From Annotation Mining to Linguistic Generalisations: The Grammar of preposition-noun combinations in German

Tibor Kiss

Sequences of prepositions and determinerless nominal projections (as e.g. *by bus, in jail)* have recently raised some attention, particularly in HPSG (cf. Baldwin et al. 2006, Trawinski 2003, Trawinski et al. 2006). Preposition-noun combinations show a variety of problematic properties, in particular that the nominal complement misses a determiner. This has led to calling the construction ‚determinerless PP', although it should be clear that nouns require a determiner, and not prepositions. The problem is particularly evident with singular count nouns. As has been pointed out by Himmelmann (1998), languages showing a determiner system as well as distinctions between singular count nouns, mass nouns and plurals, typical require a singular count noun be combined with a determiner. This universal tendency has also been implemented in the DUDEN-Grammatik for German. The following examples show a systematic violation of this condition not only with unadorned preposition-noun sequences, but also with more complex combinations of preposition and nominal projection.

(1) auf Anfrage *(after being asked),* auf Aufforderung *(on request),* durch Beobachtung *(through observation),* in Anspielung *(in allusion),* mit Vorbehalt *(with reservations),* ohne Probe *(without test),* ohne Vorwarnung *(without warning),* unter Androhung *(under threat)*

(2) auf parlamentarische Anfrage, auf diskrete Aufforderung, durch kritische Beobachtung, in untertreibender Anspielung, mit leisem Vorbehalt, ohne positive Probe, ohne mündliche Vorwarnung, unter sanfter Androhung

Dropping the determiner is not mandatory and constructions with and without determiner can be interchanged and lead to almost identical interpretations.

(3) Möglich geworden war das aggressive Vorgehen nur deshalb, weil Monica Lewinsky sich *unter der Androhung* einer langjährigen Zuchtshausstrafe (wegen Meineids) zur Ausplünderung ihrer Person bereit erklärt hat.
‚They could proceed in an aggressive manner because Monica Lewinsky accepted her self-exploitation under threat of a lasting imprisonment (for committing perjury).'

Baldwin et al. (2006) have suggested – for English and without reference to the count/mass-distinction – that preposition-noun combinations should be handled by lexical selection of a determinerless nominal projection. They admit that this specification in itself it too coarse-grained and requires further qualifications, possibly including semantic conditions. We take this assumption as a starting point and suggest identifying such qualifications not by standard linguistic methodology but by annotation mining.

Standard linguistic methodology, particularly judgements relying on introspection, assumes that speakers of a language are able to judge the construction in question and to point out criterial properties (albeit not being able to explain the workings of such properties). In preposition-noun combinations, neither condition is satisfied. Preposition-noun combinations are clearly productive and they do not show more instances of idiomaticity as other constructions (cf. Dömges et all. 2007, Kiss 2007). Yet, speakers of German are unable to judge the grammaticality of such constructions in isolation or to explain their meaning. The same holds for the criterial property of countability. Unless speakers are given explicit tests for countability (and – this being almost impossible for naive speakers – unless speakers are forced to evaluate such tests on large data sets), they are not able to state whether a noun in question should be qualified as count or mass. In a sense, this corresponds to findings since Allan (1980), where eight fine classes for mass and count nouns are suggested. A further challenge for clear-cut distinctions comes from polysemy and homonymy: for many nouns, different senses are related to count/mass-distinctions. Consider German *Kontrolle,* which in its abstract sense can be translated as *control,* but also shows an interpretation that can be translated as *passport control* or *reception control.* The latter can be classified as a count noun, the former as a mass term. Similar problems have also been addressed in Bond (2005).

Annotation mining makes explicit use of annotations on a variety of levels (parts-of-speech, morphology, chunking, parsing, semantic and ontological distinctions) to feed annotated corpora into a variety of classifiers, with the goal of grouping large sets of data into (machine-)identifiable clusters which can be related to linguistic properties and generalizations, and eventually can be turned into features and values of a HPSG analysis of these constructions.