Natural color categories are convex sets

Gerhard Jäger

University of Tübingen
Department of Linguistics
gerhard.jaeger@uni-tuebingen.de

Abstract. The paper presents a statistical evaluation of the typological data about color naming systems across the languages of the world that have been obtained by the World Color Survey. In a first step, we discuss a principal component analysis of the categorization data that led to a small set of easily interpretable features dominant in color categorization. These features were used for a dimensionality reduction of the categorization data.
Using the thus preprocessed categorization data, we proceed to show that available typological data support the hypothesis by Peter Gärdenfors that the extension of color category are convex sets in the CIELab space in all languages of the world.

1 Introduction: The World Color Survey

In their seminal study from 1969, Berlin and Kay investigated the color naming systems of twenty typologically distinct languages. They showed that there are strong universal tendencies both regarding the extension and the prototypical examples for the meaning of the basic color terms in these languages.

This work sparked a controversial discussion. To counter the methodological criticism raised in this context, Kay and several co-workers started the **World Color Survey** project (WCS, see Cook et al. 2005 for details), a systematic large-scale collection of color categorization data from a sizeable amount of typologically distinct languages across the world.

To be more precise, the WCS researchers collected field research data for 110 unwritten languages, working with an average of 24 native speakers for each of these languages. During this study, the Munsell chips were used, a set of 330 chips of different colors covering 322 colors of maximal saturation plus eight shades of gray.

The main chart is a 8×40 grid, with eight rows for different levels of lightness, and 40 columns for different hues. Additionally there is a ten-level column of achromatic colors, ranging from white via different shades of gray to black. The level of granularity is chosen such that the difference between two neighboring chips is minimally perceivable.

For the WCS, each test person was "asked (1) to name each of 330 Munsell chips, shown in a constant, random order, and (2), exposed to a palette of these chips and asked to to pick out the best example(s) ('foci') of the major terms

elicited in the naming task" (quoted from the WCS homepage). The data from this survey are freely available from the WCS homepage http://www.icsi.berkeley.edu/wcs/data.html.

This invaluable source of empirical data has been used in a series of subsequent evaluations that confirming Berlin and Kay's hypothesis of universal tendencies in color naming systems across languages (see for instance Kay and Maffi 1999), even though the controversy about universality vs. relativism continues.

2 Feature extraction

For each informant, the outcome of the categorization task defines a partition of the Munsell space into disjoint sets — one for each color term from their idiolect.

An inspection of the raw data reveals — not surprisingly — a certain level of noise. This may be illustrated with the partitions of two speakers of a randomly chosen language (Central Tarahumara, which is spoken in Mexico). They are visualized in Figure 1. In the figure, colors represent color terms of Cen-

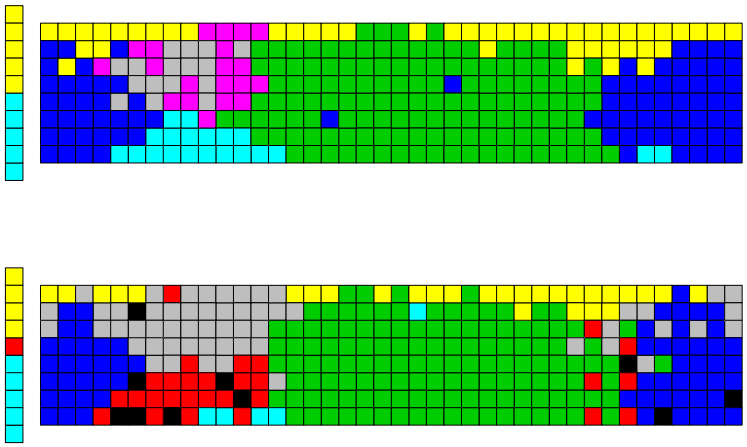

Fig. 1. Partitions for two speakers of Central Tarahumara

tral Tarahumara. We see striking similarities between the two speakers, but the identity is not complete. They have slightly different vocabularies, and the extensions of common terms are not identical. Furthermore, the boundaries of the extensions are unsharp and appear to be somewhat arbitrary at various places. Also, some data points, like the two blue chips within the green area in the center of the upper chart, seem to be due to plain mistakes. Similar observations apply to the data from other participants.

To separate genuine variation between categories (of the same or of different speakers, from the same or from different languages) on one hand from random

variation due to the method of data collection on the other hand, I employed **principal component analysis** (PCA), a standard technique for feature extraction and dimensionality reduction that is widely used in pattern recognition and machine learning.

