Stefanie Dipper
Michael Götze
Stavros Skopeteas
(eds.)

Information Structure in Cross-Linguistic Corpora:

Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure

ISIS | Working Papers of the SFB 632 | 7 (2007)

ISIS issues do not appear according to strict schedule.

Vol. 7 (2007)

Volume Editors: Stefanie Dipper, Michael Götze, Stavros Skopeteas
Universität Potsdam, SFB 632
Karl-Liebknecht-Str. 24/25, 14476 Golm
{dipper|goetze}@ling.uni-potsdam.de; skopetea@rz.uni-potsdam.de

Series Editors: Shinichiro Ishihara
Universität Potsdam, SFB 632
Karl-Liebknecht-Str. 24/25, 14476 Golm
ishihara@rz.uni-potsdam.de; mschmitz@ling.uni-potsdam.de

Stefanie Jannedy, Anne Schwarz
Humboldt-Universität zu Berlin, SFB 632
Sitz: Mohrenstr. 40-41
Unter den Linden 6, D-10099 Berlin
anne.schwarz@rz.hu-berlin.de

Published by Universitätsverlag Potsdam
Postfach 60 15 53, D-14415 Potsdam
Fon +49 (0) 331 977 4517
Fax +49 (0) 331 977 4625
e-mail: ubpub@rz.uni-potsdam.de
http://info.ub.uni-potsdam.de/verlag.htm

Foto: U. Baumann

Printed by Audiovisuelles Zentrum der Universität Potsdam
Published 2007

Contents

Introduction

Stefanie Dipper, Michael Götze, Stavros Skopeteas
University of Potsdam

The annotation guidelines introduced in this chapter present an attempt to create a unique infrastructure for the encoding of data from very different languages. The ultimate target of these annotations is to allow for data retrieval for the study of information structure, and since information structure interacts with all levels of grammar, the present guidelines cover all levels of grammar too. After introducing the guidelines, the current chapter also presents an evaluation by means of measurements of the inter-annotator agreement.

Information structure (IS) is an area of linguistic investigation that has given rise to a multitude of terminologies and theories, that are becoming more and more difficult to survey. The basic problem is that IS-related phenomena can often be observed only indirectly on the linguistic surface and hence invite competing interpretations and analyses tailored to the needs and taste of individual researchers. Thus, in contrast to syntax, where different approaches can be - more or less - systematically compared, with IS it is often not even clear whether two theories compete to describe the same phenomenon or are in fact complementary to each other, characterizing linguistic regularities on different levels of description.

In 2003, a long-term research infrastructure ('Sonderforschungsbereich', henceforth 'SFB') was established at Potsdam University and Humboldt-University Berlin (http://www.sfb632.uni-potsdam.de). Its aim is to investigate the various facets of IS from very different perspectives and to contribute to a

Interdisciplinary Studies on Information Structure 07 (2007): 1–28
Dipper, S., M. Götze, and S. Skopeteas (eds.):
Information Structure in Cross-Linguistic Corpora

broader and more general understanding of IS phenomena by bringing the various results together and promoting the active exchange of research hypotheses. Participating projects provide empirical data analyses to serve as the basis for formulating theories, which, in turn, seek to advance the state of the art and overcome the undesirable situation characterized above.

An important prerequisite for this long-term and multi-disciplinary approach is the ability to *annotate* IS data with appropriate information. From the very beginning, it has been an important goal of the SFB to develop common annotation guidelines that can be used in the annotation of SFB corpora and thus make it possible to exploit and compare data across individual SFB projects. Moreover, detailed descriptions of the criteria that were applied during annotation would render the SFB corpora a valuable resource for the research community.

Specific SFB-wide working groups dedicated to various levels of analysis were set up and met regularly over a period of several months to develop annotation guidelines. Draft versions were tested by a group of students and, in addition, reviewed by linguist experts within the SFB. The main focus of the SFB is obviously on the annotation of Information Structure, which in our guidelines builds on syntactic information (NPs, PPs, and sentential constituents). Hence, we place special emphasis on the evaluation of the Syntax and IS guidelines and performed a three-day test annotation of these sections. The results of this evaluation, including Kappa measures, are presented below.

