# Between versus Within-Language Differences in Linguistic Categorization

**Anne White (anne.white@ppw.kuleuven.be)**
**Gert Storms (gert.storms@ppw.kuleuven.be)**
Faculty of Psychology and Educational Sciences, University of Leuven
Tiensestraat 102, B-3000 Leuven, Belgium

**Barbara C. Malt (barbara.malt@lehigh.edu)**
Department of Psychology, Lehigh University
17 Memorial Drive East, Bethlehem, PA 18015-3068, USA

**Steven Verheyen (steven.verheyen@ens.fr)**
Institut Jean-Nicod, PSL Research University, École Normale Supérieure
Pavillon Jardin, 29 rue d'Ulm, 75005 Paris, France

**Abstract**

Cross-linguistic research has shown that boundaries for lexical categories differ from language to language. The aim of this study is to explore these differences *between* languages in relation to the categorization differences *within* a language. Monolingual Dutch- (N=400) and French-speaking (N=300) Belgian adults provided lexical category judgments for three lexical categories that are roughly equivalent in Dutch and French. Each category was represented by good, borderline, and bad examples. A mixture modeling approach enabled us to identify latent groups of categorizers within a language and to evaluate cross-linguistic variation in relation to within-language variation. We found complex patterns of lexical variation within as well as between language groups. Even within a seemingly homogeneous group of speakers sharing the same mother tongue, latent groups of categorizers display a variability that resembles patterns of lexical variation found at a cross-linguistic level of comparison.

**Keywords:** artifact categories; cross-linguistic differences; semantic variation; vagueness

## Introduction

People of different languages and cultures share a perception of the similarity among entities within at least some domains (e.g., common household containers: Ameel, Storms, Malt, & Sloman, 2005; Malt, Sloman, Gennari, Shi, & Wang, 1999; color: Roberson, Davies, Corbett, & Vandervyver, 2005: human locomotion: Malt, Ameel, Imai, Gennari, Saji, & Majid, 2014; and spatial relations: Munnich, Landau, & Dosher, 2001). Despite the shared non-linguistic appreciation of these domains, its relation with linguistic categorization is complex: Linguistic categories do not map directly onto similarity clusters (Ameel et al., 2005; Malt et al., 1999).

In different languages the world is carved up differently. This cross-linguistic variation has been shown for domains as varied as color, causality, mental states, number, body parts, containers, motion, direction, and spatial relations (Malt & Majid, 2013; Malt et al., 2015). Malt and colleagues, for instance, described how different languages label a set of household containers and found that not all languages observe the same distinctions, despite perceiving the similarity of the objects in the same way (Ameel et al., 2005; Malt, Sloman, & Gennari, 2003; Malt et al., 1999). For example, the Dutch word for bottle (*fles*) encompasses objects that in French are either called *bouteille* or *flacon*. Not only are there differences in the number of distinctions made in different languages, there is crosscutting in the way exemplars of a category are grouped together as well (Malt et al., 2003). The roughly equivalent French *bouteille* and Dutch *fles* demonstrate a difference in how they map onto a shared similarity space, which reflects a cross-linguistic difference in meaning representation. Additionally, the categories *fles* and *bouteille* each include a different number of objects, indicating differences in category extension as well.

Although these cross-linguistic differences have received growing attention in recent years, within-language variation exists as well (McCloskey & Glucksberg, 1978; Verheyen, Hampton, & Storms, 2010). Inter-individual differences in linguistic categorization have been described in the relation to vagueness (Black, 1937; Verheyen & Storms, 2013). A distinction is made between vagueness in criteria and vagueness in degree (Devos, 2003). The former is involved when individuals use different criteria to determine if an object belongs to a category. When individuals agree on the criteria for category membership but use a different cut-off for separating members from non-members, the latter type of vagueness is in play. In seemingly homogeneous groups of speakers of the same language, groups that display one or both types of differences have been identified (Verheyen & Storms, 2013).