The extension of a given term for a given speaker is a subset of the Munsell space. This can be encoded as a 330-dimensional binary vector. Each Munsell chip corresponds to one dimension. The vector has the value 1 at a dimension if the corresponding chip belongs to the extension of the term in question, and 0 otherwise. By using this encoding I obtained a collection of 330d vectors, one for each speaker/term pair.

PCA takes a set of data points in a vector space as input and linearly transforms the coordinate system such that (a) the origin of the new coordinate system is at the mean of the set of points, and (b) the new dimensions are mutually stochastically independent regarding the variation within the data points. The new dimensions, called **principal components**, can be ordered according to the variance of the data points along that dimension.

One motivation for performing a PCA is **dimensionality reduction**. Suppose the observed data points are the product of superimposing two sources of variation — a large degree of "genuine" or "interesting" variation and a small degree of irrelevant noise (and the latter is independent of the former). Then PCA is a way to separate the former from the latter. If the observed data live in an n-dimensional vector space but the genuine variation is m-dimensional (for $m < n$), then the first m principal components can serve as an approximation of this genuine variation.

In our domain of application, "interesting" variation is the variation between the extensions of different categories, like the difference between the extensions of English "red" and English "green" or between the extensions of English "blue" and Russian "galubòj" (which denotes a certain light blue). Inessential variation is the variation between the extensions that two speakers (of the same dialect of) the same language assign to the same term. It is plausible to assume the latter to be small in comparison to the former. So as a heuristic, we can assume that the first m principal components (for some $m < 330$ that is yet to be determined) capture the essence of the "interesting" variation.

Figure 2 depicts the proportion of the total variance in the data explained by the principal components. The graphics does not motivate a specific choice of m. For the time being, I will choose $m = 10$ because, as we will see shortly, the first ten principal components can be interpreted straightforward, while the others can't. The main result of the paper does not depend on this choice though. The first ten principal components jointly explain about 62.0% of the total variance in the data. Each of the following 320 principal components only explains a small additional proportion of variance of less than 1%.

It is worthwhile to look at the first ten principal components in some detail. Figure 3 gives a visualization. Please note that each principal component is a vector in the 330d space defined by the Munsell chips. The degree of lightness of each chip in the visualization corresponds to the value of the principal component

