) and another one who was standing close to, and directing its gaze towards this character (acting as the *addressee*). The remaining three characters were further away from these two, and facing away from the interaction (acting as *others*). Characters' colors were randomly assigned per participant and remained constant for that participant throughout the task. GIFs varied in the relative location/placement of the characters in the space. A static example of the GIF is provided in Fig. 2A.

After watching this interaction, participants were introduced to the card sorting task. Participants were presented with collection of cards depicting the meanings of interest using the same characters they had seen in the GIFs (see Fig. 2B). Participants were instructed to group the cards in two piles based on the interactions they had seen. Participants had to sort the cards successively in a top-down manner: they started with the whole set of cards and had to sort recursively until they had sorted all the cards in piles of no more than 2 cards. In other words, on the first round, participants sorted all cards into two piles. They were then shown all the cards from one of these piles and had to sort them into two new piles. Then they sorted all the cards from the second pile into two piles. And so on, until the piles consisted of two or fewer cards each. Participants were told that the piles could be different sizes, and that, if they did not see a way of grouping the cards, they could also leave one pile empty (but they should use this strategy as a last resort).

All cards depicted the 5 characters, but in each card only some of the characters were highlighted. Highlighted characters instantiate the relevant meaning. Following the space in Fig. 1A, we tested 11 referents/meanings: s, a, o, s+a, s+o, a+o, o+o, s+a+o, s+o+o, a+o+o and o+o+o. Given that three characters of different colors could represent the non-participant role (i.e. other), we generated two alternative cards for each meaning involving a non-participant other (i.e. o, o+o, s+o, a+o, o+o, s+a+o, s+o+o, a+o+o). Participants had to thus sort a total of 18 cards (Fig. 2B shows the 11 unique referents, with non-participant others shown in a random subset of colors).

At the end of the experiment, participants were asked to report any strategy they used to perform the sorting, and to answer a number of questions about their linguistic background, including their experience with other languages besides their native language.

