Interrelating Treebanks with Language-Specific Descriptions of Information Structure

Daniel Storbeck Sanghee Kwon Felix Sasaki
Andreas Witt

Bielefeld University
Department of Linguistics and Literary Studies

1 Subject of the paper

The motivation for this article is to describe a methodology for interrelating and analyzing language and theory-specific corpus data from various languages. As an example phenomeon we use information structure (IS, see [3]) in treebanks from three languages: Spanish, Korean and Japanese. Korean and Japanese are typologically close, while both are typologically different from Spanish. Therefore, the problem of annotating IS is that there are diverging language-specific formal linguistic means for the realization of IS-functions (like "topicalization / contrast") on various levels like prosody, morphology and word-order. Hence, it is necessary to describe the relations between language-specific formal means and functional views on IS, and how to operationalize these relations for corpus analysis.

2 The methodology

There are projects which deal with a corpus-based approach to IS, e.g. [2]. Nevertheless, these projects mainly deal with european languages, and — different to our approach — they don't rely on a dedicated methodology how to interrelate various levels of IS-function and formal means. Our methodology (see fig. 1) relies on (A) annotated corpora, which make use of a specific annotation format, and (B) a "conceptual level", which serves as a formal description of IS and which encompasses a set of language-specific or language-general "conceptual models".

(A) Usually, a single document grammar is used for the creation and validation of a corpus. This imposes restrictions on the possibilities of structuring the annotations and on their extensibility. [6] outlines a solution to this problem that

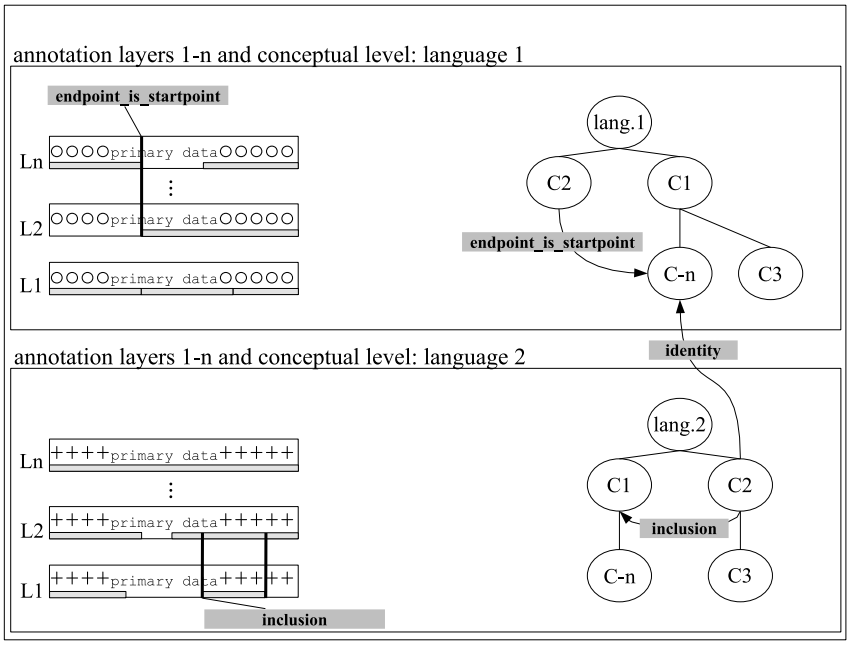

Figure 1: Overview of the methodology

allows for "multiple annotation hierarchies". Key features of this annotation format, which is visualized in the left part of fig. 1, are that it is XML-based, that modelling of alternative annotations based on different theoretical assumptions is possible, that each annotation layer can be created separately, and that new layers can be added at any time. Each annotation layer is represented in a separate XML-file and has its own document grammar. For each language, the primary data serve as an implicit link between multiple, separate annotations. Interrelations between the annotations are declared separately, i.e. on the conceptual level (see below and the right part of fig. 1), in order not to affect the annotation.

The annotations are transformed into Prolog facts. The Prolog fact base can be queried with a dedicated query language. The query language makes use of a closed set of predicates which define "positional relations" between annotations on separate layers. Such relations like "endpoint_is_startpoint"[1], "endpoint identity" or "inclusion"[2], can be derived from [1]. They are applied to textual data in the corpus by referring to character positions in the string, e.g. two annotations have the relation "identity" if they span the same range of characters.

[1] See for example the annotations on layer n and layer 2 of language 1 in fig. 1.

[2] See for example the annotations on layer 1 and layer 2 of language 2 in fig. 1.

