Information Enriched Constituents in Dialogue

Abstract

Information enriched constituents are a linguistic phenomenon typically studied under headings such as ellipsis, fragments, and short or non-sentential utterances. The term *information enriched constituent* is introduced in this thesis to emphasise that the default in spontaneous spoken dialogue is to produce only that part of a message that contributes to the context, and that remaining material in the message is not omitted, but rather already part of the context.

This thesis takes a generation perspective of information enriched constituents, and investigates how utterances can be characterised and represented with regard to information enrichment, and under what circumstances information enrichment can be relied on to different degrees. While it is recognised that information enriched constituents concern a number of different linguistic levels, this thesis is primarily concerned with pragmatic aspects. Information enriched constituents are analysed using theories of information structure, with the *focus* as that part of an utterance that is to update the context, and *ground* as that part which is a reflection of what is already in the context. A distinction is also made between contrastive and non-contrastive foci, as these differ in their relation to the context. An extensive corpus study is provided, based on recorded dialogues in six different corpora from a number of different activities, and in the three languages English, French, and Swedish. The corpus study shows that for the generation of utterances relying on information enrichment, the pragmatic focus has to be able to be smaller than the syntactic focus phrase, and that ground material that is realised in dialogue is identical, anaphoric, or reformulated in relation to the relevant part of the preceding context.

The use of corpus dialogues for the study of information structure with respect to information enrichment, also provides a context that needs to be modelled, and the study identifies a number of different contextual components that can enrich subsequent utterances, notably questions under discussion (QUD), questions no longer under discussion and answers to such questions, and domain and situational knowledge. The corpus investigation also reveals a number of constraints on the form and content of information enriched constituents, primarily in terms of the determination of information structure, and different reasons for the inclusion of ground material in dialogue.

Utterances are formalised using Head-driven Phrase Structure Grammar (HPSG), with the inclusion of the representation of information structure. The context is formalised using a QUD-based information state. Constraints on the form and content of the utterance are explored using Optimality Theory (OT), which, in particular, reveals two interesting areas of conflicting constraints: the determination of the ground in relation to different contextual components, and the determination of how much ground material is to be included in the realisation, if indeed any. The OT analysis also shows the need for bidirectionality – taking both hearer and speaker perspectives into account – for the generation of information enriched constituents. The thesis also includes a specification of possibilities for the implementation of information enriched constituents in the generation component of a dialogue system.

KEY WORDS: information enriched constituents, dialogue, information structure, dialogue corpora, dialogue context, information state, questions under discussion, natural language generation, dialogue systems

The thesis is written in English.