Results

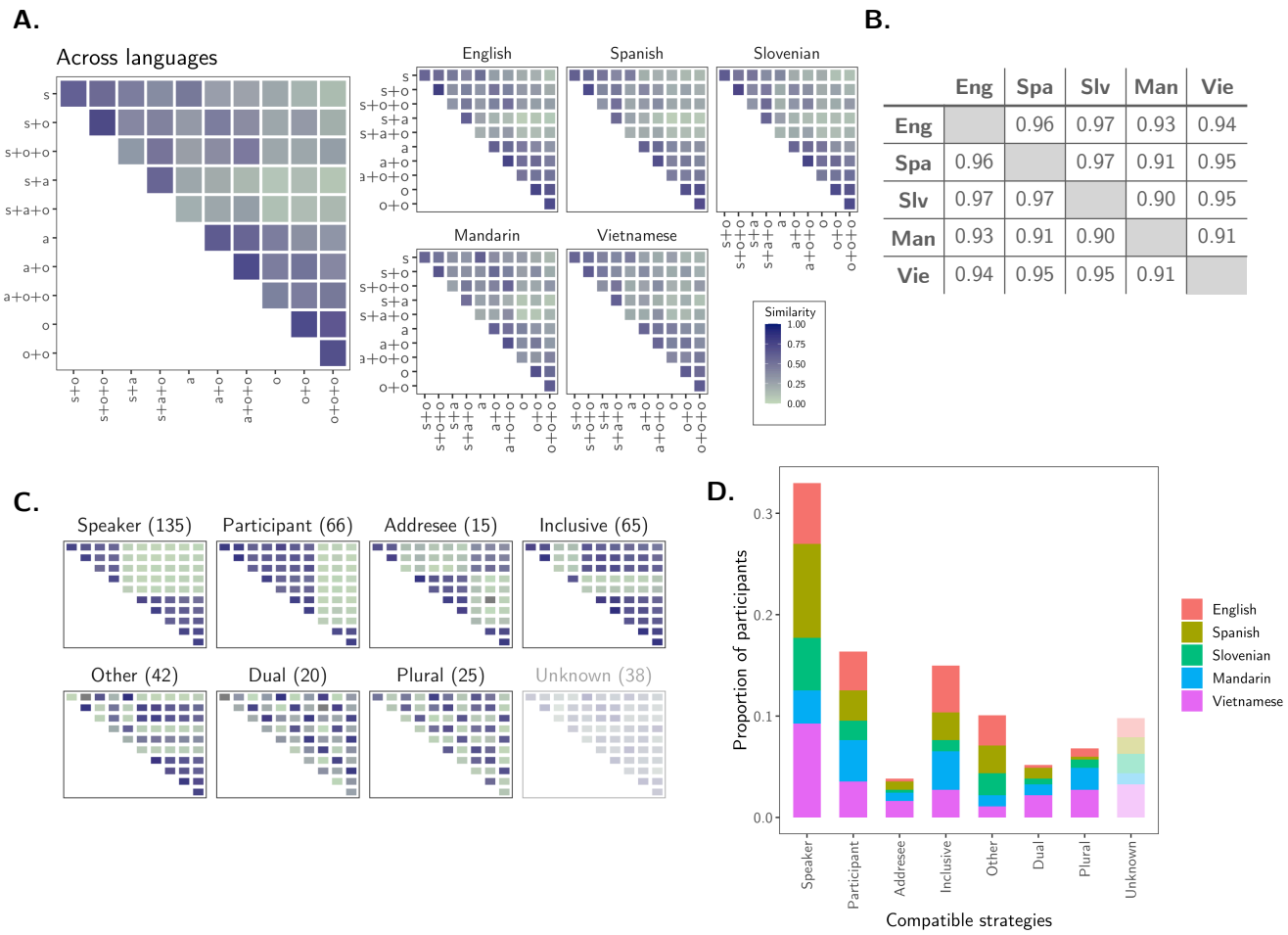
Recall that our goal is to assess the structure of the conceptual representation of pronominal referents based on participants' sorting strategies. Sorting data was thus used to compute similarity matrices, showing the frequency with which two referents (cards) were paired together in the same pile. Fig. 3A shows the resulting similarity matrices per language, as well as the overall similarity matrix, aggregated across languages.

Analysis 1: Evidence for a language-independent conceptual representation

Evidence for a representation of this conceptual space that is independent of language comes from whether sorting strategies are the same or different across the populations we tested. In order to estimate whether participants with different native languages relied on similar sorting strategies, we performed a pair-wise language comparison using a split-half technique. For each language pair, we computed the Spearman rank correlation between similarity matrices averaged over random halves of the data. We repeated this procedure 10,000 times for each pair. Fig 3B shows the averaged pairwise correlation scores for all languages in our study. Strong positive correlation scores reveal little variation in the similarity structure as a function of participants' native language. Despite using different pronominal systems, each group of participants tended to rely on similar strategies when sorting, suggesting that the task is not influenced by experience with a specific pronominal system. Crucially, this finding suggests that there is a common conceptual representation of the space, independent of the speaker's native language. Indeed, we can create a similarity matrix, averaged across languages (like the one in Fig. 3A), that is, in some sense, a proxy for this shared, language-independent representational space. Next, we wish to further explore the structure of this representational space and assess whether there are common underlying patterns, or structural primitives, that appear across speakers of different languages.

Analysis 2: Identifying structural primitives in the representational space

To explore whether there is a dominant representational structure emerging from this task that matches the hypothetical similarity structures shown in Fig. 1C, we use K-means clustering analysis on individual participant similarity matrices. We use 8 clusters since there are 8 possible representations of the domain obtained by exploiting the similarity between referents on the basis of each primitive (Fig. 1C). Fig. 3C illustrates the 8 clusters obtained from applying this method to the similarity matrices of 370 subjects. We further computed a Spearman correlation between each cluster and each of the eight hypothetical representations of the domain as defined in Fig. 1C. Clusters that have a strong positive correlation ($> .80$) with one of these hypothetical representations were classified as instantiating this representation. Fig. 3C shows the correspondence between clusters and representations, as well as how many participants fall in that cluster (between parentheses). Fig. 3D shows the proportion of participants in each cluster for each of the languages tested.



