Acquiring Lexical Information from Multilevel Temporal Annotations

Thorsten Trippel, Felix Sasaki, Benjamin Hell, Dafydd Gibbon

Department of Linguistics and Literary Studies
Bielefeld University, Germany
{ttrippel,gibbon,ben}@spectrum.uni-bielefeld.de, felix.sasaki@uni-bielefeld.de

Abstract

The extraction of lexical information for machine readable lexica from multilevel annotations is addressed in this paper. Relations between these levels of annotation are used for subclassification of lexical entries. A method for relating annotation units is presented, based on a temporal calculus. Relating the annotation units manually is error-prone, time consuming and tends to be inconsistent, and a method is presented for automatically accomplishing this task, and evaluated using German, Japanese and Anyi (W. Africa) corpora.

1. Objective

The extraction of lexical information from multilevel annotations is the objective of this paper[1]. With existing annotations on multiple levels there are at least two ways of extracting information using relations between two or more layers:

- Relations between the different annotation levels are defined, either in general for the whole level or individually relating annotation units from more than one annotation level. The annotation can be processed according to these rules.
- Relations between the different annotation levels are unknown at least partially; the levels share at least the source (the annotated signal or text), and the relation is therefore given by the sequential relation via the source.

The paper focuses on the later variant of information gain from annotation levels with only partially known relations. This will be done by applying the technique described below to time related signal annotations and text annotations, which shows that the usability of the technique is neither restricted to lexicons for spoken language systems nor to texts. The approach is related to similar analyses in constraint based phonology (see [1]). It is based on the assumption that the annotation is available in standoff style (see [2]).

The coverage of a corpus-based lexicon does not extend the coverage of the corpus if no complex generating functions such as paradigm generators are involved. Each annotation level can serve as the basis for lexical information such as wordlists extracted from an orthographic transcription, word frequency lexicons, etc. But if two or more obviously correlating annotation levels are present (such as orthographic word level transcription and canonical phonemic word level transcription), these relations have to be defined or discovered first. The temporal calculus based approach provides a method of finding assumptions in this area.

[1]This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG) in the project *Texttechnologische Informationsmodellierung* (FOR 437)

2. Design

The acquisition of a lexicon is part of an integrated approach to corpus lexicography as illustrated by Fig. 1.

The lexicon is induced from a corpus by a lexicon acquisition function. The lexicon accesses the corpus by a concordancing function. Both relatively static units, lexicon and corpus can be used for multimodal output using appropriate technology.

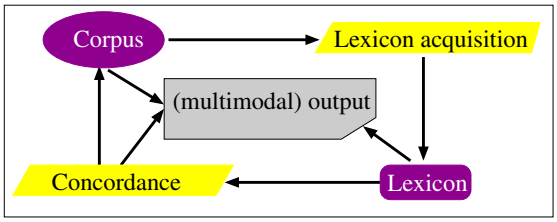

Figure 1: Integrated corpus based lexicon system

The annotations are available in linear order, either by the timeline referring to *absolute time* or by sequential enumeration of characters according to a timeline referring to *category time* (see [3], p. 68). Both are formally consistent with *Annotation Graphs* as defined in [4]. Fig. 2 illustrates the use of *category* and *absolute time* for the same utterance.

The relation of different annotation levels is not determined by the synchrony of annotated events. For example [5] mentions that there are time shifts between the annotated words and accompanying gestures. The ModeLex corpus for conversational gestures and other linguistic annotations[2] shows that hardly any gesture is synchronous with its lexical affiliate.

In the text domain the same holds true if a textual source is tagged according to different linguistic description levels, for example morphemic vs. syllabic level. In this special case the relation is regular and therefore could be inferred if the rules are known for a given language, but the segmentation is not the same on both levels, resulting in a non-contextfree structure if unified (see [6]).

Manually relating the annotated units is obviously not the preferred procedure as it is error-prone, time consuming and tends to be inconsistent. Consequently a different way of realising structured and consistent relations is needed. This is possible using a temporal calculus approach as illustrated in [7] (or a logic such as that of [8]). Fig. 3 illustrates the 13 temporal relations that can be applied to events on different annotation levels.

