Published in: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002). Las Palmas, Gran Canaria. - European Language Resources Association (ELRA), 2002. pp. 1225-1230.

# Co-reference annotation and resources: A multilingual corpus of typologically diverse languages

**Felix Sasaki, Claudia Wegener, Andreas Witt,**
**Dieter Metzing and Jens Pönninghaus**

Universität Bielefeld
Fakultät für Linguistik und Literaturwissenschaft
- Computerlinguistik und Texttechnologie -
Postfach 10 01 31
D-33501 Bielefeld
{felix.sasaki, c.wegener, andreas.witt,dieter.metzing,jens.poenninghaus}@uni-bielefeld.de

**Abstract**

This article introduces a dialogue corpus containing data from two typologically different languages, Japanese and Kilivila. The corpus is annotated in accordance with language specific annotation schemes for co-referential and similar relations. The article describes the corpus data, the properties of language specific co-reference in the two languages and a methodology for its annotation. Examples from the corpus show how this methodology is used in the workflow of the annotation process.

## 1. Introduction

This paper provides details on research from the project "Secondary information structuring and comparative discourse analysis", a part of the research group "Text-technological information modelling" (cf. http://www.text-technology.de). In our project we focus on two typologically diverse languages, Kilivila (an austronesian language) and Japanese. The phenomenon under investigation is the expression of co-reference and related phenomena, which is important for various applications like information retrieval and extraction.

Co-reference is often used as a generic term for a wide range of phenomena (see Deemter and Kibble, 2000, for a thorough classification). Co-reference in a narrow sense means that two expressions refer independently to the same extra-linguistic entity, e.g. "Einstein" and "the founder of the theory of relativity". In contrast to this the reference of anaphoric expressions depends on other linguistic units, e.g.: "**The man$_i$** came down the street. **He$_i$** was smiling". There are several subtypes of anaphoric expressions like bridging anaphora. These are semantically indirectly related to the antecedent, e.g. "the door" following the previously introduced antecedent "house". Several kinds of possible relations exist, like part - whole, cause - event etc.

To maintain the quality of co-referential and similar annotations and to make the results of automatic resolutions of co-reference comparable, several annotation schemes have already been developed, e.g. MUC-7 (Chinchor and Hirschman, 1997), MATE (Poesio, 2000) or the schemes developed by Bruneseaux and Romary (1997), or Müller and Strube (2001). Due to the complexity of defining co-reference and separating it from related phenomena, it is difficult to maintain the quality of annotation schemes and the corpora relying on them. In their review of MUC-7, Deemter and Kibble (2000) address this problem. For example, quite often co-referential and anaphoric relations are not clearly distinguished. Also MUC-7 generalizes certain predications, although the semantic extension of the nouns is not co-referential. Furthermore it is difficult to choose potential linguistic units for the co-referential relations and to classify these relations.

In the field of computational linguistics, many annotation schemes or corpora concentrate mainly on English. If other languages are investigated, these are mostly other European languages, or Japanese. The same holds true for the investigation of co-referential phenomena in this field; non-European languages and the multilingual dimension of co-reference (see section 3 of this paper) are ignored. One goal of our project is to fill this gap by contributing resources, i.e. annotation schemes and annotated data. This might also help to improve the quality of co-referential annotation schemes in general.

In this paper, we will introduce the corpus (section 2) and basic properties of co-reference[1] in the two languages under investigation (section 3). Furthermore we will introduce our methodology of annotation in general (section 4.1.) and for the two languages (section 4.2.), making use of various XML-based tools we are developing.

## 2. Description of the corpus

Our data consists of task-oriented dialogues elicited with an interactive game, namely the Tinkertoy matching game.

The Tinkertoy matching game (Brown et al., 1993) was developed as one of four interactive games used by the Cognitive Anthropology Research Group of the Max Planck Institute for Psycholinguistics in Nijmegen in order to elicit comparable data in many different languages, especially on spatial reference. The setting of this game is as follows: Two speakers (one "director" and one "matcher") sit side by side, separated by a screen preventing visual contact. The director is given photographs of objects or the objects themselves built with Tinkertoy materials. She has to describe the object and the matcher has to build it. Verbal interaction of any kind is explicitly allowed, e.g. inquiries by the matcher. The advantage of using this setting is that data in a variety

[1] According to the common terminology, in the following discussion we will use the term "co-reference" as a generic term.

of languages has already been elicited. With this data it is easy to widen the scope of future research.