The aim of this study is to quantitatively explore the extent and nature of differences in categorization between two language groups with respect to the variation existing between latent groups of categorizers within a language. More specifically this study evaluates the degree of within-language variability in relation to the degree of cross-linguistic variability for roughly equivalent categories. To this end we collected category judgment data from Belgian Dutch and French speaking participants, who share a similar environment and who perceive similarity in the tested domain in much the same way (Ameel et al., 2005).

A mixture modeling approach was used in order to identify latent groups of categorizers in a seemingly homogeneous group, that is, adult speakers of the same mother tongue. The mixture model partitions a participant sample into subgroups of individuals who display similar categorization behavior. By doing so, it identifies subgroups that use different criteria in making their category decisions (Verheyen & Storms, 2013; Verheyen, Voorspoels & Storms, 2015). We compared the categorization patterns of Dutch-speaking and French-speaking Belgian participants for the roughly equivalent categories *doos* and *boîte* (similar to English *box*), *fles* and *bouteille* (similar to English *bottle*) and *pot* and *pot* (similar to English *jar*). We also identified latent groups within each language, and the differences between the latent within-language groups were compared to those between languages.

# Method

## Participants

We collected data from approximately 400 monolingual Dutch-speaking and 300 monolingual French-speaking Belgian adult participants. Data from participants who were determined to be bilingual were discarded, as well as data of participants under the age of 17, resulting in the sample sizes displayed in Table 2.

## Materials

The stimulus material consisted of an existing set of pictures of household containers described by Ameel et al. (2005) expanded with new stimuli, totaling to 192 stimuli. The new pictures were made according to the guidelines used by Ameel et al. (2005). The objects were photographed in color against a neutral background with a constant camera distance to preserve relative size. A ruler was included in front of each object to provide additional size information.

Four lexical categories were investigated, selected based on unpublished naming data for the full set of 192 common household objects. The four most frequently generated category names across the complete set by 32 monolingual Dutch-speaking adult participants were *fles*, *pot*, *bus*, and *doos*. Because it was not feasible to conduct a categorization experiment with multiple categories for the complete set, a selection of 40 stimuli per category was made. In order to make an adequate selection of the stimuli, a pilot category judgment task with approximately 30 Dutch speaking participants was conducted. In the pilot study participants had to decide if a given name was suited for a presented object by responding 'yes' or 'no'.

The stimuli for this study were selected based on the results of the pilot study according to the following criteria. The selected set of stimuli spanned the full range of proportion of yes-responses, varying from approximately 0.10 to 0.90. This selection contained a mixture of clear members, borderline members, and clear non-members, spanning the range of shapes and sizes for the category. The earlier collected naming data of Dutch and French participants were also taken into account. Some objects that showed incongruities in the use of category labels between Dutch and French were included. For instance, a cooking *pot* was called *pot* by almost all Dutch participants, while in French this is rarely called *pot*.

For the French version of the task, four categories were selected based on naming data of adult monolingual French participants, analogous to the category selection for the Dutch version. The four most generated category names were *bouteille*, *pot*, *flacon*, and *boîte*. Because these categories do not map directly onto the categories for the Dutch task (Malt et al., 1999), the French task was composed as follows. The stimulus set for *bouteille* and *flacon* both consisted of the objects presented in the Dutch category *fles*. For the French category *boîte* the items belonging to the Dutch category *doos*, and for the French category *pot* the items belonging to the Dutch category *pot* were presented. The French speaking participants judged the same set of objects as the Dutch speaking participants, with the exception of two objects that were only presented in the Dutch category *bus*. Further descriptions will be limited to the roughly equivalent categories *doos-boîte*, *fles-bouteille* and *pot-pot*. The categories *flacon* and *bus* will be disregarded in the discussion of the results, since they do not have a roughly equivalent category in Dutch and French, respectively.

## Procedure

The linguistic categorization task was conducted via Qualtrics and the link to the task was distributed via social networks (e.g. Facebook), both for the French-speaking and Dutch-speaking participants. After informed consent, demographical information was collected (age, gender, education level, and mother tongue). The participants were instructed to decide for each item in the series whether or not it belonged to a particular category. The four categories were presented to participants in random order, as were the pictures within each category. Above every picture the question 'Is this an X?' was displayed, which participants could respond to by choosing 'yes' or 'no'. The instructions explicitly stated that pictured objects could belong to one or more categories and that there were no right or wrong answers. The full survey took between 10 to 15 minutes to complete.