Using all possible temporal relations — excluding *before* and *after*[3] — produces a huge number of related events, which

[2]See http://www.spectrum.uni-bielefeld.de/modelex/

[3]The relations *before* and *after* result in every annotation unit being

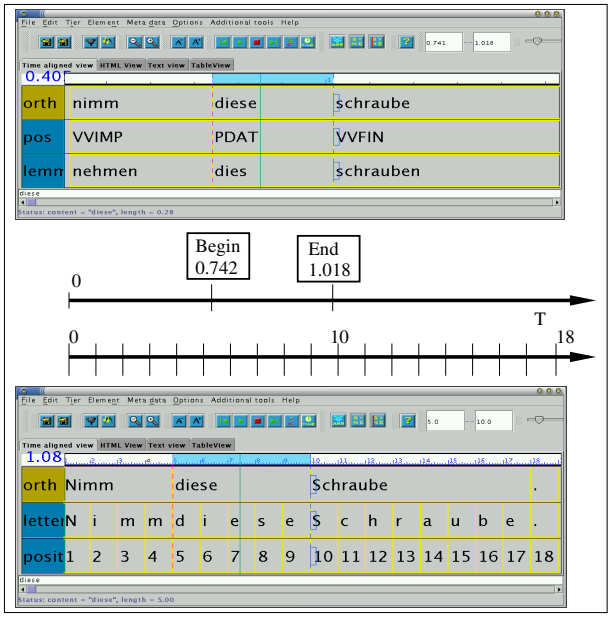

Figure 2: Relation of *absolute* and *category time* structure. From top: annotation of the utterance "Nimm diese Schraube" *Take this screw* according to the acoustic signal with orthography, POS, lemmatization, absolute timeline, category timeline, textual source, annotation in category time (enumeration of characters).

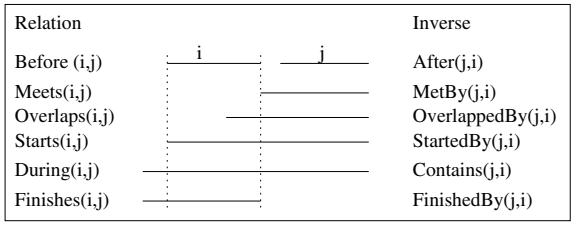

Figure 3: Selected relations of annotation units, equality not included, from [7].

can be estimated by

$$\sum_{la \in LA} |EV_{la}| * AVR * |LA|,$$

with LA being the set of layers of annotation, EV_{la} being the set of annotation events on layer la, $|...|$ indicating the cardinality of these sets, and AVR being the estimate average number of relations on every other layer. For initial estimation based on existing results an $AVR = 2$ has been assumed. However this estimation is far from being precise; for a small corpus with 4 annotation layers of 1461, 519, 98 and 79 events (annotated on word level, phrase level, gesture level and gesture function level) this results in the vast number of 17256 while the empirical results show "only" 8998 relations. However this is due to gaps in the annotation and uncertainty remaining from the variability of event length and number of events. With larger corpora the deviation does not remain constant, for example a 10 layer annotation with 3302 events is estimated to have 66040 relations while in fact there are "only" 39036 relations. Nevertheless the number is sufficiently large to stress the need for subclassification of possible relations.[4]

Three possible classifications for $ev_i \in EV$ and $la_j \in LA$ are:

1. *General* (implicational) relations that always exist: whenever there is a specific event ev on layer la_i, there is a corresponding specific event ev' on layer la_j. These are the most *general relations* and relatively easy to depict, for example that nouns are parts of noun phrases.
2. *Systematic* (tendential) relations that exist sufficiently often: if there is a specific event ev on layer la_i in more than a threshold number of cases there is a corresponding specific event ev' on layer la_j. If the threshold is sufficiently large, this could be an indicator for clustering a unit in different senses or at least to distinguish stylistic variation.
3. *Singular* relations that are unique in a given context: for one specific event ev on layer la_i, there is a corresponding specific event ev' on layer la_j. These cases give no particular information, as they might describe extremely rare examples as well as annotation inconsistency or indicate that the corpus is not sufficiently lexically saturated.