We decided to use task-oriented dialogues, because they provide relatively clear patterns of dialogue acts and their structure is rather simple (cf. Metzing and Kindt, 2001). Due to this simplicity compared with other types of dialogue it is easier to formalize the relations between several levels of description (morphemes, lexemes, phrases, utterances, speaker intentions, dialogue segments). This formalization can be used for a classification of dialogue acts. With this classification one can interfere candidates for co-referential and similar relations. Furthermore the number of possible candidates is limited by the more or less restricted discourse universe, which is due to the goal-oriented character of the dialogues.

Our corpus-data of Kilivila contains several dialogues, which were recorded in 1992 by Gunter Senft. The participants in these games are villagers on Kaile'una Island, one of the Trobriand Islands in Papua New Guinea. They range in age from approximately 25 to 55 years.

The Japanese dialogues were elicited in 2000 in Japan by Felix Sasaki. The participants in the dialogue were mainly Japanese students of Tokyo University.

All dialogues were audio recorded, the Kilivila dialogues were in addition video taped. The Kilivila data was transcribed by Gunter Senft, the Japanese data by Felix Sasaki and Mirei Maki.

The following table specifies the exact number of utterances and the number of words contained in the corpus:

| Language | # of utterances | # of words |
|---|---|---|
| Japanese | 2.125 | 15.267 |
| Kilivila | 1.141 | 6.057 |

In the corpus, annotations on several levels will take place. The basic level will be an annotation of morphemes. Other levels will include certain word classes, noun phrases, utterances etc. Certain semantic relations will be marked as well, to be able to differentiate the various aspects of co-reference. For the data structure of the corpus and the process of annotation see section 4.

## 3. Properties of co-reference in typologically diverse languages

### 3.1. Co-reference in Japanese

In Japanese, co-referential relations can be expressed with several means.

First, after a referent is introduced explicitly, the default way is not to realize subsequent referring expressions. This is the case of so-called 'zero-pronouns'[2] (cf. Kameyama, 1985). Unlike other languages (e.g. Italian, Finnish) there is no verbal inflection indicating syntactic arguments. Hence, one has to use other cues of information to resolve the antecedent of a zero-pronoun. The syntactic positions of the explicitly mentioned referent and the omitted argument of the verb in the following utterances are important clues. In addition the honorific marking gives information about potential referents especially if they are human or related to certain persons or groups. The honorific marking takes place on several levels like lexical items, certain morphemes and syntactic constructions, and on many syntactic positions. There are also agreement-like relations between several honorific markers, e.g. between subject and verb (cf. Siegel, 2000).

The second means of referring to nominal referents are overt pronouns. Different to many other languages, such explicit anaphora are not the default way of topic continuity. Often they indicate a focus- or thematic-shift. Hence, when considering the various applications for a corpus with co-reference and similar annotations, e.g. text summarization, pronouns should be classified separately.

The third means of referring to a nominal referent in Japanese are numeral classifiers. They always occur with a numeral, specifying certain semantic properties like spatial dimensions:[3]

*kiiroi no* ***ni*** ***hon no*** ***bou$_i$*** *wo*
yellow GEN two NC GEN stick ACC
'two yellow sticks …'
…
***i*** ***ppon me$_i$*** *ha*
one NC CN THEME
'one of (the sticks) …'

Example 1: Numeral classifiers in Japanese

The example above, taken from our Japanese data, contains parts of two utterances. In the first utterance the nominal entity "yellow stick" is introduced and specified as two units – two sticks. The numeral classifier *hon* implies the semantic properties 'long, round'. In the second utterance, the allomorph *ppon* of the same classifier is used with a different numeral to specify a quantitative subset: 'one of them'. In contrast to the (zero-) pronouns, the two numeral classifiers do not refer to the same 'world' or discourse entity, but to two different sets which are semantically interrelated – two sticks vs. one stick. Following Deemter and Kibble (2000), for a co-reference annotation scheme such a semantic difference compared to 'normal' co-reference has to be taken into account.

### 3.2. Co-reference in Kilivila

As in Japanese, anaphoric and referential relations can be expressed in Kilivila in a number of ways (cf. Senft, 1986).

The nominal referent can be omitted after having been introduced. If this is not the case, pronouns can be used to express subsequent anaphoric relations. However, this is quite uncommon. Comparable to Japanese, they have certain discourse functions, e.g. emphasis.

Much more often the referent is omitted. In this case the valence of the verb and verbal inflection facilitate the inference and unique identification of the corresponding referent.

[2] Because they are not realized at a certain position, zero-pronouns lead to the problem how to annotate them in a corpus.

[3] The following abbreviations are used: GEN: genitive particle; NC: numeral classifier; ACC: accusative; CN: counter noun. CP: classificatory particle; 2.: second person.

