Contents lists available at SciVerse ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

Circularity effects in corpus studies – why annotations sometimes go round in circles

Manfred Consten [a,*], Annegret Loll [b]

[a] *Institute of Germanic Linguistics, University of Jena, D-07737 Jena, Germany*
[b] *Cologne, Germany*

ARTICLE INFO

Article history:
Available online 28 May 2012

Keywords:
Anaphor
Annotation
Circularity
Corpus study
Definiteness
Reference

ABSTRACT

Linguistic corpus research mainly deals with annotated data rather than raw data. This contribution investigates the status of annotated corpus data in empirical linguistics.

We argue that annotators should be regarded as co-producers of data; annotations depend on certain theoretical categories, hence they are theory-laden. Annotation categories differ with respect to different (structural and functional) levels of description and different degrees of canonisation, e.g. annotating a corpus item as a *noun* at a structural level is a highly canonised decision in most cases whereas the allocation of a cognitive-functional annotation category like *expression with identifyable referent* is subject to specific theories that often lack established definitions. As a minimal requirement, annotated data have to allow the reconstruction of the original raw data and annotations should be constrained by guidelines in order to avoid that the annotator's decisions are arbitrary.

Annotation problems resulting from the close relation between annotation categories and their theoretical prerequisites are exemplified using a newspaper corpus study and a study on a second-language acquisition corpus, both studies dealing with anaphora as a discourse-functional phenomenon.

It is shown that the problems discussed have their origins in two circles: the first one results from the interplay of deductive and inductive procedures that causes an impact of theory on annotation; the second circle originates from the relations between language structures and their discourse functions, the latter failing to be observable independently from the structural features of the utterance.

1. Introduction

In empirical linguistics, corpus studies are a widespread method employed not only for the exploration of phenomena, but also for the testing of hypotheses. The annotation of raw data (i.e. the assigning of theoretically acquired categories to raw data segments)[1] is an essential step in the process of conducting corpus studies. This contribution will show by example some problems arising in the course of this process.

* Corresponding author.
E-mail addresses: manfred.consten@uni-jena.de (M. Consten), lolla@smail.uni-koeln.de (A. Loll).

[1] Raw data are data which have not undergone transformations or manipulations by researchers. Strictly speaking, oral speech is represented as real raw data only in original audio/video tapes; transcriptions of oral speech are no raw data in a strict sense since they are data which have been transformed from an auditive mode into a written mode. However, *raw data* is a relative term. It depends on the object of investigation which data are considered as raw data. Transcriptions must not cause a loss of information relevant for the investigation, and they must not add any information. Only on that condition, transcriptions can be considered as raw data input for further annotation steps at which additional information will be generated. These requirements are met in the study discussed in Section 4.2.

0388-0001/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.langsci.2012.04.010

The article is organised as follows: in Section 1 we will postulate the interaction of both induction and deduction in empirical research. Section 2 characterises properties of data: it is shown that empirical research mainly deals with annotated data rather than raw data, a process in which annotators should be regarded as quasi co-producers of data.

Annotations are located at various levels that differ with respect to the theoretical complexity and different degrees of canonisation of the annotation categories and of the theories which these categories are based on. Annotated data have to fulfil the postulate of reconstructability (i.e., raw data must be reconstructable from annotated data) in order to guarantee transparent annotations.

In Section 4, two corpus studies are introduced: both show methodical problems, mainly with respect to the annotation of some doubtful data.

These problems are discussed more closely in Section 5.1 and revealed as problems of circularity. Here, two circles are discovered: the first circle has its origin in the interplay of deductive and inductive procedures (especially the impact of theory on annotation); the second circle has its origin in the characteristic of functional theories (i.e. theories reflecting the language function in human communication) which aim to specify relations between certain structures and certain functions: in some cases, functional categories (in the sense of *categories of natural language use*) cannot be specified and annotated independently from the structural features of the utterance.

In Section 5.2, characteristics of functional categories and their annotation are highlighted. Strategies based on plausibility often turn out to be the only way to avoid arbitrary annotations, and provide for reconstructability as claimed in Section 3.

Sadly, these considerations fail to resolve the circles described in Section 5.1.

2. The place of empirical data in linguistics: induction and deduction

In the recent history of linguistics, there has been one important topic in the debate between the opposing communities of Generative Grammarians and corpus linguists: the question of whether introspective data as they are preferred by generativists are empirical data at all (Kertész and Rákosi, 2008, p. 27).

We do not intend to pursue this topic, which has already been discussed extensively (cf. Lehmann, 2004; Geeraerts, 2006, as well as the contributions in Sternefeld, 2007, and Stefanowitsch and Gries, 2007).

If (*pure*) *empirical data* is defined as “data based on pure observation without any theoretical impact”, it will be of greater interest whether purely empirical data – apart from raw data – actually exist in linguistics and what problems arise in empirical linguistics from a questionable state of data. Certain problems of data annotation in corpus studies do not seem to be accounted for by the exponents of purely empirical data analyses.

Geeraerts (2006) may serve as an example of such a position. At first, he reports the viewpoint of those who claim that purely empirical research is impossible:

“Isn’t any attempt to reduce the role of introspection and intuition in linguistic research contrary to the spirit of Cognitive Linguistics, which stresses the semantic aspects of the language – and the meaning of linguistic expressions is the least tangible of linguistic phenomena. Because meanings do not present themselves directly in the corpus data, will introspection not always be used in any cognitive analysis of language?” (Geeraerts, 2006, p. 42)

It has to be added that the present contribution does not only deal with meaning but also with referential structures, which contain mental objects, exceed the levels of language structure and, thus, are even less tangible than semantic structures.

Geeraerts, then, turns against the sceptic position and claims that introspective and intuitive elements are part of theory formation only, and not part of data work.

> “Empirical research does not imply that intuition and interpretation have no role to play in linguistic research: rather, it implies that interpretation is but one step in the empirical cycle of successful research.” (Geeraerts, 2006, p. 45)

Probably this view is due to the common assumption that empirical work consists in either inductive or deductive use of data: “In the context of empirical scientific research, a datum serves either as the basis for the inductive construction of a hypothesis or as the test for a theorem arrived at deductively.” (Lehmann, 2004, pp. 180–181)

With this notion, however, it remains unclear how theorems which are tested deductively arise. Some sort of induction is required anyway – at least induction is provided by utterances linguists happen to pick up in every day life.

Therefore, we prefer a different notion of empirical work, assuming that data are used at inductive as well as deductive stages in a cyclic way.[2]

A sketch of this notion is shown as Fig. 1.

An empirical cycle consists of a process of the following-up of inductive and deductive steps[3]:

[2] For a differentiation of *cycle* versus *circle* see Rákosi (this issue), and Kertesz/Rakosi (2009, pp. 718–719). Following Rescher (1976), they define a cyclic process as a fruitful process that will lead back to a starting point (e.g. a theory) but at a higher level of insight (here: the initial theory is now confirmed and possibly refined by data). A circle, on the other hand, is a kind of fallacious argumentation. We will discuss circular argumentation in 4.

[3] Similar to the debate about the status of introspective data, the question of whether a strict deductive or a strict inductive approach is adequate for linguistics was the subject of a debate between generative and corpus linguists: Chomsky generally rejects inductive work whereas it is considered indispensable by corpus linguists (for a closer discussion see Kertész/Rákosi, 2008, pp. 26–27). This debate, however, seems to be quite ridiculous, since all linguists are capable of at least one language that will *always* inductively influence their theory formation.