In Section 1, we present the general requirements and design decisions of our annotation guidelines. Section 2 gives overviews of the individual annotation layers, in Phonology, Morphology, Syntax, Semantics and Information Structure. Section 3 contains the details of the Syntax/IS evaluation.

A fully-annotated sample is provided in the appendix to the book along with an overview of all tagsets.

We would like to thank all the members of the SFB who actively participated in the development of the guidelines, as authors and/or reviewers.[1]

1 Requirements and Design Decisions

Due to the diverse goals and methods of the individual SFB projects, the SFB corpora do not represent a homogeneous set of data. First, the corpora differ with regard to the language of the primary data. There are corpora ranging across 18 different languages, including typologically diverse languages such as Chinese, Dutch, English, Canadian and European French, Georgian, German, Greek, Hungarian, Japanese, Konkani (India: Indo-European), Manado Malay, Mawng (Australia: Non-Pama-Nyungan), Niue (Niue Island: Austronesian), Old High German, Prinmi (China: Tibeto-Burman), Teribe (Panama: Chibchan), and Vietnamese. Second, primary data may consist of written texts or spoken/spontaneous speech, complete or fragmentary utterances, monologues or dialogues. The heterogeneity of the data resulted in the following requirements.

- The annotation guidelines should be language independent. For instance, they must provide criteria for agglutinative as well as isolating languages. Hence, in addition to English examples, many of the annotation instructions are supplemented by examples from other languages.
- The guidelines should be as theory independent as possible. Researchers within the SFB come from different disciplines and theoretical backgrounds, and the guidelines should therefore rely on terms and concepts that are commonly agreed on and whose denotations are not

[1] Special thanks are also due to the students who tested different versions of the guidelines: Anja Arnhold, Sabrina Gerth, Katharina Moczko, and Patrick Quahl.

disputable *in general*. For instance, notions such as “subject” are obviously still difficult to define exhaustively. However, in the majority of the cases, subjecthood can be determined straightforwardly. That is, the *core concept* of subjecthood is sufficiently well-defined to be a useful notion in the annotation criteria.

- The guidelines should be easy to apply. Often the guidelines provide criteria in the form of decision trees, to ease the annotation process. Similarly, the guidelines focus on the annotation of *relevant* information. For instance, the exact details of the form of a syntactic tree are often irrelevant for IS applications, whereas information about the arguments of the verbal head of the sentence will be extremely useful for many users. As a result, syntactic annotations according to the guidelines do not result in fully-fledged trees but in a detailed labeling of all arguments in a sentence, including the syntactic category, grammatical function, and theta role.
- The guidelines presuppose basic linguistic knowledge. For instance, it is assumed that the user knows the difference between ordinary verbs, modal verbs, and auxiliaries.
- The guidelines should cover both coarse- and fine-grained annotations. Most of the SFB guidelines specify a *core tagset* and an *extended tagset*. The core part is the obligatory part of the annotation, whereas the extended part provides instructions for the annotation of more fine-grained labels and structures. The user is free to opt for either one, according to her/his needs.
- The guidelines should cover all IS-related information. Information Structure is interweaved with various, if not all, linguistic levels. For instance, word order (i.e., syntax), pitch accent (phonology) and particles

(morphology) etc., all play important roles in structuring information in an utterance. Accordingly, there are guidelines for the annotation of phonology, morphology, syntax, semantics/pragmatics, as well as information structure itself.

2 The Annotation Layers

Each of the individual guidelines in this book consists of the following components:

- Preliminaries and general information
- Tagset declaration of the annotation scheme
- Annotation instructions with examples

In this section, we present a general picture of each annotation layer, by summarizing the most important features and principles of the annotation criteria.

2.1 Phonology

The annotation guidelines for phonology and intonation include general orthographic and phonetic transcription tiers (the ‘words’ and ‘phones’ tiers), which are essential for all users of the data, as well as tiers for more specific transcriptions of information relating to the phonetics, phonology and prosody of the utterance.