Furthermore a technique of nominal classification is used: Several word classes (including demonstrative pronouns, numerals and some adjectives) have to be marked with respect to the class of the noun they refer to. This is done by attaching or inserting 'classificatory particles' (CPs, cf. Senft, 1996). The corresponding CP is used to refer to the omitted referent, even beyond sentence boundaries, and sometimes several utterances after the introduction of this referent, thus securing coherence in discourse. This is comparable to the function of numeral classifiers in Japanese.

The following example, taken from our corpus, exemplifies the function of the CPs in establishing referential relations:

***ke*$_i$**-  *ta*  ***kai*$_i$**  *ku-*  *kau*
CP.wooden-  one  stick  2.-  take
'take one stick...'
...
***ke*$_i$**-  *bwabwau*
CP.wooden-  blue
'the blue (stick)...'

Example 2: CPs in Kilivila

In the first part the noun *kai* 'stick' is explicitly mentioned. The referent hereby introduced is taken up later in what is the second part of the example, but is now only referred to by a nominal phrase consisting of an adjective containing the CP *ke* 'wooden'. The noun is not mentioned again. This CP also occurs in the first part of the example, attached to a numeral modifying the noun *kai* 'stick'. The occurrence of this CP in both nominal phrases bridges the gap between them and creates a referential relation.

## 4. Annotation of co-reference in our project

The properties of Japanese and Kilivila introduced in the last section shows clearly that there are language specific configurations of certain linguistic expressions, referential relations and discourse functions. The facts to be considered can be summarized as follows:

(1) If pronouns or other referential expressions are not explicitly mentioned, language specific clues like honorific or syntactic marking have to be used to infer possible antecedents.

(2) There are linguistic expressions like numeral classifiers and classificatory particles which are important for co-reference, but which are not yet taken into account in language-independent annotation schemes.

(3) There are certain *quantitative* referential relations between discourse units. In the languages under investigation these are expressed by classifiers used with numerals.

(4) Overt pronouns, which are regarded as a common means of expressing referential relations in many annotation schemes, are an uncommon means in Japanese and Kilivila, related to certain discourse functions like focus-shift.

These points support our motivation to declare language specific annotation schemes for co-reference. We will now describe our methodology of how to create such schemes.

### 4.1. Process of annotation and creation of annotation schemes

To get a language-specific annotation scheme of co-reference, we use the methodology of multiple annotations which is described in detail in Witt (2002). At the beginning, the same data is marked up separately for each annotation unit (certain linguistic units, co-referential relations, discourse functions etc.). To enhance this process we wrote an a XML-based tool which can be used in an ordinary internet browser:


Figure 1: Multiple annotation of primary data

Figure 1 shows two separate annotations of the first utterance in example 1. One annotation is made for referential expressions in general. These units are called `refEx`. The other annotation is made for numeral classifiers, called `nc`.

In the next step, the separate annotations, each represented as a single XML-document, are unified. The unification algorithm is implemented in the Python programming language. When applied to the annotation in figure 1, the unified document looks like this:

```
<refEx><nc>nihon</nc>nobou</refEx>
```

This can be read as "the referential expression `refEx` *nihon no bou* 'two sticks' contains a numeral classifier `nc` *nihon* 'two round things'".

Each separate annotation can be viewed as an instance of a simple XML document grammar, for example DTD (document type definition). The simple document grammar declares the respective annotation unit, e.g. the declaration of one element in the corpus. The default structure, expressed in the format of a XML-DTD, is as follows[4]:

```
<!ELEMENT corpus(#PCDATA|elementname)+>
<!ELEMENT elementname (#PCDATA)>
```

All these 'atomic' document grammars are collected in a pool, without being interrelated:

[4] We are working with mixed-content, i.e. the content model is unrestricted with respect to the order of unparsed data and elements.

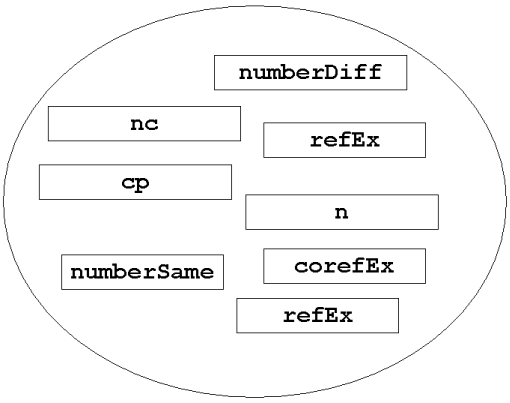


Figure 2: Pool of unrelated document grammars

In the final step for the creation of the annotation scheme, a more complex document grammar is written for the unified document. Regarding our example it contains not only the simple definitions of the referential expression and the numeral classifier, but also their interrelation:

```
<!ELEMENT corpus(#PCDATA|refEx)+>
<!ELEMENT refEx  (#PCDATA|nc)>
<!ELEMENT nc     (#PCDATA)>
```

This example for a complex document grammar declares an element `corpus` which may contain one or more elements `refEx`, or character data. The element `refEx` might contain an element `nc` or character data.