The lexicon includes the general and systematic relations. The systematic relations need further explanation and subcategorization. The singular relations must be dealt with separately.

3. The lexicon structure

To be able to access the lexicon in a distributed environment and to ensure reusability, we use the formalisms provided by the *Resource Description Framework* (RDF, see [10]). Fig. 4 visualises the generic structure of the lexicon.

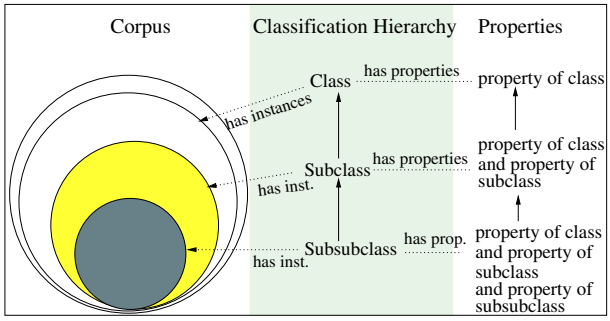

Figure 4: Basic lexicon structure

The lexicon is based on three kinds of inclusion relations:

1. the corpus with its subsets (left part of Fig. 4),
2. a hierarchical classification, depending on the linguistic categories of the annotation (middle part of Fig. 4), and
3. differentiating properties for classification (right part of Fig. 4).

At the top of the classification hierarchy, a particular set of annotation units on a given annotation level is chosen to be the object of classification, for example, all annotation units with

related to every other annotation unit (generalised Cartesian product) and are therefore omitted for practical work before evaluating specific hypotheses.

[4] This concept of classification has been applied to textual data with generic markup in [9].

the tag w on the word annotation level. The classification hierarchy is then built for this annotation unit. For each class in the hierarchical classification, a name is given, for example 'class of nouns' or 'class of pronouns' and so on. Each class has some properties, in our case some relations between annotation units. In Fig. 4, the connection between properties and the respective classes is expressed via the dotted lines between the two. Each subclass inherits the properties of its superordinate class. That is: a class1 has the property properties-class1, a class subclass-of-class1 has the properties properties-class1 and properties-of-subclass-of-class1 and so on. The inheritance relation is visualised via the arrows pointing upwards.

The connection from the classification hierarchy to the corpus (from the middle to the left in Fig. 4) is given by building subsets of instances of annotation units in the corpus, based on whether a certain property is given for a corpus instance of the annotation unit or not. Before the classification hierarchy is constructed, no property is given, so all instances of annotation units in the whole corpus are part of the set. As a specific annotation unit on a particular annotation level is chosen, all instances of this annotation unit are part of a subset of the corpus. In Fig. 4, this is visualised by the arrow from the highest class in the classification hierarchy to the outermost circle in the left. Next, for each instance of the annotation unit, the properties of the following subclass, i.e. structural relations between annotation units, are tested. If a relation exits, the instance of the annotation unit will be part of the respective subset, otherwise the instance of the annotation unit will be neglected. As there is an inheritance relation between the properties attached to the different classes, there is a subset relation between the instances of the annotation units in the corpus: the instances of a subclass are also instances of its superordinate class, and the instances of the subsubclass are also instances of the subclass. This subset relation is visualised with the Venn diagram in the left of Fig. 4. Every circle, i.e. every set, contains the instances of the circles inside.

The basic lexicon structure described above is abstract in respect to the 'meaning' of a certain class in a classification hierarchy. That is, a class — in the classification hierarchy — can be interpreted as constitutive for a class — in the lexicon — of lexemes (e.g. all nouns), a subclass of lexemes (e.g. all pronouns), a certain lexical entry (i.e. the pronoun 'this') or certain subtypes of the lexical entry. We do not commit ourselves to particular machine learning algorithms to construct classes and properties for the systematic (tendential) dependencies in a given corpus, but acknowledge that several such algorithms may be usefully deployed; this discussion does not fall within the scope of the present paper, however. In the following section we demonstrate how linguistic knowledge can be combined with (semi-)automatic analysis of the corpus data.