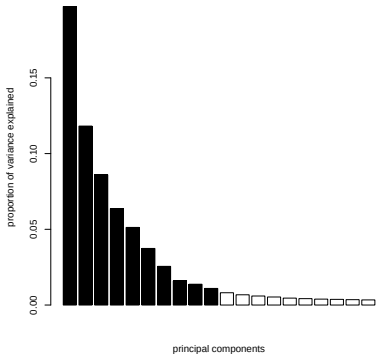

Fig. 2. Proportion of total variance explained by prinicpal components

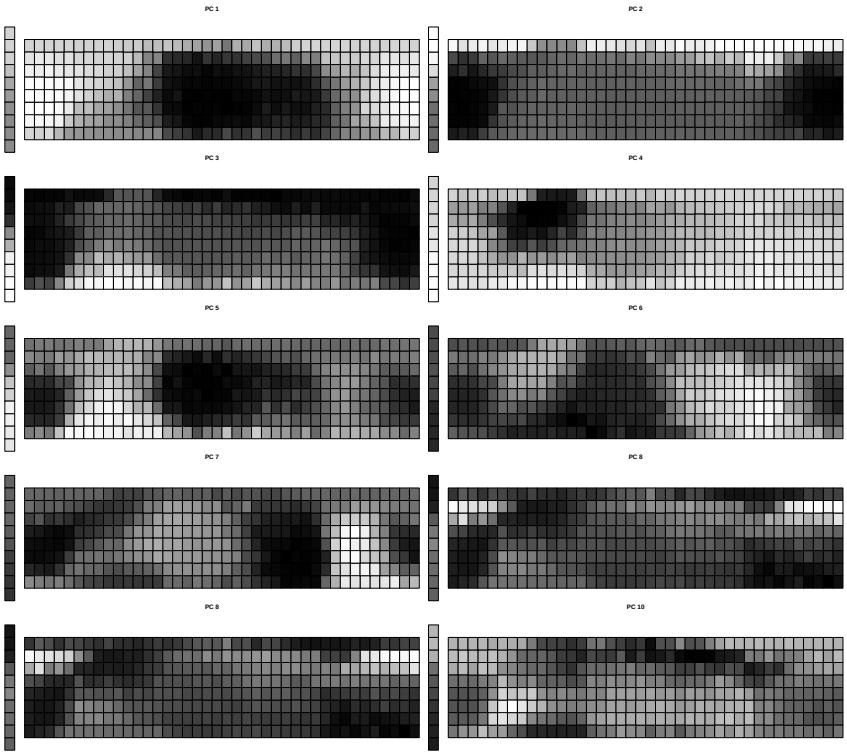

Fig. 3. Visualization of the first ten principal components

in question in the corresponding dimension. The values are scaled such that black stands for the maximal and white for the minimal value, whatever their absolute numerical value may be. Also note the directionality of principal components is arbitrary — so inverting a chart would result in a visualization of the same principal component. The important information is where the regions of extreme values (black or white) are located, in opposition to gray, i.e. the non-extreme values.

In all ten charts, we find clearly identifiable regions of extreme values. They are listed in Table 1. With very few exceptions, the thus identified regions ap-

Table 1. Oppositions defined by the first ten principal components

PC	extreme negative values	extreme positive values
1	red, yellow	green, blue
2	white	red
3	black	white, red
4	black, red, blue, purple	yellow
5	black, brown	red, green, blue
6	blue	red, black, green
7	purple	red, orange, blue
8	pink	red, orange, yellow, white, purple
9	pink, orange	black
10	brown	black, light green, light blue

proximately correspond to (unions of) ten of the eleven universal basic color terms identified by Berlin and Kay (1969). (The only universal basic color that does not occur is gray. This is likely due to the fact that shades of gray are under-represented in the Munsell chart in comparison to shades of other basic colors. The absence of gray is thus likely an artefact of the way the data in the WCS were collected.) Remarkably, the first six principal components jointly define exactly the six primary colors black, white, red, green, blue and yellow. (Purple has extreme values for PC4, but it is not distinguished from the neighboring red and blue.) The 7th – 10th principal components additionally identify the composite colors purple, brown, orange and pink. The 10th prinicipal component furthermore identifies another composite color between green/blue and white.

As can be seen from this discussion, the 10th principal component is less clearly interpretable than the first nine. The remaining principal components starting with the 11th lend themselves even less to an intuitive interpretation.

3 Dimensionality reduction

The first ten principal componentens define a linear 10d subspace of the original 330d space. We are operating under the assumption now that most of the "interesting" variation between color categories takes place within this low-dimensional

subspace, while variation outside this subspace is essentially noise. As the next step, I projected the original 330d data points to that subspace. Technically this means that in the transformed coordinate system defined by PCA, only the first ten dimensions are considered, and the values of all data points for the other 320 dimensions are set to 0. The resulting vectors are transformed back into the original coordinate system.

If visualized as a chart of gray values, the original data points correspond to black-and-white pictures where the extension of the corresponding category is a black region with jagged edges. After dimensionality reduction, we get dark regions with smooth and fuzzy gray borders. Put differently, while the original data points are classical binary sets with sharp and jagged boundaries, the projected data points are fuzzy sets with smooth boundaries.[1] (Technically speaking this is not entirely true because the values of the vectors after dimensionality reduction may fall slightly outside the interval $[0, 1]$, but the notion of a fuzzy set is still a good conceptual description.) Figure 4 contains two randomly chosen examples of data points before and after dimensionality reduction.

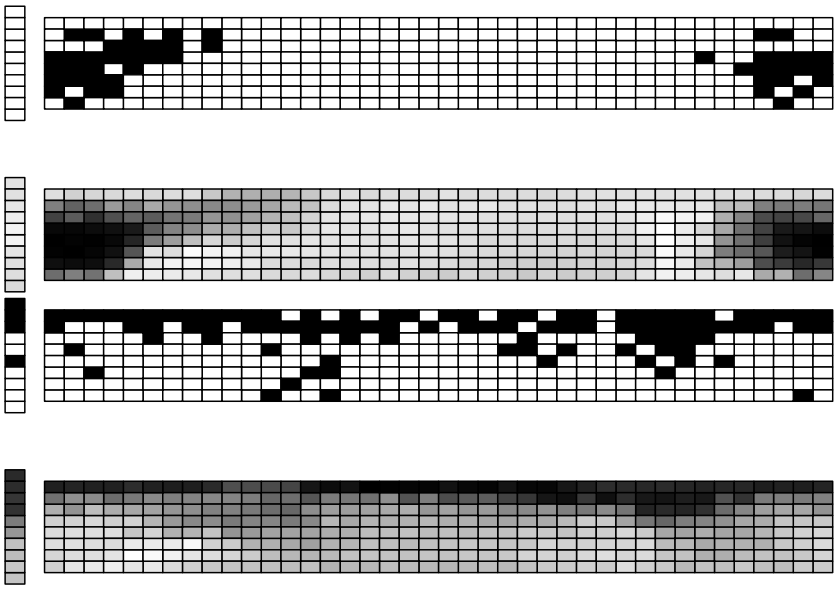

Fig. 4. Dimensionality reduction

For a given speaker, we can now determine for each Munsell chip which category has the highest value (after dimensionality reduction). In this way we