## Model analyses

Item response theory (IRT) modeling can be used as a formalization of the Threshold Theory of Semantic Categorization (Hampton, 1995; Verheyen, Hampton, & Storms, 2010). The Threshold Theory (Hampton, 1995, 1998, 2007) accounts for vagueness in degree by allowing individuals to use a different cut-off or threshold along the criterion for category membership. In that case participants diverge only with respect to the category's extension.

The use of a mixture IRT model allows for the identification of subgroups within a seemingly homogeneous group, that use different criteria in their

category judgment. By allowing for subgroups, the assumption that all participants employ the same criterion is relaxed, and thus vagueness in criteria is accounted for. Within each of the identified groups, individuals can still differ in terms of their categorization cut-off.

Individual categorization decisions $Y_{ci}$, where $c$ refers to a categorizer and $i$ to an item, serve as input for the mixture model. When an item $i$ is endorsed as a category member by categorizer $c$, $Y_{ci}$ takes value 1; when it is not endorsed as a member of the target category it takes value 0. Each of these categorization judgments is regarded as an outcome of a Bernoulli trial with equation (1) modeling the probability of a positive response.

$$\Pr(Y_{ci} = 1) = \frac{e^{\alpha_g(\beta_{gi} - \theta_c)}}{1 + e^{\alpha_g(\beta_{gi} - \theta_c)}} \quad (1)$$

For each latent group $g$ of categorizers a separate criterion is extracted. The values for the parameters $\beta_{gi}$, $\theta_c$, and $\alpha_g$ are estimated by application of the model to an item by participant categorization matrix. The position of each item $i$ along the criterion of group $g$ is indicated by the estimate $\beta_{gi}$ and represents the extent to which that item meets the group's criterion. $\theta_c$ represents categorizer $c$'s threshold or cut-off: the extent to which items need to meet the criterion to be endorsed as category members. The relative position of $\theta_c$ and $\beta_{gi}$ defines the probability of endorsement: the more $\beta_{gi}$ exceeds $\theta_c$, the higher the probability that the item will be endorsed as a category member. Conversely, the more to the left of $\theta_c$ $\beta_{gi}$ is positioned, the lower the probability that the item will be endorsed. A separate $\alpha_g$ for each group determines the shape of the response function (Verheyen & Storms, 2013; Verheyen, Voorspoels, & Storms, 2015).

For each of the six (3 categories x 2 languages) data sets, the model in Equation (1) was estimated with 1, 2, 3, 4, and 5 different groups. The parameters in Equation (1) were estimated in a Bayesian manner, using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) running 3 chains of 10,000 samples each with a burn-in of 4,000 samples. The chains were checked for convergence and label switching. The reported results are based on the posterior means for the models that yielded the smallest Bayesian Information Criterion (BIC). These models yield the best approximation of the categorization data when both model fit and complexity are taken into account. See Verheyen, Voorspoels, and Storms (2015) for example code, rationale for prior specification, and model selection simulations.

## Results

### Cross-linguistic differences

We start by presenting the results for all participants of a language group. Figure 1 displays per item the categorization proportion of French participants for the French category *boîte* and of Dutch participants for the Dutch category *doos*. This graph confirms the idea that *boîte* and *doos* are roughly equivalent categories since the categorization proportions for French and Dutch seem to follow a roughly similar rising trend. However, several notable differences in categorization proportions can be observed as well.

The categorization proportions for the French *boîte* (M=0.45) are on average higher than the categorization proportions for the Dutch *doos* (M=0.40) ($t(39) = 3.287$, $p < 0.01$). The observation that *boîte* has a larger category extension than *doos* suggests a degree difference: the category *boîte* is somewhat larger than the category *doos*.