This additional detailed prosodic information is vital for analysis of information structure because many languages are known to make use of prosodic means, either partially or exclusively, for the expression of information structure categories. A range of tiers is provided from which annotators may select a subset appropriate for the language under investigation. For example, in a tone language, underlying and/or surface tonal behaviour can be captured on different

tiers (‘lextones’ and ‘surface’, respectively), whereas in an intonational language, pitch events of all types (pitch accents, phrase tones, or both) can be labeled on the ‘int-tones’ tier using a language-specific prosodic transcription scheme (cf. Ladd 1996, Jun 2005), alongside information about word- and sentence-stress (‘stress’ and ‘accent’). In a language for which an intonational analysis is not yet available, provision is made for a more phonetic labeling of intonation (in the ‘phon-tones’ tier). Finally, since prosodic phrasing is common to all languages, regardless of prosodic type, phrasing at two layers corresponding to the Phonological Phrase and Intonational Phrase layer can be annotated (‘php’ and ‘ip’).

2.2 Morphology

This level contains the three elementary layers necessary for interpretation of the corpus. It provides the user of the database with information about the morphological structure of the archived data, a morpheme-by-morpheme translation, as well as information about the grammatical category (part of speech) of each morpheme. This level is vital for linguists that aim at syntactic analysis or semantic interpretation of data from object languages that they do not necessarily speak.

The information within this level is organized as follows: First, a morphemic segmentation of the data is given, in which the boundaries between morphemes are indicated (‘morph’). The next layer includes morphemic translations and corresponds in a one-to-one fashion to the segmentation of morphemes in the previous layer (‘gloss’). Each morphemic unit of the object language is either translated into English or “glossed” with a grammatical label. Finally, the morphological category of each word is given in a third layer (‘pos’). The guidelines for morphology follow existing recommendations in

language typology (see *Leipzig Glossing Rules*, Bickel et al. 2002, *Eurotyp*, König et al. 1993) and norms for the creation of language corpora (see *EAGLES*, Leech & Wilson 1996; *STTS*, Schiller et al. 1999).

2.3 Syntax

Based on the morphological information which is given at the previous level, the level of syntax gives a representation of the constituent structure of the data, including syntactic functions and semantic roles. Since information structural generalizations are often correlated with particular constituent types, this layer is designed to enable the retrieval of data that display particular syntactic properties; for instance, to set queries for preverbal constituents, subjects or agents, or for a combination of these categories.

Syntactic information is organized in three layers. The layer "constituent structure" ('cs') provides a number of simplified and theory independent conventions for the annotation of maximal projections. The layer "function" contains information about different types of constituents such as main vs. subordinate clauses, arguments vs. adjuncts, subjects vs. objects, etc. Finally, the layer "role" contains an inventory of semantic roles (agent, theme, experiencer, etc.) which are annotated in relation to the syntactic functions. The syntactic guidelines are partially related to other syntactic annotation standards such as the Penn Treebank (Santorini 1990), GNOME (Poesio 2000), TIGER corpus (Albert et al. 2003), and Verbmobil (Stegmann et al. 2000).

2.4 Semantics

The annotation guidelines for Semantics focus on features that are decisive for the semantic interpretation of sentences and are often related to or even act together with information structural properties. These include in particular quantificational properties (e.g. quantifiers and scope relations, in the layers

'QuP' and 'IN'), but also more general semantic/pragmatic features such as definiteness ('DefP'), countability ('C'), and animacy ('A').

2.5 Information Structure

For the annotation of Information Structure (IS), three dimensions of IS were selected: Information Status (or Givenness) ('infostat'), Topic ('topic'), and Focus ('focus'). The choice was driven by the prominence of these dimensions in linguistic theories about IS, and by their usage across different theoretical frameworks and in the research center. The single dimensions distinguish further subcategories, e.g. aboutness and frame-setting topic within 'Topic', or new-information focus and contrastive focus within Focus.

Aiming at applicability of the annotation scheme to typologically diverse languages, the annotation instructions use functional tests to a large degree - without reference to the surface form of the language data. Furthermore, we annotate the features of the IS dimensions independently from each other, thus avoiding postulation of relationships between potentially different aspects of IS. Hierarchical annotation schemes and decision trees facilitate a consistent annotation.

Other approaches to the annotation of IS differ from ours by being language and theory specific (e.g., Hajicova et. al 2000) or by focussing on the annotation of only one aspect of IS (e.g., Calhoun et al. 2005 for Information Status). Indeed often, the detailed annotation guidelines are not published.

