SPRIK Conference
Explicit and Implicit Information in Text - Information Structure across Languages
University of Oslo, June 8-11 2006

Investigating nominal coreference in originals and translations

Kerstin Kunz

Saarland University

k.kunz@mx.uni-saarland.de

The present paper deals with the elaboration of a method for analysing various types of nominal coreference in originals and translations. We focus on designing a model for the manual annotation of coreferential noun phrases to capture differences on the lexicogrammatical level as well as on a more semantic and conceptual level. Note that, in this paper, we concentrate on the analysis of German and English texts only.

(1) A relation of coherence and cohesion

Coreference is a strategy of text creation and text reception linking pieces of text in a coherent way. It goes beyond the text surface as it is not only a relation between linguistic elements but also a relation of meaning - a cognitive connection between mental concepts. The recipient of a text establishes this connection not only on the basis of linguistic expressions in the text, but also by inferring knowledge stored in cognitive systems other than the language system, e.g. world knowledge. The relation of coreference consists in identity of reference as the same cognitive concept is activated and reactivated. However, this mental relation is evoked by linguistic expressions and cohesive devices which create coreference on the text surface in a more or less explicit way. Nominal coreference therefore constitutes both, a textexternal and a textinternal relation:

- textexternal with two or more noun phrases in a text referring to the same cognitive or extralinguistic concept
- textinternal with a directional pointer being set by a cohesive marker from one coreferring noun phrase to another

(2) Types of nominal coreference

Consider the following example of succeeding noun phrases as the may appear in a text:

A blue Porsche .... The car ..... it.... The Porsche

In the example, an extralinguistic concept or referent is activated for the first time by the noun phrase *A blue Porsche* - the antecedent. The same referent is reactivated by several other noun phrases (*'The car, it, The Porsche*) in the succeeding text – the anaphors. Personal pronouns, such as *it* can be stated as a typical and very common cohesive device to create identity of reference. However, as the preceding example displays, the system of a language provides other cohesive devices like hyperonymy, repetition, etc. The occurrence and use of these various types in specific texts may not only be instantiations of the language system but may also depend on the accessibility of the respective referents: different types of nominal coreference may indicate differences in information status (e.g. discourse old or discourse new) and attentional state (e.g. topic or focus). Referent accessibility not only may affect form and semantics of the coreferring

noun phrases but also their position and function in the sentences as well as in the text. In addition, the realization of reference identity via nominal coreference may also be a question of register.

(3) Nominal coreference in translation

Creating identity of reference constitutes a fundamental strategy to establish meaning in a given text as it provides thematic continuity. We consider the creation of coreference relations to be an essential part of the translation process as they help to create connectivity between textual elements in original and translation, and at the same time, retain the meaning of an original in its translation. Relations such as identity of reference on the mental level, which, in the original are expressed by cohesive devices in the source text language have to be preserved in its translation expressing them by adequate cohesive devices in the target text language. More precisely, the translator has to:

- identify noun phrases as referring expressions in the source text and identify cohesive devices indicating relations between these noun phrases
- identify reference markers as well as position and function of coreferring noun phrases signalling the accessibility of referents
- interprete the relation between the referring expressions to be identity, e.g. establish a mental relation of identity of cognitive concepts
- indicate referent accessibility by form, function and position of coreferring noun phrases in the target text
- transform the relation of sense into wording, e.g. express identity of reference by coreferring noun phrases in the target text language to create thematic continuity

(4) Analysing coreference in translations and originals

In order to analyse nominal coreference in originals and translations with corpuslinguistic methods we have to concentrate on how the same cognitive relation of identity is realized by noun phrases and cohesive markers within these noun phrases in the original and translation, respectively. Furthermore, we have to investigate if and how referent accessibility is expressed differently on the surface of source and target text. The sentences below, extracted from a German original and its English translation serve as an example. As the example displays, in the German source text the coreference relation is expressed by several noun phrases, which form a coreferential chain (underlined noun phrases). In the English translation, a coreferential chain is set up for the same extralinguistic concept but differs in number of anaphoric noun phrases and realization via lexicogrammatical elements within these noun phrases. In few cases, the coreferring noun phrases in original and translation also differ in their syntactic function and position. The example below can hardly reflect the network of different coreferential chains being set up in a complete text, nor does it display all possible varieties in form, semantics, position and syntactic function of coreferring noun phrases in source text and target text language.