This is not the only difference between the two categories. There are also indications that the two language groups use different categorization criteria since the shape of the proportion curves is not the same. If the x-axis of the graph would be organized according to the categorization proportions of the French participants, one would obtain a different order of the items along the axis. This is demonstrated by an imperfect correlation of 0.78 between *doos* and *boîte*. Some of the French proportions are even lower than the corresponding Dutch ones, despite the established degree difference. So although the French category *boîte* includes more objects in general, some objects are not considered to be as good a category member as they are in the smaller Dutch category *doos*.

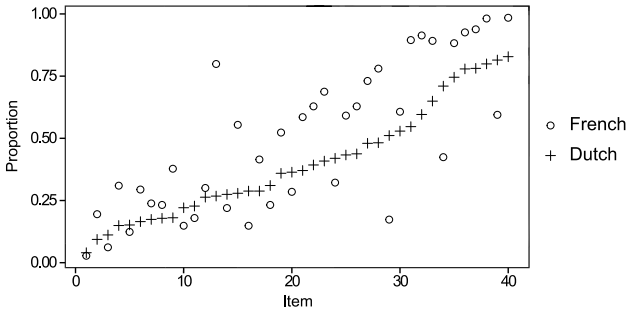


Figure 1: Categorization proportions per item for the French category *boîte* and the Dutch category *doos*. The order of the items is determined according to the Dutch categorization proportions.

As to the other categories[1], *bouteille* (M=0.24) displays significantly lower categorization proportions than *fles* (M=0.38) ($t(39) = -6.213$, $p < 0.001$), and *pot* (F) (M=0.33) does not display a significant difference with *pot* (D) (M=0.35) ($t(39) = -1.354$, $p = 0.1837$). This indicates a difference in degree for the categories *fles-bouteille*, but not for *pot-pot*. The correlations between *fles-bouteille* (0.83) and *pot-pot* (0.74) suggest that both category pairs show a difference in criteria.

[1] Due to space limitations only the graphs for the categories *doos* and *boîte* are displayed. Similar graphs were made for the categories *fles–bouteille* and *pot–pot*. These figures were similar to those presented and lead to the same conclusions. A discussion of the results using *doos–boîte* was preferred since this category yields the same number of latent groups in Dutch and French and thus allows for a straightforward comparison.

## Within-language differences

To study within-language differences we identified latent groups of categorizers using the mixture IRT-approach. Table 1 shows the BIC values for every category, for partitionings[2] in one to five groups. The solution with the lowest BIC (indicating the appropriate number of subgroups to consider) is in bold typeface.

Table 1: BIC values for five partitions of the categorization data with the number above each column representing the number of groups.

| category | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| *boîte* | 11256 | 10870 | **10695** | 10917 | 11159 |
| *doos* | 16838 | 16142 | **16000** | 16213 | 16221 |
| *pot* (F) | 10121 | **10097** | 10290 | 10530 | 10772 |
| *pot* (D) | 14170 | **13788** | 14040 | 14296 | 14549 |
| *bouteille* | 9015 | **8645** | 8859 | 9088 | 9323 |
| *fles* | 15650 | 14844 | **14802** | 15054 | 15310 |

Table 2 presents the total number of participants per category and also the total number of participants per latent group. Since the latent groups were determined by the mixture model, the participants are not necessarily evenly divided over the identified groups corresponding to the partitionings with the lowest BIC value.

Table 2: Overview of the number of respondents per group.

| category | **all** | Group 1 | Group 2 | Group 3 |
|---|---|---|---|---|
| *boîte* | **322** | 60 | 71 | 191 |
| *doos* | **448** | 77 | 167 | 204 |
| *pot* (F) | **310** | 27 | 283 | / |
| *pot* (D) | **424** | 192 | 232 | / |
| *bouteille* | **308** | 110 | 198 | / |
| *fles* | **436** | 47 | 180 | 209 |

Figure 2 and Figure 3 display the categorization proportions in the three latent subgroups for the French category *boîte* and the Dutch category *doos*. The amount of variability within each language is striking. Even within a language, there appears to be only limited consensus with respect to categorization decisions. While one may expect some disagreement in the middle part of the curve for the borderline objects, there is also considerable disagreement at the ends of the curve, which should hold the clear members and clear non-members of the category for which one would expect to find strong agreement among speakers of the same mother tongue.