3 Evaluation[2]

We investigated inter-annotator agreement for syntax and information structure by calculating *F-scores* as well as *Kappa* (Cohen 1960, Carletta 1996) between two annotators.

The annotators, two students of linguistics, took part in a three-day test annotation. The students started with an intensive half-day training for annotation of both syntax and IS. In the actual test annotation, they first annotated syntactic constituent structure (constituents and their categorial labels). The annotations were then checked and corrected by us. Next, the students annotated IS, based on the corrected syntactic constituents. The annotation tool that we used in the evaluation was EXMARaLDA.[3]

As described in Section 1, the data of the SFB is highly heterogeneous and includes both written texts and spontaneous speech, complete and fragmentary utterances, monologues and dialogues. As a consequence, annotators face various difficulties. For instance, written newspaper texts often feature complex syntactic structures, such as recursively-embedded NPs. In contrast, the syntax of spoken language is usually less complex but it exhibits other difficulties such as fragmentary or ungrammatical utterances. Similarly, the annotation of IS in running text differs a lot from question-answer pairs. We therefore decided to select a sample of test data that reflects this heterogeneity:

- 20 question-answer pairs from the typological questionnaire QUIS (Skopeteas et al. 2006) (40 sentences)
- 2 dialogues from QUIS (60 sentences)

---

[2] Many thanks to Julia Ritz for invaluable help with the evaluation.

[3] http://www1.uni-hamburg.de/exmaralda/. EXMARaLDA uses annotation tiers, so that constituents (or segments) can be annotated by one feature only. For annotating multiple features of a segment, such as “NP” and “given”, the student annotators had to copy the segment from the syntax tier to the information-status tier.

- 7 texts of newspaper commentaries from the Potsdam Commentary Corpus (100 sentences)

Altogether, the test data consisted of 200 German sentences with approx. 500 nominal phrases (NP) and 140 prepositional phrases (PP). The following table displays the annotated features and their (core) values. For a description of these features and the complete set of values, see the Annotation Guidelines for Syntax (Chapter 2) and Information Structure (Chapter 6), respectively.

Table 1: Annotated features and core values

	Feature	Values
Syntax		S, V, NP, PP, AP
Information Structure	Information Status	acc, giv, new
	Topic	ab, fs
	Focus	nf, cf

Usually, annotations are evaluated with respect to a *gold standard*, an annotated text whose annotations are considered "correct". For instance, automatic part-of-speech tagging can be evaluated against a manually-annotated, "ideal" gold standard. In our case, however, we want to evaluate *inter-annotator consistency*, that is, we compare the results of the two annotators.

We distinguish two tasks in the evaluation: (i) *bracketing*: determining the boundaries of segments; and, (ii) *labeling*: annotating a feature to some segment (e.g., "NP"). Labels for the annotation of IS can be taken (a) from the *core* set or (b) from the *extended* set of labels.

3.1 Calculating F-scores

For F-score calculation, we used the following measures: Segments that have been bracketed (and labeled) the same way by both annotators are considered as

"exact matches". *Overlapping* segments, i.e., segments that share some tokens while the left and/or right boundaries, as marked by the two annotators, do not match exactly, are considered "partial matches". All other segments marked by one of the annotators (but not by the other) are considered as "not matching".

We calculate "precision", "recall", and "F-score" (the harmonic mean of precision and recall) of the annotators A1 and A2 relative to each other (Brants 2000). In addition, we *weight* the matches according to their matching rate, which is the ratio (F-score) of shared and non-shared tokens. This means that exact matches are weighted by 1, not-matching segments by 0. The weighting factor f of partial matches, a kind of 'local' f-score, depends on the amount of shared tokens, with $0 < f < 1$.[4]

(1) $$Precision(A1,A2) = Recall(A2,A1) = \frac{AMR \times \#matches(A1,A2)}{\#segments(A1)}$$

(2) $$Recall(A1,A2) = Precision(A2,A1) = \frac{AMR \times \#matches(A1,A2)}{\#segments(A2)}$$

(3) $$F\text{-}score(A1,A2) = \frac{2 \times Precision(A1,A2) \times Recall(A1,A2)}{Precision(A1,A2) + Recall(A1,A2)}$$

The average matching rate *AMR* is calculated as the average of all matching rates (*matchRate*). The matching rate of individual matches $match_{A1,A2}$ is:[5]