(B) A key characteristic of the methodology is that these positional relations can be hypothetically declared at the conceptual level and / or heuristically derived from the corpus data, cf. [4]. The relations can be applied in three ways. First, as described above, they allow for the interrelation of annotations on different layers. Second, as can be seen in the right part of fig. 1, the relations can be used to describe "interconceptual relations" in a conceptual model for a language. For example the relation "endpoint_is_startpoint" between some annotations for language 1 can be interpreted as a relation between the concepts C2 and C-n, which are part of a conceptual model for that language. And third, the relations serve as interconceptual relations between different conceptual models, i.e. for different languages. For example in fig. 1, the concept C2 from the model for language 2 and the concept C-n from the model for language 1 are related via the relation "identity".

[5] illustrate the application of this methodology to japanese data, which are annotated according to several heterogeneous, theory-specific models of linguistic phenomena. The description of relations between such heterogene annotation units is done within the conceptual level. In the remainder of this paper, after describing the properties of IS in the three languages in question (section 3), several use cases for the methodology in the area of IS will be exemplified:

1. to describe relations between linguistic forms and IS-functions within one language, making use of annotations on several layers (see example 1 in section 4.1);
2. to interrelate language-specific, corpus-based descriptions of IS-functions on the conceptual level (see example 2 in section 4.2); and
3. to explicate typological differences of two given languages on the conceptual level (see example 3 in section 4.2).

3 Language-specific characteristics of IS

In terms of morphosyntax, it is usually said that Korean and Japanese are typologically closely related. Grammatical functions are normally morphologically marked by particles or verbal suffixes which contain their own meaning features. Referring to morphological elements such as the topic particles neun/eun (kor.) and wa (jap.) and case particles i/ga/(l)eul (kor.) and ga/o (jap.), categories like definiteness, genericity, topic/focus and contrast in a sentence can be described. The basic korean and japanese word oder is SOV but it is relatively flexible, i.e. scrambling is permitted. Spanish has a mixed morphology using analytic as well as synthetic formation principles. This is especially true of the verbal system with person, number

and tense inflection but Spanish also makes extensive use of analytic tense and periphrastic aspect formation. The basic constituent order is considered SVO. While Japanese and Korean have postpositions and are of the "complement–verb" and "modifier–modified" order types, Spanish has the reverse order types and prepositions. In the following section we focus on the means each language uses to realize the topic-focus-articulation and the related phenomenon of contrast. Spanish relies primarily on constituent order and intonation to realize these phenomena, while Japanese and Korean additionally have explicit morphological means for this purpose.

4 Application of the methodology to IS

In order to illustrate what can be done with our methodology, we will present possible cases for intra- and interlingual comparisons of the specific and general sets of IS categories.

4.1 Description of intralingual relations

Case 1: In the first case we propose to interrelate annotations of language-general, IS-functional categories with annotations of language-specific, formal categories. This can be done describing the positional relations between annotations with respect to the primary textual data.

```
   Example 1: A korean sentence with possible annotation layers.
   Transl.: ``The child drinks _tea_.'' (And not someting else.)
   Morph. : child-nom  tee-topic  drink-present-deklarative

1:Primary data: ai------ga cha-neun masi--n----da
2:Sentence    : ---------------S-----------------
3:Word        : ----W----- ----W--- ------W------
4:Particles   :        nom     top-
5:IS-function :            topical./contrast
```

In this example there are the primary textual data (visualized as layer 1) and four annotation layers. Layers 2 and 3 express sentence and word segmentation. Layer 4 contains the annotations of language-specific morphemes while layer 5 contains annotations of IS-functions. Bearing in mind that there are not enough data available to derive a stable language-specific definition of IS-functions, e.g. "topicalization / contrast", this annotation permits a hypothetical and partial definition, though, as follows: Topicalization / contrast is positionally included in a

stretch of text representing a sentence and it has a positional identity relationship with a word unit which has a positional inclusion and endpoint identity relationship with a topic marker. Furthermore the word unit has neither starting point nor endpoint identity with the including sentence unit. Since IS-functions can be realized via several combinations of language-specific formal means, their resulting definitions as derived from the annotation are actually sets of definitions, each one of them describing a pattern of a characteristic configuration of language-specific means.

The advantage of the approach of multiple annotation is that such definitions can be established and refined during every phase of elaboration of the corpus. More layers can be added when more annotations are available or when diverse, new theoretical viewpoints on the same primary data are adopted. The annotations can be flexibly extended because they are not tied to a single, immutable document grammar.

