Building a Corpus of Indefinite Uses Annotated with Fine-grained Semantic Functions

Maria Aloni, Andreas van Cranenburgh, Raquel Fernández, Marta Sznajder

Institute for Logic, Language & Computation
University of Amsterdam

Abstract content

1. Introduction

Natural languages possess a wealth of indefinite forms. English, for example, has at least four different indefinite determiners: *a, some, any, one*. Italian has many more including *uno, nessuno, (un) qualche, (uno) qualsiasi/qualunque, qualsivoglia*. These various forms seem to have a common logical/semantic core (their main function is to express indefinite reference), but typically differ in distribution and interpretation. For instance, there are contexts where English *some* and *any* can be interchanged (e.g. conditionals: *If you hear something/anything, call me*), while in other contexts using one or the other leads to different interpretations or to ungrammatical sentences (e.g. direct negations: *I didn't meet someone/anyone*; permissions: *You may kiss someone/anyone*; and episodic sentences: *I kissed someone/#anyone*). Italian determiner *qualsiasi* behaves like *any* in permissions and episodic sentences, but unlike *any*, it is ungrammatical in direct negations. German *irgendein* exemplifies yet another distribution/meaning pattern, resembling *any* in permissions, but being closer to *some* in episodic sentences.

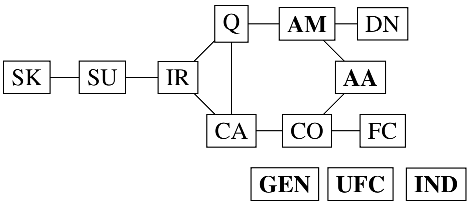
Many theoretical questions arise from these observations. For instance, why is there so much cross-linguistic and language-internal variation in indefinite forms? What exactly is the common core of these various forms and what is specific to each of them? Why did some typological patterns emerge rather than others? As a starting point towards a principled answer to these questions, our group has conducted a number of synchronic and diachronic corpus studies of various indefinite forms cross-linguistically (Aguilar-Guevara et al., 2011). The main goal of this research is to understand and compare the meaning and distribution of these forms and to develop some hypotheses on their historical development. The point of departure for the identification of the relevant categories was the typological survey by Haspelmath (1997), who identified 9 main functions (context/meaning) for indefinite forms and organised them in an implicational semantic map (see Figure 1). Haspelmath proposes that an indefinite will always express a set of functions that are contiguous on the map (i.e. that form a connected sub-graph). One of the aims of our previous work (Aguilar-Guevara et al., 2011) was to test this hypothesis. We extended the original map with a more detailed classification of negative and free choice uses of indefinites and developed a set of explicit logico-semantic tests organised in a binary decision tree, that would allow us to systematically assign particular functions on the map to instances of indefinites in context. We then conducted several pilot corpus studies where indefinite forms in Dutch, German, Spanish, Italian, and Czech where annotated by one annotator per language using the decision tree. These preliminary corpus studies confirmed Haspelmath's hypothesis of function contiguity.

The present paper makes the following contributions: (i) we extend previous work by developing annotation guidelines for the logico-semantic tests that can be used by non-expert annotators; (ii) we conduct a small corpus study on English indefinites *some* and *any* and report results of inter-annotator agreement; (iii) we make the resulting annotated corpus available to the research community through a searchable online database.

2. Data and Procedure

The aim of the methodology proposed by Aguilar-Guevara et al. (2011) was to come up with a systematic way of labelling uses of indefinites with one of the functions in the extended Haspelmath map in Figure 1. However, the logico-semantic tests that constitute the nodes of the decision tree were hardly usable by non-expert annotators. In order to evaluate this methodology with multiple non-expert coders, we first set ourselves the task of defining annotation guidelines that would elucidate the tests. The guidelines include a description of each function in the map, the decision tree proposed by Aguilar-Guevara and colleagues, and a description of the test to be applied at each non-terminal node of the tree, including an intuitive description of how to apply the test and some examples. The guidelines are available at `http://staff.science.uva.nl/~maloni/Indefinites/corpus.html`

We constructed our annotation dataset by extracting 100 instances of indefinites from the British National Corpus through the BYU-BNC web interface (Davies, 2004). Of these, 80 items were instances of *any* and 20 of *some*. All examples were independently annotated by 5 annotators, all of them graduate students at the ILLC. Only one of them was a native English speaker; the remaining 4 were proficient English speakers whose native language was Dutch (2 annotators), Russian (1) or Polish (1). The annotation scheme consisted of the functions in the extended Haspelmath map plus an additional label UN for "unclear", intended for cases where a test in the decision tree was inconclusive. Annotators were provided with the guidelines and received a few sessions of training where the guidelines were discussed. The annotation was done through an online interface that showed each indefinite to be annotated in context (100 tokens left and right, respectively).