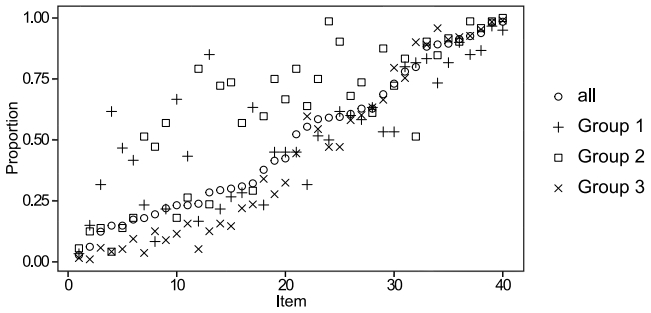


Figure 2: Categorization proportions for the French participants for the category *boîte* per item and per latent group with *all* referring to the average categorization proportion of the entire sample of French participants.

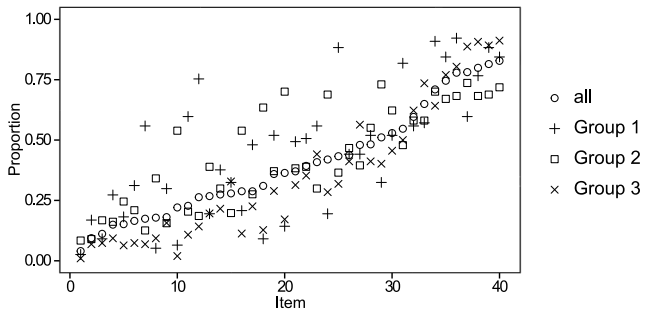


Figure 3: Categorization proportions for the Dutch participants for the category *doos* per item and per latent group with *all* referring to the average categorization proportion of the entire sample of Dutch participants.

This finding is not the result of the mixture analysis yielding smaller groups of categorizers who show more variability in their categorization data. The reliabilities[3] shown in Table 3 refute the possibility that the variability displayed in the graphs is due to unreliability. All reliabilities exceed 0.95 demonstrating that the participants in a subgroup performed in much the same way.

Table 3: Reliabilities per latent group of categorizers and for the complete group of participants per category.

| category | **all** | Group 1 | Group 2 | Group 3 |
|---|---|---|---|---|
| *boîte* | **0.995** | 0.961 | 0.981 | 0.995 |
| *doos* | **0.993** | 0.975 | 0.981 | 0.992 |
| *pot* (F) | **0.993** | 0.956 | 0.999 | / |
| *pot* (D) | **0.995** | 0.991 | 0.993 | / |
| *bouteille* | **0.992** | 0.983 | 0.990 | / |
| *fles* | **0.995** | 0.968 | 0.991 | 0.992 |

[2] For the categories *fles* and *pot* (D) a group consisting of, respectively, only one and ten participants was identified. In those cases the analyses were repeated without these participants, resulting in the numbers shown in Table 2.

[3] Reliability was evaluated by applying the split-half method, followed by the Spearman–Brown correction. The displayed reliability is the average reliability across 10,000 random splits.

If the mixture analyses succeeded in identifying homogeneous groups of categorizers, the reliability within each group of latent categorizers should be higher than the reliability for the language group as a whole (if the groups are equated for number of participants). To evaluate whether this is true, a sampling procedure was used. Table 4 displays the average split-halves reliability across 10,000 random splits of the data of 25 randomly drawn participants from either the complete language group (all) or one of the latent groups within a language.

With the exception of two groups (*doos* group 2 and *boîte* group 1) the data indeed show the pattern we expected, showing that the variability we observe in the graphs is not random variation but reflects meaningful between-group-differences within a language. This is the case for the latent groups within both the Dutch and French language groups.