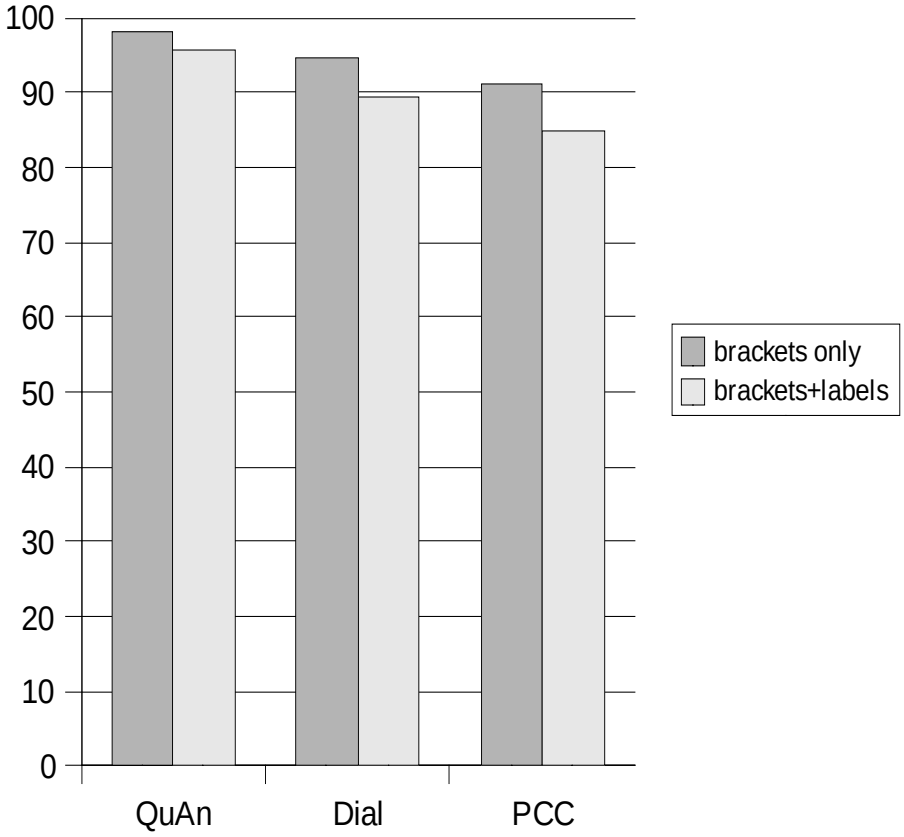
(4) $$matchRate(match_{A1,A2}) = \frac{2 \times \#sharedTokens(A1,A2)}{\#tokens(A1) + \#tokens(A2)}$$

[4] Since *Precision(A1,A2) = Recall(A2,A1)*, it holds that *F-score(A1,A2) = F-score(A2,A1)*.

[5] For constituent-based annotations such as syntax, it would make sense to compare the number of shared and non-shared dominated *nodes* rather than *tokens*. However, the tier-based annotation tool EXMARaLDA does not easily allow for infering constituent structure.

The average matching rate can be computed (i) for *all* matches, i.e., including exact and partial matches as well as non-matching segments, or else (ii) for the *partial* matches only.

Figure 1: Syntax evaluation results across text types (F-scores)



3.1.1 Syntax evaluation

Figure 1 shows the results of the syntax evaluation for the different text types. The first column pair encodes the results for the question-answer pairs (QuAn), the second for the dialogue data (Dial), the third for the data from the Potsdam Commentary Corpus (PCC). The columns in dark-grey correspond to the F-score of task (i), i.e., the bracketing task, while ignoring the labeling of the segments. The F-scores for the three text types are 98.04%, 94.48%, and

91.03%, respectively. The columns in light-grey show to what extent agreement decreases when labeling is also taken into account (task (ii)). The respective F-scores are 95.74%, 89.37%, and 84.79%.

Figure 1 shows that the question-answer pairs are the least controversial data with regard to syntax, while the PCC newspaper texts turned out to be considerably more difficult to annotate.

