Exploratory and confirmatory methods of investigation of causative constructions in a multilingual parallel corpus

1 Theoretical background

This study investigates the role of iconicity and frequency in the use of causative constructions. Using a multilingual parallel corpus of film subtitles, we compare the division of labour between causative constructions in eight European languages from the Germanic and Romance groups. More specifically, we focus on the factors that determine the use of lexical and analytic causatives, which are illustrated by (1a) and (1b), respectively:

(1) a. *The sheriff killed Bob*.
b. *The sheriff caused Bob to die*.

It has been suggested that the division of labour can be explained by the iconicity principle: lexical causatives convey more 'direct' causation with a stronger conceptual integration of the cause and effect, whereas analytic causatives convey indirect causation (e.g. Haiman 1985; Dixon 2000). However, according to some other proposals (e.g. Haspelmath 2008), more frequently occurring events have more basic, unmarked forms, whereas rare events are encoded with the help of more complex forms. This study puts the competing hypotheses to test by integrating the conceptual and usage factors in a multifactorial model based on data from film subtitles.

2 Data and method

The data come from a self-compiled corpus of film subtitles, aligned at the sentence level with the help of Tiedemann's Uplug software (Tiedemann 2003). We demonstrate that translationese and spatiotemporal restrictions in subtitling have a smaller impact on the use of causative constructions than one might think. The data set will contain 300 multilingual exemplars: analytic causatives and a sample of lexical causatives extracted from all originals and translations. The exemplars are coded for the type of causative constructions (analytic, lexical or other) and a number of conceptual and usage variables, such as the semantic classes of the main arguments and the relative frequencies of the construction in the corpus in comparison with its more (less) direct synonyms, which are identified on the basis of their distributional similarity with the help of semantic vector spaces. The frequencies are extracted from large comparable monolingual reference corpora.

The study employs both exploratory and confirmatory statistical methods. The exploratory analyses involve Multidimensional Scaling on a matrix of Hamming distances between the exemplars. The result is a probabilistic semantic map (Wälchli 2010), which helps us pinpoint the most important conceptual dimensions of causative constructions. In addition, Kriging, an additive polynomial model applied in geostatistical studies and spatial analysis, is employed to compare the prototypes of analytical and lexical causatives in different languages (Cysouw & Forker 2009). Figure 1 shows the results of Kriging for a subset of causative exemplars in four Germanic languages. The confirmatory analysis is based on a mixed-effect logistic regression model (Baayen 2008) with the semantic and usage variables as fixed effects, the exemplars and languages as random effects, and the type of causative as the response.

3 Preliminary results

Both the semantic maps and the mixed-effect model reveal a complex interplay of the conceptual and usage-related factors. In accordance with the iconicity hypothesis, interpersonal causation, which is usually indirect, has indeed a higher probability of analytic constructions than other types of causation. However, the frequency is by far the strongest predictor in the model. We also observe substantial cross-linguistic differences in the prototypes of analytic causatives, which affects the division of labour between the two constructional types.

References

Baayen, R. H. (2008) *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Cysouw, M. & D. Forker (2009) Reconstruction of Morphosyntactic Function: Non-spatial Usage of Spatial Case Marking in Tsezic. *Language*, 85 (3): 588-617.

Dixon, R. M. W. (2000) A typology of causatives: form, syntax and meaning. In R. M. W. Dixon and A. Aikhenvald, eds., *Changing valency: Case studies in transitivity*, 30–83.

Haiman, J. (1985) *Natural Syntax: Iconicity and Erosion*. Cambridge University Press, Cambridge.

Haspelmath, M. (2008) Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, 19(1): 1–33.

Tiedemann, J. (2003) Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing. Ph.D. Thesis, Uppsala University, Uppsala, Sweden.

Wälchli, B. (2010) Similarity semantics and building probabilistic semantic maps from parallel texts. *Linguistic Discovery*, 8(1): 331–371.