4. Sample lexicon entry

Figure 5 describes the lexicon entry of the Japanese demonstrative pronoun それ *sore* .

The object of the classification is the annotation unit w with the attribute $type = pronoun$, which is the top of the classification hierarchy, that is, all instances in the corpus of this annotation unit are part of the set to be classified. The first property to be tested is whether the pronouns are inside nominal phrases, which are represented as annotation units NP on the annotation level *phrases*. This property is grounded on basic linguistic knowledge and the starting point for creating new properties and further subclassifications. As this is true for all cases, this relation can be classified as a general relation as mentioned before. Subsumed under the pronoun class is a pronoun called *sore*. For this purpose, we construct a subclass Pronoun_Sore which owns the property $identity(Pronoun, \text{それ})$. In other words, the subset of instances in the corpus are all text strings それ, which are tagged as pronoun and which are part of an NP.

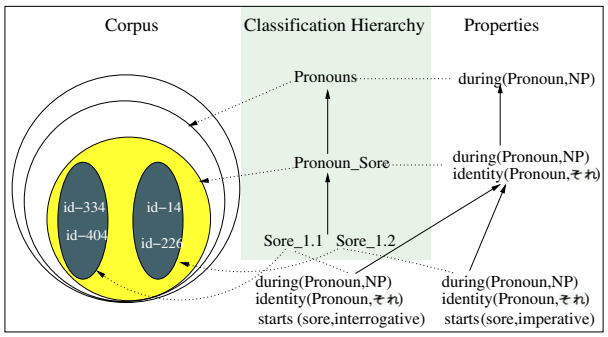

Figure 5: Example instance of a Japanese Lexicon

The next subclasses are constructed semi-automatically using the relation, *starts*, and looking for instances of annotation units on all given annotation levels which own this relation. As a result, we find two relevant instances on the level of sentence types, namely the interrogative sentence type and the imperative sentence type. These relations to the sentence type level is true for sufficiently many cases, so it is an indicator for clustering different *senses*. We use them as properties of two new subclassess $Sore\_1.1$ and $Sore\_1.2$, which also are connected to certain instances of annotation units in the corpus with unique identifiers, for example `id-334` as an instance of $Sore\_1.2$.

The position of a temporal relation within the hierarchical dependency structure (whether implicational or tendential) expresses the degree of universality of the relation, ranging from being a prototype for the whole lexeme to being a specific subclass of it.

5. Application and Evaluation

A tool to investigate the relations of corpora encoded in TASX format, described in [11], is implemented using XSLT, providing an algorithm of relating every annotated event to all other events on other annotation layers and classifying them according to the relations mentioned in Section 2. The result is output using XML syntax and is processed further for categorisation, especially for purposes of word sense clustering for relations that exist sufficiently often. The timestamps provided in the annotation layer are used to define the linearity and baseline for the annotations. In textual data, this can be achieved by enumeration of the characters, which yields category time as defined above.

Another implementation was described by [12] using Prolog, which enables further predicate logic inferences and restrictions, by which the available corpus can be checked for the existence of general rules (i.e. to see if there are other cases as well if a number of properties are available) and for the evaluation of existing rules against the corpus. This tool is especially useful for evaluating general relations (see Section 2).

The lexicon tools were initially tested with four different, relatively small corpora:

- Anyi-Story-Teller corpus: A recorded story by a story teller in the West African language Anyi recorded on

video, annotated in the audio modality, using orthographic transcription for the story teller and the responding audience, syllable segmentation, prosodic annotation and gloss; TASX format.
- German multimodal corpus: A recorded story by a German story teller annotated for audio and visual modalities, using an orthographic transcription, phonemic transcription, lemma assignment, and gesture transcription; TASX format.
- Japanese test utterance: annotated on 7 layers, including gloss, phrases and sentence type; TASX format.
- Japanese textual corpus: annotated on 32 layers, including part of speech, phrases, sentence types, coreferential relations; only partially available in TASX format, hence number of relations estimated and no classification numbers given in Table 1 for this corpus. This corpus has been tested with a Prolog tool for classification hypothesis testing.