Die Sojabohne zählt zu den bekanntesten Nutzpflanzen der Erde. Kaum eine andere Pflanze ist so reich an Wirkstoffen wie sie. Die Pflanze gehört als Hülsenfrucht zur Familie der Schmetterlingsblütler und stammt ursprünglich aus dem nordöstlichen Asien. In China und Japan haben Anbau und Verarbeitung der Sojabohne seit Jahrtausenden Tradition. Sie wird dort als eines von fünf "heiligen Getreiden" gepriesen und ist die Basis für eine Vielzahl unterschiedlicher Lebensmittel. Soja genießt in Asien in etwa den gleichen Stellenwert wie Fleisch und Kuhmilch in Europa.	*The soybean is one of the most well-known agricultural products on earth with more beneficial components than virtually any other plant. Soybeans are legumes belonging to the pea family and originally came from northeast Asia, where they have traditionally been cultivated and processed for thousands of years in countries like China and Japan. Here they are praised as one of the five "holy grains" and serve as the basis for a variety of different foods. In Asia, soybeans enjoy approximately the same status as meat and dairy products in Europe.*

(4) Model for an annotation scheme

In order to track down as many differences and commonalities in nominal coreference as possible our annotation scheme, which is designed for the manual annotation of corpora, is very fine grained.

We set up the following categories and subcategories in order to differ types of anaphora:

- **recurrence**: lexical and orthographical identity between anaphor and antecedent (total or partial recurrence)
- **pronominal relations**: the anaphor is a personal, possessive**,** demonstrative or relative pronoun
- **is-a relations**: the anaphor constitutes a synonym, a hyperonym, or a hyponym of its antecedent

With the following category, we intend to differ items indicating the type of reference

- **reference marker:** definite, indefinite and zero article, pronoun, possessive determiner, demonstrative determiner

We analyse the noun phrase by establishing the following categories:

- **modification:** to differ pre, post and zero modification
- **noun phrase head**: to indicate the level of embedding of a subordinate anaphoric noun phrase
- **number:** to differ plural and singular of the noun phrase
- **noun phrase type:** to differ noun phrases referring to generic or specific referents
- The category **function** looks into the syntactic function of coreferring noun phrase differing subject, different types of objects, complement adverbial, predicator, etc.

- The categories **positionGER** and **positionEN** analyse the position of coreferring noun phrases in the sentence, considering differences in German and English constituent structure.

Apart from these classifications, we number all coreferring noun phrases and assign each coreferring noun phrase to its respective coreferential chain in the text. Finally, we align a coreferring noun phrase in the translation to the corresponding coreferring noun phrase in the original.

(5) Conclusions and outlook

The classifications described above serve as a scheme for the manual annotation of nominal coreference e.g. with an XML editor. They help us to trace and compare coreferential chains referring to the same concept in originals and translations. For instance, the number of noun phrases in a coreferential chain may vary in original and translation. In addition, we may detect differences resulting from various modifications on the lexicogrammatical level e.g. different cohesive markers, different semantics of the head noun, different types of modification in the coreferential noun phrases, or differences in position and syntactic function of the coreferring noun phrases. We may also capture changes due to the more or less explicit marking of coreference on the text surface. These shifts in nominal coreference on the text level can cause shifts of coreference on the semantic and cognitive level: They may affect referent accessibility or cause the creation of new or different coreferential chains. Eventually, this may result in changes in the coherence and thematic continuity of the translation in comparison to its original.

References

Blum-Kulka, S. 1986. Shifts of Cohesion and Coherence in Translation. In: J. House & S. Blum-Kulka (eds.), *Interlingual and Intercultural Communication*. Tübingen: Narr.17-35

Gundel, J., N. Hedberg & R. Zacharski. 1993**.** Cognitive Status and the Form of Referring Expressions in Discourse. *Language*, 69/2: 274-307

Halliday, M.A.K. & R. Hasan. 1976. Cohesion in English. London: Longman

House, J. 1997. Translation Quality Assessment: A Model Revisited. Tübingen: Narr.

Martin, J.R. 1992. English Text: System and Structure. Amsterdam: Benjamins

Steiner, E. 2004. Translated Texts: Properties, Variants, Evaluations. Frankfurt/M. etc.: Lang

Strube, M. & U. Hahn. 1999. Functional Centering: Grounding Referential Coherence in Information Structure. *Computational Linguistics* 25/3: 309-344

Grosz, B. J. , A. K. Joshi & S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21: 203-225.