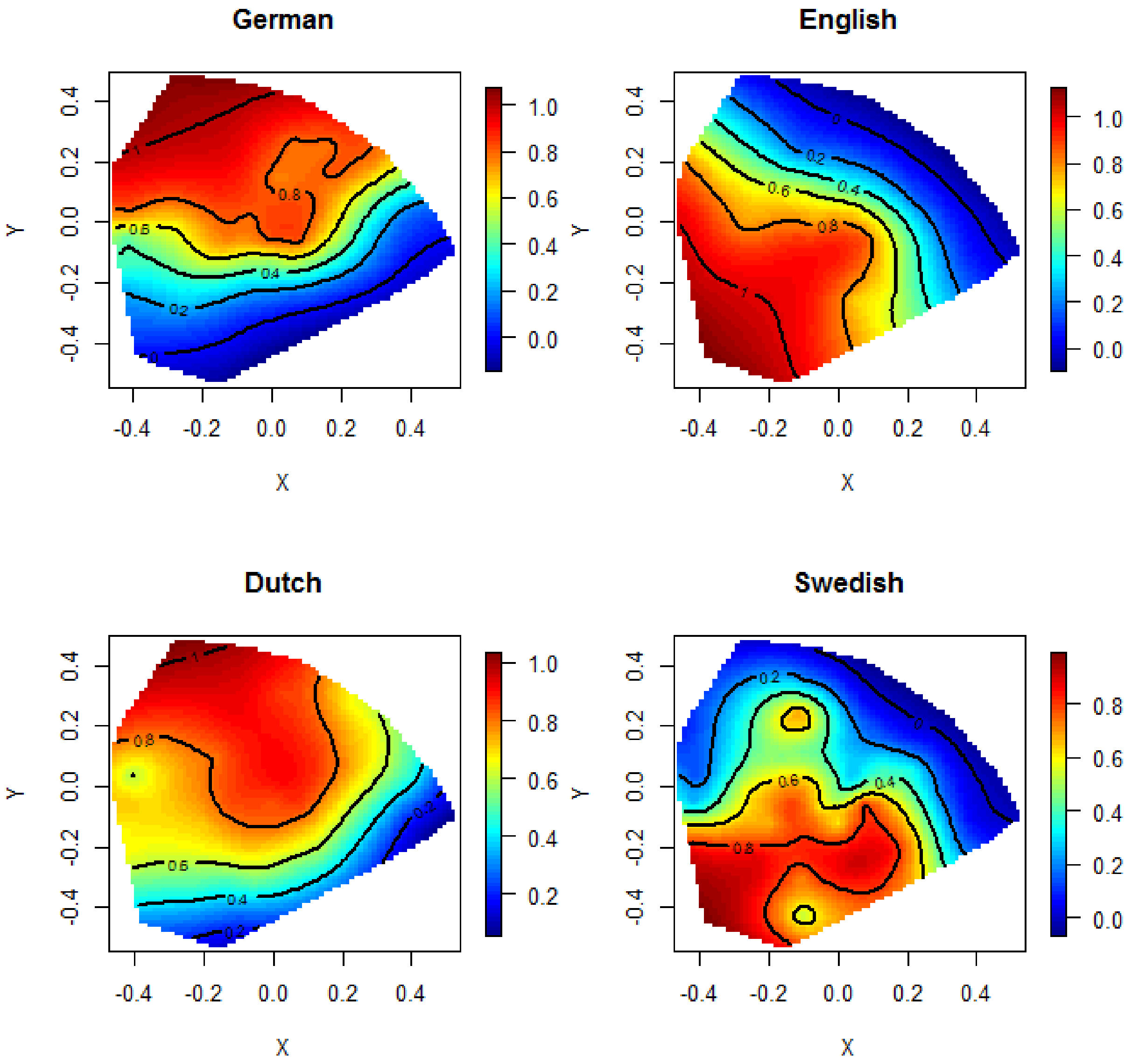

Figure 1. Kriging plots of four Germanic languages.The warmer colours indicate a higher density of analytic causatives (subtype: letting). The axes *x* and *y* represent two dimensions of MDS, which are interpreted conceptually in the paper.