In the following section, we will describe the properties of preliminary versions of complex document grammars for co-reference in Kilivila and Japanese.

## 4.2. Annotation schemes for Japanese and Kilivila

As for the Japanese example (example 1), the following annotation units will be defined:

```
numeral classifier – nc
noun – n
referential expression – refEx
coreferential expression – corefEx[5]
the same number – numberSame
different number – numberDiff
```

Other annotation units could be defined as well, like for noun phrases, syntactic functions, honorifics etc. Nevertheless, to show the language-specific properties of the annotation scheme, this small set of annotation units will be sufficient.

The multiple annotation of the Japanese example is shown in figure 3:

[5] This unit will be used to annotate co-referential expressions in general, e.g. referential identity, anaphora etc.

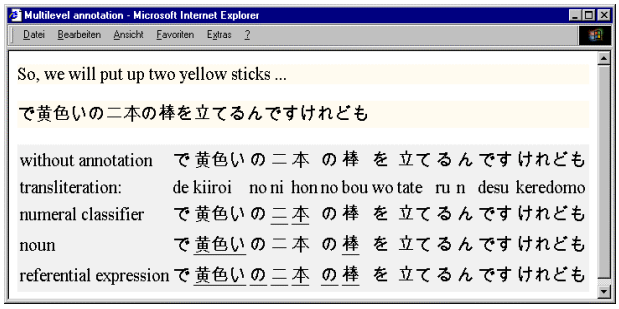


Figure 3: Annotation of Japanese data

Only the first of the two utterances in example 1 is shown here. The first line contains the English translation of the utterance, the second line the original transcription, followed by the primary data without annotations. A transliteration in a Latin script is supplied as well. Three separate annotations are added: for the unit 'noun' the annotation of *kiiroi* "yellow" and *bou* "stick", for the unit 'numeral classifier' the annotation of *ni hon*, and for the unit 'referential expression' the annotation *kiiroi no ni hon no bou.* The smallest segments in the data are morphemes which can be seen by the separated, underlined passages.

The unified version of the document will look like this:

```
<corpus><refEx>...<nc>nihon</nc>no
<n>bou</n></refEx>...
<corefEx>
<numberDiff><nc>ipponme</nc>
</numberDiff>ha</corefEx>
</corpus>
```

The following complex document grammar can be written for the unified document:

```
<!ELEMENT corpus   (#PCDATA|
                    refEx|corefEx)+>
<!ELEMENT refEx    (#PCDATA|n|nc)+>
<!ELEMENT corefEx(#PCDATA|numberDiff)+>
<!ELEMENT numberDiff (nc)>
<!ELEMENT n        (#PCDATA)>
<!ELEMENT nc       (#PCDATA)>
```

It is possible that the unification of the simple annotations leads to a document that cannot be validated with the complex document grammar defined above. In such a case the complex document grammar has to be reformulated. For example, the same number of units can be specified several times with a numeral classifier, like "two sticks … two sticks …". Therefore another element will be introduced, namely `numberSame`. This leads to a new declaration of the `coref` - Element:

```
<!ELEMENT corefEx(#PCDATA|
                  numberDiff|numberSame)+>
```

For the Kilivila example, the following annotation units will be defined:

```
classifier – cp
noun – n
referential expression – refEx
coreferential expression – corefEx
```

The unification of the separate annotations leads to the following document:

```
<corpus>
<refEx><cp>ke</cp>ta<n>kai</n>
</refEx>...
<corefEx><cp>ke</cp>bwabwau</corefEx>
</corpus>
```

For this document, the following complex document grammar can be written:

```
<!ELEMENT corpus
(#PCDATA|refEx|corefEx)+>
<!ELEMENTrefEx    (#PCDATA|n|cp)+>
<!ELEMENT corefEx (#PCDATA,cp)>
<!ELEMENT cp (#PCDATA)>
<!ELEMENT n (#PCDATA)>
```

The complex document grammars for the two languages match up to a certain level: The declarations of the element `corpus` are the same, and the declarations of the `refEx` element both contain the element `n`. The difference can be seen in the `refEx` element, which in the case of Kilivila might contain a `cp` element, and in the case of Japanese an `nc` element. Furthermore the `corefEx` for Japanese might contain a `numberDiff` element or a `numberSame` element.

