There is no end to it
Subjects in non-canonical constructions

Lea Cyrus, Münster

1 Introduction

The linguistic annotation of natural language corpora is one of the main areas of computational linguistics. Much energy has been devoted to building large syntactically annotated corpora, which are also called treebanks after the phrase-structure trees they contain. For some time now, functional information, i.e. information on whether a constituent functions as e.g. subject, object or adverbial, has also been included in the annotation. Yet, while at first glance this may not seem to be a venture too complicated, matters are not always as easy as they seem.

In this paper, we will take a look at the annotation of the subject in certain non-canonical constructions, namely sentences with extraposed subjects, existential sentences and cleft constructions. We will investigate how these constructions are analysed in theory, and also how the task of assigning the proper functions to each of their constituents is tackled in two syntactically and functionally annotated English language corpora, namely the Penn Treebank (see Marcus et al. 1993, 1994) and the SUSANNE corpus (see Sampson 1995).

2 Extraposition

The first and apparently the most straightforward construction we will investigate is commonly referred to as extraposition. The structure is taken to be the non-canonical equivalent of a sentence with a clausal or infinitival subject (1-a), in which the subject is shifted from clause initial to clause final position. The now empty pre-verbal subject slot is filled with the anticipatory or dummy pronoun it, which carries formal grammatical subject properties but is void of meaning (1-b).

(1) a. To choose a higher-salaried job is not unethical.
    b. It is not unethical to choose a higher-salaried job. [PENN]

Contrary to what the terminology suggests, the non-extraposed basic construction is by no means the more frequent of the two. While this seems to have been common knowledge for some time (see Quirk et al. 1985, p. 1392, Ernmann 1987, p. 41, and HUDDELESTON and Pullum 2002, p. 1404), Biber et al. (1999) have substantiated this assumption with statistical corpus findings: that-clauses in subject position occur only 10–20 times per million words in written registers (academic prose and news), and hardly

---

1 The example sentences for the Penn Treebank [PENN] have been taken from Bies et al. 1995a, those for the susannee corpus [SUSANNE] from Release 5, which is freely available via anonymous ftp at ftp://ftp.cogs.susx.ac.uk/pub/users/geoffs/SUSANNE.tgz. In both cases, sentences have been abbreviated where appropriate. The sentences have been extracted with the tool Tyrep2 (Rohde 2002). As this tool was originally developed for the Penn Treebank, the susannee corpus has been converted to a compatible format. However, the brackets and labels have remained unchanged.

ever in spoken communication, while the extraposed variant occurs 200–500 times per million words, depending on the register (p. 676). The same goes for to-infinitives: being comparatively rare in sentence-initial position (< 50 times per million words), they occur ten times as much in extraposition (p. 724). Biber et al. ascribe this preference for the sentence-final position to processing difficulties when the clausal subject, which can become very complex, has to be analysed prior to the matrix predicate (p. 677 and 725).

The main issue from the functional perspective is which of the two constituents, anticipatory it or extraposed infinitive/clause, is to be analysed as subject of the matrix clause. It is a widely held opinion that constructions with extraposition have two subjects, the anticipatory it as grammatical or dummy subject and the extraposed constituent as logical, postponed or notional subject (see e.g. Quirk et al. 1985, p. 1391). This solution worthy of a Solomon is supported by the annotation in the SUSANNE corpus (2).

In the functional annotation scheme of the SUSANNE corpus, there are three subject tags: :S for the surface subject, :s for the logical and :a for the passivised subject (Sampson 1995, p. 362). It has to be noted that “functional annotation” in the SUSANNE corpus is really what Sampson calls logical grammar (chapter 5). It follows from this that the logical subject is the default, while the surface subject is only explicitly annotated when these two functions do not coincide, as is the case in sentences with extraposed subjects. Hence, in the example above, the anticipatory it is tagged :S and the extraposed to-infinitive to think about the unthinkable is marked :s, logical subject.

While, as mentioned before, this functional analysis of sentences with extraposition is fairly common, this view is nevertheless disputed by some grammarians. Huddleston and Pullum (2002) take the view that it is only the dummy it that functions as subject. They do acknowledge that the extraposed elements correspond semantically to the subject in the canonical sentence, but state clearly that “they are not to be interpreted as kinds of subject. The subject is a syntactic function, and these elements are no more subjects than a former president is a president” (p. 243).

Huddleston and Pullum reach this conclusion by testing both the dummy it and the extraposed clause for their prototypical subject properties such as, among others, position and agreement, with the outcome that it is the dummy it alone that behaves syntactically like a subject. Seppänen et al. (1995) argue along the same lines, and are equally explicit as far as the status of the extraposed element is concerned: “we must