	Eng	Spa	Slv	Man	Vie
Eng		0.96	0.97	0.93	0.94
Spa	0.96		0.97	0.91	0.95
Slv	0.97	0.97		0.90	0.95
Man	0.93	0.91	0.90		0.91
Vie	0.94	0.95	0.95	0.91	

Figure 3: Results. **A**. Cross-language and language-specific similarity matrices estimated from the card sorting task. For every two referents in the domain, the corresponding cell in the similarity matrix indicates their average similarity across participants. **B**. Pair-wise correlations between the similarity matrices of different languages **C**. Cluster analysis of individual participants across all language yielded the eight clusters shown here. The label assigned to each cluster corresponds to the hypothetical strategy that most strongly correlates with it (see Fig. 1C). The number of participants assigned to each cluster is indicated in parentheses. **D**. Proportion of participants in each cluster for each of the languages in the study.

Notably, 7 out of the 8 clusters strongly correlate with some hypothetical representation of the domain. This suggests that most participants sort based on representations of the domain that make use of the hypothesized primitives. Put differently, most participants are sorting the cards on the basis of some feature of the referent, either the speech role(s) involved or its cardinality. Even more significantly, the most common strategy is based on whether the meaning involves the speaker. This supports what has been sometimes called a 'speaker'-bias, whereby referents that include the speaker in the conversational context are perceived as more similar to each other than those that do not (Zaslavsky et al., 2021). This is the most pervasive strategy both across and within languages, even for speakers of languages where the pronominal system does not exploit a speaker/non-speaker distinction, such as Vietnamese.[2]

Discussion

Theories and models of categorization in language have largely been built on the assumption of a share representational space for the domain in question. These spaces are generally designed to capture aspects of cross-linguistic variation (and constraints on that variation). However, in most cases, there is no evidence, other than the cross-linguistic data, for a universal mental representation. In other words, these shared representations are assumed without independent evidence. Here, we aimed to test whether there exists evidence for a

[2]This trend does not show up when looking specifically at Mandarin speakers, who are more uniformly distributed across Speaker, Participant and Inclusive strategies. However, given the high correlation between all the different languages, it is not entirely clear whether dividing participants based on language in the cluster analysis is justified.

shared, language-independent representation of person referents. We used a (non-linguistic) card sorting task to show that speakers of different languages indeed represent the structure of this space in similar ways. This suggests that there are at least some aspects of the mental representation of these concepts which are language-independent. This result has important implications for theories of person, as it constitutes the first evidence, outside of language, for the idea that there are *universal* constraints on pronominal systems: all humans share ways of representing this semantic domain, even when these structures do not immediately show up in their native languages. Importantly, this supports the intuition that the cross-linguistic regularities in the distribution of pronominal systems may be at least partially determined by representational constraints or biases.

Our findings are also informative about the specific properties of these common mental representations. A closer look into participants' strategies in our sorting task suggest that this domain is represented by relying on primitives such as the ones proposed in Fig. 1A; namely, speech roles and cardinalities. Moreover, some of these primitives seem to have a more prominent role than others. In line with previous work (Harbour, 2016; Zwicky, 1977; Zaslavsky et al., 2021; Maldonado & Culbertson, 2020), we find evidence for the speaker/non-speaker distinction being the most important, followed by the participant/non-participant distinction.

Two aspects of our findings require further discussion. First, participants in our experiment are not consistently relying on a single sorting strategy. There is a range of variation in strategies, which does not correlate with participants' native language. How can we interpret this variability? And does it really support the notion of a universal mental representation? One possibility is that the way speakers structure this domain depends on multiple biases—i.e., multiple competing notions of similarly, some of which are weighted more heavily then others. Different speakers might have different relative weightings, and different contexts might (probabilistically) elicit different weightings. But importantly, what we have shown is that the relatively weighting is *not* different for speakers of different languages. It is a further question whether this kind of highly non-uniform representation results in a better fit to the cross-linguistic data than the stipulated representation used in Zaslavsky et al. (2021). A natural next step, which we are currently exploring, is comparing this empirically-grounded representation with non-empirically grounded alternatives in order to test whether it lends a better account of the distribution of person systems.