	Abbr	**Label**	**Example**
a.	SK	specific known	*Somebody* called. Guess who?
b.	SU	specific unknown	I heard *something*, but I couldn't tell what it was.
c.	IR	irrealis	You must try *somewhere* else.
d.	Q	question	Did *anybody* tell you anything about it?
e.	CA	conditional antecedent	If you see *anybody*, tell me immediately.
f.	CO	comparative	John is taller than *anybody* else.
g.	DN	direct negation	John didn't see *anybody*.
h.	**AM**	anti-morphic	I don't think that *anybody* knows the answer.
i.	**AA**	anti-additive	The bank avoided taking *any* decision.
j.	FC	free choice	You may kiss *any* woman.
k.	**UFC**	universal free choice	John kissed *any* woman with red hair.
l.	**GEN**	generic	*Any* dog has four legs.
m.	**IND**	indiscriminative	I don't want to sleep with just *anybody* anymore.

Figure 1: An extended version of Haspelmath's map (new functions in boldface) and a short description of the functions.

3. Results and Analyses

Although the distributions of functions assigned to *some* and *any* differed across annotators, none of them violated Haspelmath's hypothesis of functional contiguity. Figure 2 shows the distribution of functions assigned to the 80 instances of *any* by each annotator.

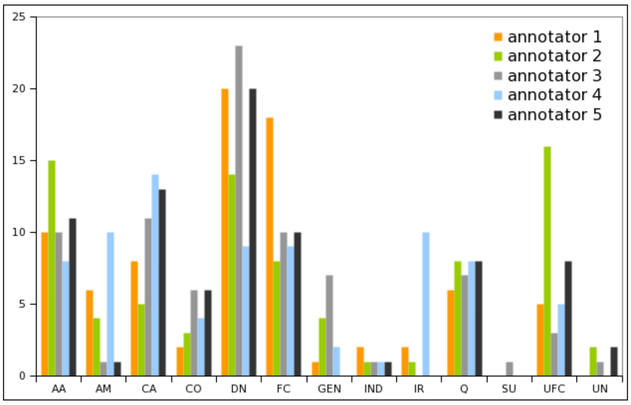


Figure 2: Distribution of functions for *any*

We compared the annotations using *kappa*. The overall *kappa* score was 0.52, with a standard deviation of 0.07 (*alpha* scores were equivalent up to the second decimal). Inter-annotator agreement was therefore moderate, which was not surprising given the fine-graininess of the annotation scheme. An analysis of the pair-wise confusion matrices showed that for *some* there were frequent disagreements between SU, SK, and IR. The confusion between SU and SK is to be expected in English, since there isn't a grammaticalised distinction between these two functions in this language. When these two functions are collapsed into one, we obtain a *kappa* score of 0.56 (with 0.07 standard deviation). As for IR, ambiguity seems to have played an important role. Some uses of *some* are often ambiguous between IR and SU/SK. Ambiguity, however, was difficult to detect by the annotators, which sometimes led to situations where annotators had focused on different readings.

Regarding *any*, most of the disagreements concerned the fine-grained functions that had been added to the original map proposed by Haspelmath: in the extended version of the map the classification of negative uses had been made more precise by adding AM and AA in place of Haspelmath's IN (indirect negation), while FC had been complemented by UFC, GEN, and IND. If we collapse these two groups of functions and thus consider the original Haspelmath's map, inter-annotator agreement increases substantially, with a *kappa* score of 0.62 (and a standard deviation of 0.05). The tests developed to distinguish AA from AM, and FC from UFC/GEN rely on intuitions about entailments of embedded disjunctive sentences (see the annotation guidelines). Reasoning tasks involving disjunction are known to be cognitively hard. Furthermore there is a lot of cross-linguistic variation with respect to the possibility of embedding disjunction. This might explain why these newly introduced distinctions led to disagreements between our annotators, who had a variety of native languages.

4. The Corpus

The work presented in this paper is part of our ongoing effort to create a cross-linguistic corpus of indefinite uses annotated with fine-grained functions as identified by formal semanticists. We expect the corpus to be a valuable resource for conducting cross-linguistic and typological studies of the different form/function mappings exhibited by lexical items used to express indefinite reference. For now, we make available the English corpus described in this paper together with the multi-coder annotation. The corpus is accessible through an online interface that allows users to browse the corpus restricting several parameters, including document genre; to search for items annotated with particular functions (by one or more annotators); and to download the dataset and/or the annotations. A beta version of the online interface is available from: `http://staff.science.uva.nl/~maloni/Indefinites/corpus.html`.

5. References

A. Aguilar-Guevara, M. Aloni, A. Port, R. Šimík, M. de Vos, and H. Zeijlstra. 2011. Semantics and pragmatics of indefinites: methodology for a synchronic and diachronic corpus study. In *Proceedings of the DGfS Workshop "Beyond Semantics: corpus-based investigations of pragmatic and discourse phenomena"*. BLA.

M. Davies. 2004. BYU-BNC: The British National Corpus. Available at `http://corpus.byu.edu/bnc/`.

M. Haspelmath. 1997. *Indefinite Pronouns*. OUP.