Figure 2: F-scores of individual categories (PCC data)

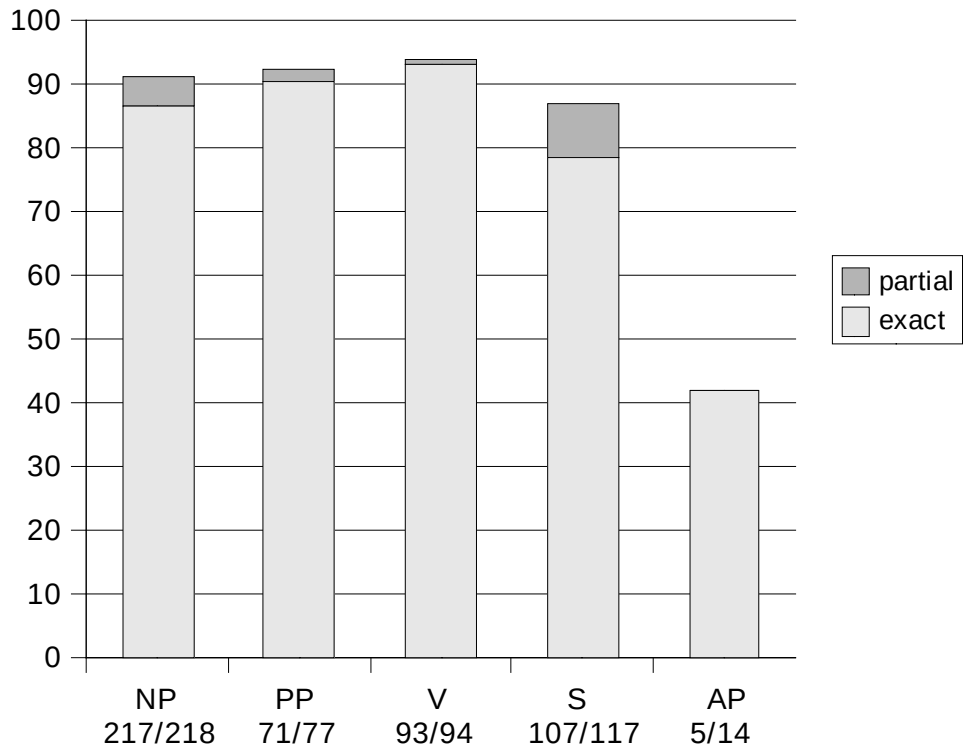


Figure 2 displays the results for use of individual labels within the PCC dataset.[6] For each category, we report the number of times it was used by each annotator (e.g., the label “NP” was used 217 times by one of the annotators, and 218 times by the other). The F-scores of NP, PP, and V are comparably high (> 90%), while S reaches 86.85% only. The agreement on annotation of AP is even lower,

[6] We did not include discontinuous constituents, annotated as “NP_1” etc., in this evaluation.

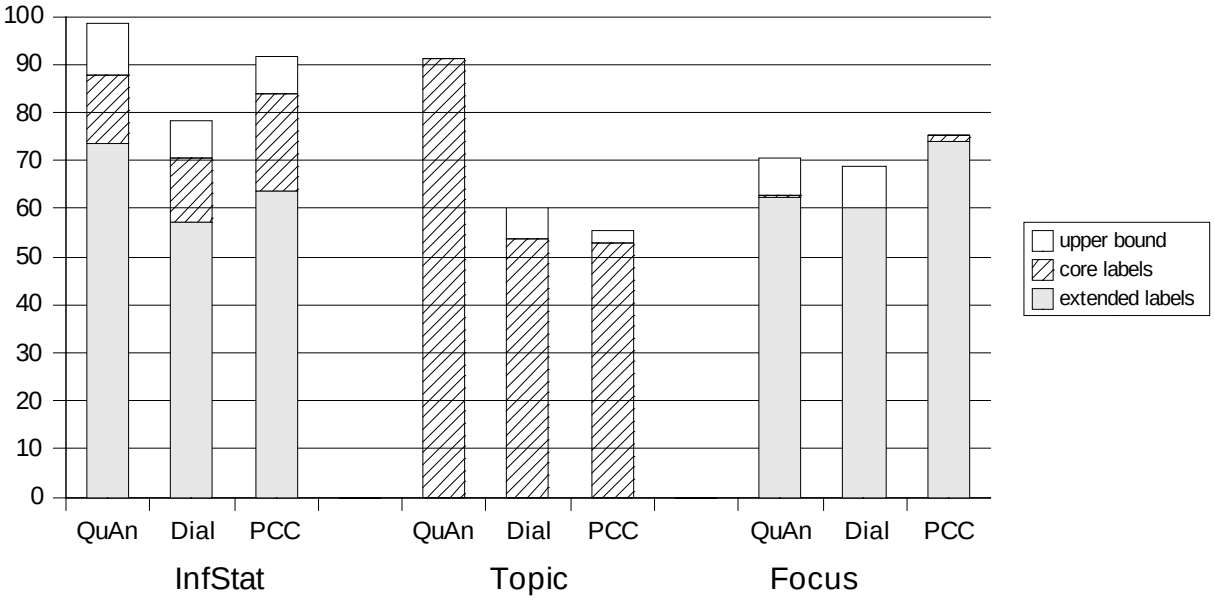
with an F-score of 42.11%, which can be attributed to the fact that one of the annotators found 14 APs and the other only 5. The top parts of the columns, which correspond to the (weighted) portions of partial matches, indicate that partial agreement occurs more prominently with S and NP segments than with the other categories.