Table 1 shows the relation of annotation layers, events, relations between events on the different layers and the number of classes neglecting singular relations. The number of classes does not include further subclassification.

Table 1: Corpora size and number of relations

Corpus	Annot. Layer	Annot. Events	Relations	Classes (without singular rel.)
Anyi Story Teller	10	3302	39036	6094
German Story Teller Multimodal	4	2162	8998	741
Japanese test utterance	7	58	770	41
Japanese construction dialogue texts	38	7468	283784	—

Evaluation was conducted based on principles laid out in [13] and [14]. Qualitative testing was started using a technical testing corpus that only contained two elements in all possible variations which was possible to relate by hand as well as by using the available tools. As the temporal relation method ideally requires a lexically saturated corpus, a hand-annotated test relation set cannot be generated due to the number of relations.

Generating all 39036 relations for the 10 layer, 3302 event Anyi corpus on an AMD Duron 1000 running Linux using the XSLT processor $xsltproc$ takes about 18 minutes, categorisation using the same technology less than 12 minutes. Initial tests indicate that the processing time is almost linear in the size of the corpus.

6. Summary and future work

In this paper we present a method of extracting lexical information from time-aligned multilevel annotations using a temporal calculus approach. A design for a lexicon classification system and a hierarchy extraction system was presented and a resulting lexicon structure was introduced. A sample lexicon entry with information based on temporal annotation relations was discussed. The lexicon acquisition tools were evaluated for their functionality and timing requirements as well as applied for sample corpora.

The next step in the development involves the integration of the temporal relations approach with a linguistic knowledge based approach.

7. References

[1] S. Bird, *Computational phonology: a constraint-based approach*, ser. Studies in natural language processing. Cambridge University Press, 1995.

[2] D. McKelvie and H. S. Thompson, "Hyperlink semantics for standoff markup of read-only documents," in *Proceedings of SGML Europe'97*, Barcelona, May 1997.

[3] J. Carson-Berndsen, *Time Map Phonology*, ser. Text, Speech and Language Technology. Dordrecht: Kluwer Academic Publishers, 1998.

[4] S. Bird and M. Liberman, "A formal framework for linguistic annotation," *Speech Communication*, no. 33 (1,2), pp. 23–60, 2001.

[5] P. Morrel-Samuels and R. M. Krauss, "Word familiarity predicts temporal asynchrony of hand gestures and speech," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1992.

[6] A. Witt, "Multiple Informationsstrukturierung mit Auszeichnungssprachen. XML-basierte Methoden und deren Nutzen für die Sprachtechnologie." Ph.D. dissertation, Universität Bielefeld, 2002.

[7] J. F. Allen and G. Ferguson, "Actions and events in interval temporal logic, Tech. Rep. TR521, 1994. [Online]. Available: http://www.cs.rochester.edu/u/james/

[8] J. F. A. K. van Benthem, *The logic of time*. Dordrecht: D. Reidel Publishing Company, 1983.

[9] F. Sasaki and J. Pönninghaus, "Testing structural properties in textual data: beyond document grammars," *Literary and Linguistic Computing*, vol. 18, forthcoming.

[10] O. Lassila and R. R. Swick, "Resource description framework (RDF) model and syntax specification," World Wide Web Consortium, Tech. Rep., 1999, http://www.w3.org/TR/REC-rdf-syntax/.

[11] J.-T. Milde and U. Gut, "The tasx-environment: an xml-based toolset for time aligned speech corpora," in *Proceedings of LREC 2002*, Las Palmas, 2002, pp. 1922—1927.

[12] D. Goecke, D. Naber, and A. Witt, "Query von Multiebenen-annotierten XML-Dokumenten mit Prolog," in *Proceedings of the GLDV-Frühjahrstagung 2003*, Köthen, Germany, 2003.

[13] D. Gibbon, R. Moore, and R. Winski, Eds., *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter, 1997, ch. SL system assessment.

[14] D. Gibbon, I. Mertins, and R. Moore, Eds., *Handbook of multimodal and spoken dialogue systems: resources, terminology, and product evaluation*. Dordrecht/New York: Kluwer Academic Publishers, 2000, ch. Terminology for spoken language systems.