[1] The idea that the extensions of color categories are best modeled as fuzzy sets has been argued for on the basis of theoretical considerations by Kay and MacDaniel (1978).

can assign a unique category to each chip, and we end up with a partition of the color space again. The boundaries of the categories are sharp again, but in most cases not jagged but smooth. As an illustration, the cleaned-up versions of the partitions from Figure 1 are given in Figure 5.

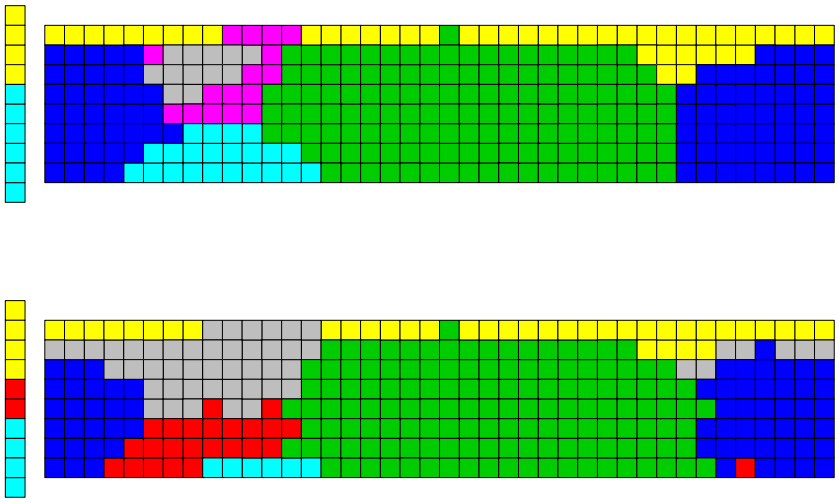

Fig. 5. Cleaned-up partitions for the two speakers of Central Tarahumara

4 Convexity in the CIELab space

The visualizations discussed so far suggest the generalization that after dimensionality reduction, category extensions are usually contiguous regions in the 2d Munsell space. This impression becomes even more striking if we study the extensions of categories in a geometrical representation of the color space with a psychologically meaningful distance metric. The CIELab space has this property. It is a 3d space with the dimension L* (for lightness), a* (the green-red axis) and b* (the yellow-blue axis). The set of perceivable colors forms a three-dimensional solid with approximately spherical shape. Figuratively speaking, white is at the north pole, black at the south pole, the rainbow colors form the equator, and the gray axis cuts through the center of the sphere. The CIELab space has been standardized by the "Commission Internationale d'Eclairage" such that Euclidean distances between pairs of colors are monotonically related to their perceived dissimilarity.

The 320 chromatic Munsell colors cover the surface of the color solid, while the ten achromatic chips are located at the vertical axis. Visually inspecting CIELab representations of the (dimensionality-reduced) partitions led to the hypothesis that boundaries between categories are in most cases approximately linear, and extensions of categories are convex regions. This is in line with the main

claim of Gärdenfors' (2000) book "Conceptual Spaces". Gärdenfors suggests that meanings can always be represented geometrically, and that "natural categories" must be convex regions in such a conceptual space. The three-dimensional color space is one of his key examples.

We tested to what degree this prediction holds for the partitions obtained via dimensionality reduction. The algorithm we used can be described as follows. Suppose a partition $p_1, \cdots, p_k$ of the Munsell colors into k categories is given.

1. For each pair of distinct categories p_i, p_j (with $1 \leq i, j \leq k$), find a linear separator in the CIELab space (i.e. a plane) that optimally separates p_i from p_j. This means that the set of Munsell chips is partitioned into two sets $\tilde{p}_{i/j}$ and $\tilde{p}_{j/i}$, that are linearly separable, such that the number of items in $p_i \cap p_{j/i}$ and in $p_j \cap p_{i/j}$ is minimized.
2. For each category p_i, define

$$\tilde{p}_i \doteq \bigcap_{j \neq i} p_{i/j}$$

As every $p_{i/j}$ is a half-space and thus convex, and the property of convexity is preserved under set intersection, each $\tilde{p}_i$ is a convex set (more precisely: the set of Munsell coordinates within a convex subset of R^3).