It is also worth discussing whether the sample of languages we have tested is sufficient to be confident in positing a kind of *universal* mental representation (or set of shared representations). We tested speakers of 5 languages, which instantiate different pronominal systems, but of course, there are many other out there. Thus, while this is a first approximation of what a language-independent representation would look like, we have not included speakers of languages with pronominal systems which are maximally different from each other. An instance of a pronominal system maximally different from e.g. English can be found in Slavey (an Athabaskan language spoken in Canada). In Slavey, groups that involve any participant in the conversation (speaker, addressee, or both) are referred to by the same pronominal form. One could imagine that the inclusion of speakers like Slavey might shift the pervasiveness of certain strategies. Future research should look to replicate the task on speakers of an even broader set of languages.

Finally, in addition to providing an empirically-grounded, language-independent representation of similarity in the domain of pronominal referents, this study also makes a methodological contribution. It illustrates a method for testing mental representations on language-independent grounds which could be extended to other semantic domains with similar characteristics, including additional cardinalities/number systems, or the kind of meanings targeted by gender or noun class systems.

Acknowledgments

We thank Zhang Dingning, Ahn La, Jialing Liang, Trang Phan, Hongyuan Sun, Blaž Istenič Urh, Fang Wang, and Ruizhe Zhou, for help with translations and gathering of participants. This project was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757643) to J.C and by a K. Lisa Yang Integrative Computational Neuroscience (ICoN) Postdoctoral Fellowship to N. Z.

References

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of Memory and Language*, *53*(1), 60–80.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley and Los Angeles: University of California Press.

Boster, J., et al. (1994). The successive pile sort. *Cultural Anthropology Methods*, *6*(2), 7–8.

Denić, M., Steinert-Threlkeld, S., & Szymanik, J. (2022). Indefinite pronouns optimize the simplicity/informativeness trade-off. *Cognitive Science*, *46*(5), e13142.

Fincher, S., & Tenenberg, J. (2005). Making sense of card sorting data. *Expert Systems*, *22*(3), 89–93.

Gardenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. MIT press.

Harbour, D. (2016). *Impossible persons*. Cambridge, MA: MIT Press.

Harley, H., & Ritter, E. (2002). Person and Number in Pronouns: A Feature-Geometric Analysis. *Language*, *78*(3), 482–526. doi: 10.1353/lan.2002.0158

Kay, P., & McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. *Language*, *54*, 610-646.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1).

Lee, N. (2020). Learning (im)possible number syncretisms: Investigating innate featural representations. In *Proceedings of the 50th annual meeting of the north east linguistic society.* University of Massachusetts: Amherst.

Maldonado, M., & Culbertson, J. (2020). Person of interest: Experimental investigations into the learnability of person systems. *Linguistic Inquiry*, 1–71.

Maldonado, M., & Saldana, C. (2022). Interpretations of plurality in personal pronouns differ across person categories: An experimental study. In *Proceedings of the 23rd amsterdam colloquium.*

Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology*, *9*(2), 165–208.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *PNAS*, *104*(4), 1436-1441.

Steinert-Threlkeld, S. (2020). Quantifiers in natural language optimize the simplicity/informativeness trade-off. In *Proceedings of the 22nd amsterdam colloquium.*

Weller, S. C., & Romney, A. K. (1988). *Systematic data collection* (Vol. 10). Sage publications.

White, A., Malt, B. C., & Storms, G. (2017). Convergence in the bilingual lexicon: A pre-registered replication of previous studies. *Frontiers in Psychology*, *7*.

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57–70.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *PNAS*, *115*(31), 7937–7942.

Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Zwicky, A. M. (1977). Hierarchies of person. In *Proceedings from the Chicago Linguistic Society* (Vol. 13, pp. 714–733). Chicago, MI: Chicago Linguistic Society.