3.1.2 IS evaluation

The IS evaluation considers annotation of Information Status, Topic, and Focus. As described above, the annotations of IS were performed on gold-standard syntactic constituents. That is, for the segments to be marked for Information Status and Topic, which most often correspond to NP or PP segments, the segment boundaries were already given. Nevertheless, the two student annotators disagreed from time to time with respect to the bracketing task. This is in part due to the fact that they had to manually copy the syntactic segments that they wanted to annotate using IS features to the respective IS tiers (see footnote 3). Hence, whenever one of the annotators decided that some NP or PP was referential and, hence, had to be copied and annotated, while the other decided that it was non-referential, this resulted in bracketing disagreement. Obviously, such disagreements must be classified as *labeling* disagreements, since they are connected to the status of referentiality of some NP, not to its extension. Agreement on *bracketing* thus puts an upper bound on the labeling task: obviously, only segments that both annotators decided to copy can be labeled the same way by both of them.

Figure 3 displays F-scores for both the core set (task (iia)) and the extended set (task (iib)) of features (for Topic annotation, an extended tagset has not been defined). Figure 3 also marks the upper bound, as given by the “same extension” (identical bracketing) condition.

Figure 3: IS labeling (F-scores)



The figure displays the labeling results for all test data. The first group of columns encodes the results for the annotation of Information Status ("InfStat"), the second for Topic, and the third for Focus. Within each of the groups, the first column displays the results for the text sort question-answer pairs ("QuAn"), the second the dialogues ("Dial"), and the third the PCC texts. In the following, we point out the most prominent differences in Figure 3.

- Looking at the results of core labeling, we see that *on average* the annotation of InfStat is the easiest task, yielding agreements between 87.90% (with the QuAn data) and 70.50% (with Dial data).
- The overall highest agreement is achieved with Topic annotation of the QuAn data: 91.14%. Interestingly, Topic annotations with Dial and PCC result in the overall worst agreements: 53.52% and 52.72%. That is, the F-scores of Topic annotation vary enormously depending on the text type, whereas InfStat and Focus annotations result in rather uniform F-scores. The Topic results for the QuAn data might be attributed to the fact that