4.2 Description of interlingual relations

The following two cases illustrate how an interrelation of IS across languages is possible. Different to the example in the last section, these cases mainly rely on the conceptual level.

Case 2: For structurally similar languages like Japanese and Korean it can be useful to query treebanks of the two languages using only one of the specific sets of categories. In this case a treebank of Japanese could be transparently queried with categories of a specific model for Korean. A common problem in defining correspondences between the two sets of categories is that one of them has more finegrained definitions of some categories than the other. Suppose, a given model for Korean annotates two types of topicalization constructions: 1. topicalization with object fronting (T1) and 2. topicalization including a topic morpheme (T2), see example 2. In a given model for Japanese only a general category "topicalization" (T) might be annotated. A plain identification of the topicalization types in the following example leads to a loss of information.

Example 2:

```
*     Korean
1:Prim. dat.: cha---reul ai----ga masi--n----da
2:Sentence  : ---------------S-----------------
3:Word      : ----W----- ----W--- ------W------
4:IS-catego.: ----T1----                        (object fronting)
5:Particles :       akk-      nom    pres.decl.
```

```
*   Japanese
1:Prim. dat.: cha-----wa kodomo--ga nomu
2:Sentence  : -------------S------------
3:Word      : -----W---- ----W----- -W--
4:IS-catego.: -----T----
                                        *Added layer:
5:Particles :        top        nom
                                        *Reconstructed categories:
6:IS-catego.: ----T2----
```

To solve this problem, a distinction of the two types of topic like in the korean annotation is created in the japanese annotation. This is done creating an additional annotation layer describing particles. Then, using this morphological information, the definition of topicalization in the japanese model can be refined to distinguish the types T1 and T2. In the conceptual level, these definitions are mapped on corresponding general concepts. These concepts allow for the interrelation of language-specific categories across languages.

Case 3: The last example focusses on the conceptual level and illustrates how the methodology can be used to explicate typological differences between two languages. Suppose, Spanish and Korean are to be compared and for both languages the same language-general concepts have been defined. Then, these concepts can be redefined into several subordinate concepts, in order to make the typological differences explicit.

Example 3:

```
*     Korean
1:Prim. dat.: cha---reul ai----ga masi--n----da
2:Particles :       akk-      nom
3:Stress    : ---

1:Prim. dat.: cha---neun ai----ga masi--n----da
2:Particles :       top-      nom
3:Stress    :

*    Spanish                    (_tea_ drinks the child)
1:Prim. dat.: té bebe el niño
2:Stress    : --
3:Gram. role: O- -V-- ---S---
```

Continuing with an example of the IS-function concept “topicalization”, in Korean the subconcepts “topicalization with fronted and stressed object” as well

as “topicalization with object fronting plus topic marker” can be created. As for Spanish, only the first concept can be applied. In this way, general superordinate concepts are linked to the corresponding language-specific subordinate concepts, i.e. definitions for each language. Thus, this simple conceptual level of topicalization types expresses a typological difference between Korean and Spanish, i.e. that the korean concept “topicalization with object fronting plus topic marker” has no corresponding realization in Spanish.

References

[1] Allen, James F. and Ferguson, George (1994): “Actions and events in interval temporal logic”. Technical report 521. URL http://www.cs.rochester.edu/u/james/.

[2] Baumann, Stefan; Brinckmann, Caren; Hansen-Schirra, Silvia; Kruijff, Geert-Jan; Kruijff-Korbayová, Ivana; Neumann, Stella; Steiner, Erich; Teich, Elke and Uszkoreit, Hans (2004): “The MULI Project: Annotation & Analysis of Information Structure in German & English”. In: Proceedings of LREC 2004, Lisbon.

[3] Lambrecht, Knud (1994): “Information structure and sentence form. Topic, focus, and the mental representation of discourse referents”. Cambridge University Press.

[4] Sasaki, Felix (2004): “Secondary Information Structuring - A Methodology for the Vertical Interrelation of Information Resources”. In: Proceedings of Extreme Markup Languages 2004, Montreal.

[5] Sasaki, Felix; Witt, Andreas and Metzing, Dieter (2003): “Declarations of relations, differences and transformations between theory-specific treebanks: a new methodology”. In: Proceedings of Second Workshop on Treebanks and Linguistic Theories (TLT 2003), Växjö, Sweden.

[6] Witt, Andreas (2004): “Multiple hierarchies: new aspects of an old solution”. In: Proceedings of Extreme Markup Languages 2004, Montreal.