To perform the linear separation in a first step, I used a soft-margin Support Vector Machine (SVM). An SVM (Vapnik and Chervonenkis 1974) is an algorithm that finds a linear separator between two sets of labeled vectors in an n-dimensional space. An SVM is soft-margin if it tolerates misclassifications in the training data.[2] As SVMs are designed to optimize generalization performance rather than misclassification of training data, it is not guaranteed that the linear separators found in step 1 are really optimal in the described sense. Therefore the numerical results to be reported below provide only a lower bound for the degree of success of Gärdenfors' prediction.

The output of this algorithm is a re-classification of the Munsell chips into convex sets (that need not be exhaustive). The *degree of convexity* "conv" of a partition is defined as the proportion of Munsell chips not re-classified in this process. If $p(c)$ and $\tilde{p}(c)$ are the class indices of chip c before and after re-classification, and if $\tilde{p}(c) = 0$ if $c \notin \bigcup_{1 \leq i \leq n} \tilde{p}_i$, we can define formally:

$$\text{conv} \doteq |\{c | p(c) = \tilde{p}(c)\}|/330$$

The mean degree of convexity of the partitions obtained via PCA and dimensionality reduction is 93.9%, and the median is 94.5% (see the first boxplot in Figure 6). If the above algorithm is applied to the raw partitions rather than to those obtained via dimensionality reduction, the mean degree of convexity is 77.9%.

[2] The main reasons for the popularity of SVMs in statistical learning are that they are easily adaptable to non-linear classification tasks and that they find separators that generalize well to unseen data. These features are of lesser importance here. See (Schölkopf and Smola, 2002) for a comprehensive account.

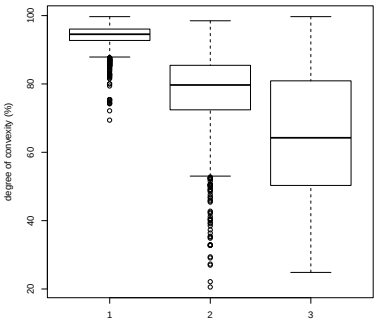

Fig. 6. Degrees of convexity (in %) of 1. cleaned-up partitions, 2. raw partitions, and 3. randomized partitions

Since the difference between these values is considerable, one might suspect the high degree of convexity for the cleaned-up data actually to be an artifact of the PCA algorithm and not a genuine property of the data. This is not very plausible, however, because the input for PCA were exclusively categorization data from the WCS, while the degree of convexity depends on information about the CILab space. Nevertheless, to test this hypothesis, I applied a random permutation of the category labels for each original partition and applied the same analysis (PCA, dimensionality reduction, computation of the degree of convexity) to the thus obtained data. The mean degree of convexity for these data is as low as 65.3% (see the third boxplot in Figure 6). The fact that this value is so low indicates the high average degree of convexity to be a genuine property of natural color category systems.

The choice of $m = 10$ as the number of relevant principal component was motivated by the fact that only the first ten prinicpal components were easily interpretable. As this is a subjective criterion, it is important to test to what degree the results from this section depend on this choice.

Therefore I performed the same analysis with the original data for all values of m between 1 and 50. The dependency of the mean degree of convexity on m is displayed in figure 7. It can be seen that the degree of convexity is not very sensitive to the choice of m. For all values of $m \leq 35$, mean convexity is above 90%. The baseline is the degree of convexity of 77.9% for the raw data (or, equivalently, for $m = 330$), which is indicated by the horizontal line.

So I conclude that the data from the WCS provide robust support for Gärdenfors' thesis.

References

Berlin, B., Kay, P.: Basic color terms: their universality and evolution. University of California Press, Chicago (1969)

Cook, R., Kay, P., Regier, T.: The world color survey database: History and use. In Cohen, H., Lefebvre, C., eds.: Handbook of Categorisation in the Cognitive Sciences. Elsevier (2005) 223–242

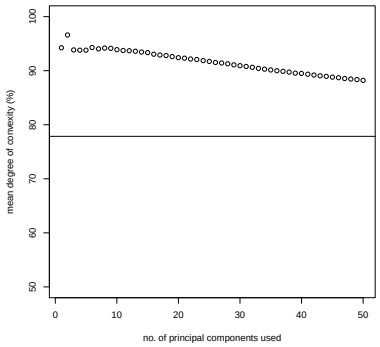

Fig. 7. Mean degree of convexity as a function of m

Kay, P., Maffi, L.: Color appearance and the emergence and evolution of basic color lexicons. American Anthropologist (1999) 743–760