Table 4: Average split-half reliability across 10 000 random samples of 25 participants.

| category | **all** | Group 1 | Group 2 | Group 3 |
|---|---|---|---|---|
| *boîte* | **0.940** | 0.911 | 0.946 | 0.963 |
| *doos* | **0.892** | 0.926 | 0.883 | 0.938 |
| *pot* (F) | **0.920** | 0.952 | 0.924 | / |
| *pot* (D) | **0.924** | 0.934 | 0.938 | / |
| *bouteille* | **0.908** | 0.930 | 0.922 | / |
| *fles* | **0.917** | 0.940 | 0.936 | 0.939 |

The sampling procedure was repeated for the two language groups together, that is, drawing 10,000 random samples of 25 participants out of the complete set for both languages together. The split-half reliability is 0.904 for all category pairs (*doos-boîte*, *fles-bouteille*, and *pot-pot*). One would expect that adding another language group to the dataset adds considerable variability to the data. Therefore, reliability should show a notable decrease in comparison with the reliability calculated within a language group (Table 4, first column). However, the reliability calculated over language groups is only slightly lower compared to the reliability calculated within one language group. This finding suggests that the within-language variability is comparable to the cross-linguistic variation.

Describing the identified variability between latent groups of categorizers in terms of differences in criteria and degree can be done both within and between languages. The strength of the correlations between categorization proportions can be interpreted as the extent to which different criteria are used. Differences in means between categorization proportions reflect a difference in degree.

The correlations in Table 5 vary from 0.20 to 0.86. Within-language correlations are displayed in the light gray area, and cross-language correlations are displayed in the dark gray area. The mixture analyses indeed succeeded in separating maximally different groups within one language, since the highest correlation between the categorization proportions of two groups of the same language is 0.77. It is quite striking that the correlations found for latent groups within a language do not exceed the correlation between the language groups (0.78). It also becomes clear that there are groups of categorizers who show a higher correlation with latent groups of the other language compared to correlations within their language group. For example, the categorization proportions for the Dutch *doos* of Group 1 resemble the proportions for the French *boîte* of Group 2 more than they resemble the other groups of their own language. Table 5 is a clear demonstration of the complexity of within-language and cross-linguistic variation in categorization patterns.

Table 5: Correlations of the categorization proportions for Dutch (D) and French (F) participants per latent group for the category pair *doos-boîte*.

| | D2 | D3 | F1 | F2 | F3 |
|---|---|---|---|---|---|
| D1 | 0.369 | 0.748 | 0.417 | 0.856 | 0.596 |
| D2 | | 0.706 | 0.825 | 0.200 | 0.473 |
| D3 | | | 0.717 | 0.707 | 0.818 |
| F1 | | | | 0.436 | 0.772 |
| F2 | | | | | 0.757 |

Taking into account differences in degree makes the comparison even more complex, since the strength of the correlation is not related to whether or not there is a difference in means. For example, comparing the average categorization proportions of D1 with those of F1 results in $t(39) = -4.7616$, $p < 0.0001$, whereas the comparison D2 – F3 results in $t(39) = 0.3549$, $p = 0.7246$. Both show a clear difference in criteria, but only the former shows a significant difference in degree. Of the groups that show a better correspondence in the used criteria, comparing the average categorization proportions of D3 and F3 shows no degree difference ($t(39) = -0.1741$, $p = 0.8627$), whereas the comparison D1-F2 does ($t(39) = -5.843$, $p < 0.0001$). Similar observations can be made for the comparison of latent groups of the same language.

Drawing a straightforward conclusion regarding cross-linguistic differences becomes more complex if one takes into account that there are latent groups that show higher correspondence with latent groups of another language group than they do with latent groups of the same language. Averaging over latent groups will in this case distort the comparison on a cross-linguistic level. This applies for both differences in criteria and degree.