We want to point out that there is a need to create document grammars for other languages or for more detailed descriptions of Japanese and Kilivila, focusing upon different domains or other phenomena. With the concept of a pool for document grammars introduced above (cf. figure 2), this task can be fulfilled while maintaining the comparability of annotation schemes which are based upon the same pool. The relations between specific configurations of document grammars and the unstructured pool is shown in figure 4.

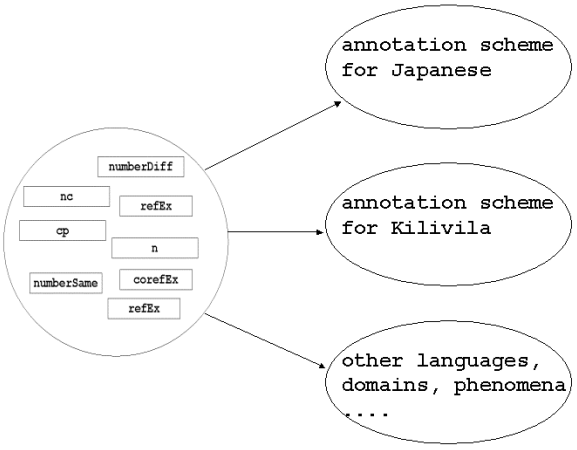


Figure 4: Several complex document grammars, based upon the same pool of unrelated document grammars.

## 5. Conclusion

We have described data of Japanese and Kilivila which was used for the annotation of co-reference and related phenomena. To improve the reliability of our annotation schemes for co-reference, we used multiple annotations in separate documents. These annotations were then unified, and a complex document grammar for each language was written.

Our methodology was demonstrated with a small set of data. Naturally, the language-specific annotation schemes are still too specific and have to be extended, to make use of more annotations and more simple document grammars. Also, the process of creating a complex document grammar, which is done manually at the current stage of our project, could be done semi-automatically. This will be a task in the continuing project.

## 6. Acknowledgements

We would like to thank Gunter Senft for providing his data and sharing his knowledge about Kilivila. The project "Secondary information structuring and comparative discourse analysis", part of the research group "Text-technological information modelling", is founded by the German research council (DFG). The collection of the Japanese data by Felix Sasaki was supported by the German Academic Exchange Service.

## 7. References

Brown, P., G. Senft, and L. Wheeldon, 1993. *Annual report 13, 1992*. Nijmegen: Max Planck Institute for Psycholinguistics.

Bruneseaux, F., and L. Romary, 1997. Codage des references et coreferences dans le dialogues homme-machine. In: *Proceedings of ACH-ALLC*, Kingston.

Chinchor, N., and L. Hirschman, 1997. MUC-7 coreference task definition, version 3.0. Available from the authors, chinchor@gso.saic.com

Deemter, K. v., and R. Kibble, 2000. On Coreferring: Coreference in MUC and related annotation schemes. *Computational linguistics*, 26 (4).

Kameyama, M., 1985. *Zero anaphora: the case of Japanese*. Ph.D. thesis, Stanford University.

Metzing, D., and W. Kindt, 2001. Strukturbezogene Methoden. In: K. Brinker, G. Antos, W. Heinemann, and S. F. Sager (eds.), *Linguistics of text and conversation*. Berlin, New York: Walter de Gruyter.

Müller, C., and M. Strube, 2001. Annotating anaphoric and bridging expressions with MMAX. In: Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue. Aalborg, Denmark.

Poesio, M., 2000. Coreference. In A. Mengel, L. Dybkjaer, J. M. Garrido, U. Heid, M. Klein, V. Pirrelli, M. Poesio, S. Quazza, A. Schiffrin, and C. Soria (eds.), *MATE (multilevel annotation tools engineering). Deliverable D2.1 – Dialogue annotation guideline*. http://www.ims.uni-stuttgart.de/projekte/mate/mdag/cr/cr_1.html

Senft, G., 1986. *Kilivila: the language of the Trobriand islanders*. Berlin: Mouton de Gruyter.

Senft, G., 1996. *Classificatory particles in Kilivila*. New York: Oxford University Press.

Siegel, M., 2000. Japanese honorification in an HPSG Framework. In: *Proceedings of the 14th Pacific Asia*

*Conference on Language, Information and Computation*.

Witt, A., 2002. Meaning and interpretation of concurrent markup. In: *ALLCACH2002, Joint Conference of the ALLC and ACH, Tübingen 2002.*