Kay, P., McDaniel, C.K.: The linguistic significance of the meanings of basic color terms. Language **54**(3) (1978) 610–646

Gärdenfors, P.: Conceptual Spaces. The MIT Press, Cambridge, Mass. (2000)

Vapnik, V., Chervonenkis, A.: Theory of pattern recognition [in Russian]. Nauka, Moscow (1974)

Schölkopf, B., Smola, A.J.: Learning with Kernels. Support Vector Machines, Regularization, Opimization, and Beyond. MIT Press, Cambridge (Mass.) (2002)

Pluralities in Concealed Questions, Interrogative Clauses and Individuals

Maribel Romero
University of Konstanz

Concealed question Noun Phrases (NPs) like *the capital of Italy* in (1) have been analysed as contributing their intension --an individual concept-- to the semantic computation, as sketched in (2)-(4) (Heim 1979, Romero 2005, Aloni 2008):

(1) Mary knows / guessed / revealed / forgot the capital of Italy.

(2) $[\![\textit{the capital of Italy}]\!] = \lambda w.\ \iota x_e\ [\text{capital-of-Italy}(x,w)]$

(3) $[\![\textit{know}_{CQ}]\!](x_{<s,e>})(z)(w) = 1$ iff $\forall w'' \in \text{Dox}_z(w)\ [\ x(w'') = x(w)\ ]$

(4) $\textit{Know}_{CQ}$ + INTENSION of the NP:
$[\![\textit{Mary knows the capital of Italy}]\!] =$
$\lambda w.\ \forall w' \in \text{Dox}_m(w)\ [\ \iota x_e[\text{capital-of-Italy}(x,w')] = \iota x_e[\text{capital-of-Italy}(x,w)]\ ]$

However, the individual concept approach encounters problems when we consider concealed question NPs with quantifiers: (5). Combining the generalized quantifier's intension with the verb does not yield the correct truth conditions (Nathan 2005, Frana to appear). This has lead researchers to deviate from the core individual concept approach in different ways (Schwagger 2007, Roelofsen and Aloni 2008, Frana to appear).

(5) a. Mary knows / guessed / revealed / forgot **most** European capitals.
b. Mary knows / guessed / revealed / forgot **few / some** European capitals

The present paper proposes a solution to this problem within the invididual concept line. The key idea is that, in the same way that adverbials like *to some extent* and *for the most part* quantify over subquestions of an embedded question (Berman 1991, Lahiri 2002, Beck and Sharvit 2002), *some* and *most* can quantify over sub-individual concepts of a concealed question, as sketched in (6). Furthermore, it will be shown that certain constraints on determiner and adverbial quantification over concealed questions are parallel to those on determiner and adverbial quantification over (plain) plural individuals.

(6) The waiter knows / remembers [$_{CQ}$ **some / most** dishes you ordered].
≈
The waiter **to some extent / for the most part** knows / remembers [$_{InterrCP}$ what dishes you ordered].

BIBLIOGRAPHY

Aloni, M. 2008. Concealed questions under cover. In Franck Lihoreau (ed.), *Knowledge and Questions. Grazer Philosophische Studien*, 77, pp. 191--216
Beck, S. and Y. Sharvit. 2002. Pluralities of questions, *J. of Semantics* 19.
Berman, S. 1991. *On the semantics and Logical Form of Wh-clauses*, UMass PhD diss.
Frana, I. (to appear). Concealed questions and de re attitude ascriptions, UMass PhD diss.
Heim, I.: 1979, 'Concealed Questions', in R. Bäuerle, U. Egli and A. von Stechow (eds.), *Semantics from different points of view*, Springer, Berlin, pp. 51-60.
Lahiri, U. 2002. *Questions and answers in embedded contexts*, Oxford Univ. Press.
Nathan, Lance. 2005. *On the interpretation of concealed questions*, Doctoral Dissertation, MIT.
Roelofsen, F. and M. Aloni. 2008. Perspectives on concealed questions, *Proceedings of SALT XVIII*.
Romero, M. 2005. Concealed questions and specificational subjects, *L&P* 28.5.
Schwager, M. 2007. Keeping prices low: an answer to a concealed question, *Proceedings of Sinn und Bedeutung XII*.