## Conclusion and discussion

The purpose of this study was to evaluate cross-linguistic lexical categorization differences relative to the categorization differences that exist between latent groups within each language. When comparing the variability displayed in Figure 1 versus Figures 2 and 3, the degree of variability within a language is higher than expected. One would expect the variability within a language to be less pronounced in comparison to cross-linguistic differences. Non-linguistic appreciation of properties of domains seems

to be universal (at least for some domains, including the one studied here), but the relation of this non-linguistic understanding to linguistic categorization is complex. That is, linguistic categories do not map directly onto similarity clusters (Malt et al., 1999). These complex patterns of lexical variation for categories of everyday objects emerge not only between languages but within a language as well. Especially the latter differences seem to be more complex than earlier assumed. Vagueness in degree and criteria seem to cause complex patterns of lexical variation between latent groups of categorizers that resemble the patterns of lexical variation at a cross-linguistic level.

The amount of variability observed within one language poses a challenge for cross-linguistic research. It is common practice in cross-linguistic research not to take into account within-language differences and to average across all individuals within a language, provided the sample comes from a restricted geographic region, implying a shared dialect. This may lead to conclusions that do not hold for the latent groups a language might harbor. For example, based on Figure 1 one might believe that the difference between the categories *doos* and *boîte* mainly consists of a difference in degree. The correlation of 0.78 between the language groups is imperfect but points out that there is a substantial agreement between both language groups as well. However, taking into account the within-language differences, it becomes clear that this conclusion could vary a great deal depending on the combination of latent groups, since these correlations vary from 0.20 till 0.86.

Future research will pinpoint possible causes for the observed variation. A possible path involves relating personal characteristics (age, gender, education level) and item characteristics to the parameter estimates of the different latent groups.

How are we able to manage the considerable inter-individual differences during the communication process and prevent a breakdown in communication? Possible answers to this question may lie in the way polysemy or words with new meanings such as eponyms are dealt with during communication using processes of sense creation and selection (Clark & Gerrig, 1983; Foraker & Murphy, 2012). Even for common nouns, referring to familiar objects in their most literal sense, these processes are relevant in the context of inter-individual differences.

## Acknowledgments

Anne White is a research assistant at the Research Foundation-Flanders (FWO-Vlaanderen). Steven Verheyen was funded by ANR project TriLogMean (ANR-14-CE30-0010).

## References

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

Black, M. (1937). Vagueness. An exercise in logical analysis. *Philosophy of Science, 4*, 427-455.

Clark, H. H., & Gerrig, R. J. (1983). Understanding old words with new meanings. *Journal of Verbal Learning and Verbal Behavior*, 22(5), 591-608.

Devos, F. (2003). Semantic vagueness and lexical polyvalence. *Studia Linguistica, 57*, 121-141.

Foraker, S., & Murphy, G. L. (2012). Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language*, 67(4), 407-425.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language, 34*, 686–708.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*, 137–165.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, 31, 355-384.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhaltery, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-337.

Malt, B., Ameel, E., Imai, M., Gennari, S. P., Saji, N., & Majid, A. (2014). Human locomotion in languages: Constraints on moving and meaning. *Journal of Memory and Language, 74*, 107–123.

Malt, B., Sloman, S. A., & Gennari, S. P. (2003). Universality and language specificity in object naming. *Journal of Memory and Language, 49*(1), 20–42.

Malt, B. C., Gennari, S., Imai, M., Ameel, E., Saji, N., & Majid, A. (2015). Where are the concepts? What words can and can't reveal. In E. Margolis & S. Laurence (Eds.), *The conceptual Mind: New directions in the study of concepts*. Cambridge, MA: MIT Press.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language, 40*, 230-262.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6*, 462-472.

Munnich, E., Landau, B., & Dosher, B. A. (2001). Spatial language and spatial representation: a cross-linguistic comparison. *Cognition, 81*, 171-208.

Roberson, D., Davies, I. R. L., Corbett, G. G., & Vandervyver, M. (2005). Free-Sorting of colors across cultures: Are there universal grounds for grouping? *Journal of Cognition and Culture, 5*, 349 – 386.

Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta Psychologica, 135*, 216-225.

Verheyen, S., & Storms, G. (2013). A Mixture approach to vagueness and ambiguity. *PLoS ONE, 8*.

Verheyen, S., Voorspoels, W., & Storms, G. (2015). Inferring choice criteria with mixture IRT models: A demonstration using ad hoc and goal-derived categories. *Judgment and Decision Making, 10*, 97-114.