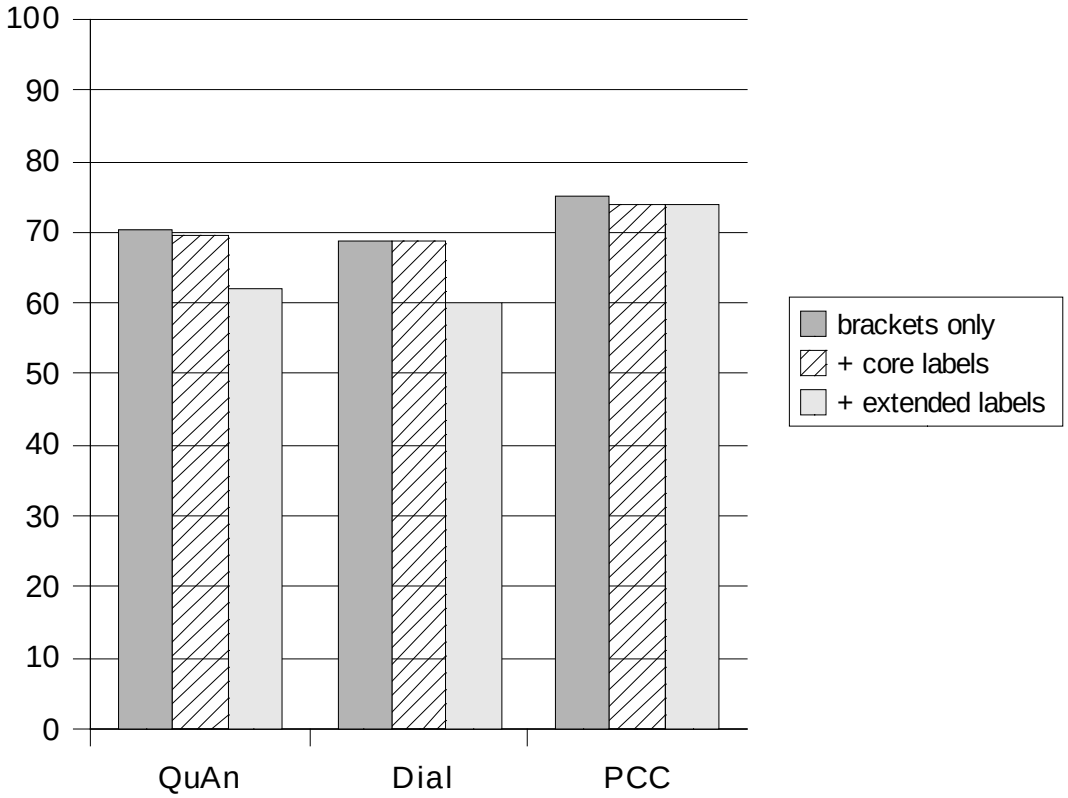
this text type contains highly constrained language content, in the form of short question-answer pairs, which appear to be suitable input for the Topic annotations.

- In contrast to syntax, annotating IS gives rise to discrepancies more in the Dial data than in the PCC data. Surprisingly, highest annotation agreement is reached for Focus in the PCC data.
- Comparing core and extended tagsets, we have to look at the portions in different colors (for InfStat and Focus only). The shaded part indicates to what degree the fine-grained, extended tagset introduces disagreement among the annotators. It turns out that this makes some difference with InfStat annotations but not with Focus annotations.
- Finally, looking at the upper bound of possible agreement, indicated by the white-marked portion at the top of each column (for InfStat and Topic[7]), we see that for InfStat annotation, the annotators quite often agreed in general on the referential status of some NP or PP, while disagreeing on the exact label, whilst this happened less often for Topic annotation.

In contrast to Information Status and Topic, Focus annotation does not rely on NP or PP segments. Hence, it makes sense to look more closely at the difficulty of task (i) which involves defining the scope of the various Focus features. Figure 4 displays the three tasks, (i), (iia), and (iib) in groups of columns for Focus annotation only.

[7] For interpretation of the "upper bound" for Focus annotation, see below.

Figure 4: Focus annotation, IS evaluation results



The figure shows that within each group of columns, the differences between the three tasks are rather small, especially in the core tagset, that is, annotators tend to label identical segments in the same way. Put differently: the difficult task is to determine the *scope* of some Focus feature, not its type.[8]

Weighting partial matches: We penalize partial agreement by multiplying the numbers with the average matching rate. With InfStat and Topic annotation, this does not have much impact on the final results, since the annotations rely on pre-defined NP and PP segments and rarely deviate in their extensions. With Focus annotation, however, the annotators had to mark the boundaries by themselves, hence, the proportion of partial-only matches is considerably higher.

[8] The differences between the measures "brackets only" and "+ core labels" are very subtle and thus hard to distinguish in the figure: 0.74 percentage points for QuAn (brackets only: 70.39%; core labels: 69.65%), 0.00 for Dial (brackets and core labels: 68.69%), and 1.09 for PCC (brackets: 75.09%; core labels: 74.00%).

Figure 5: Focus annotation, exact and partial agreement

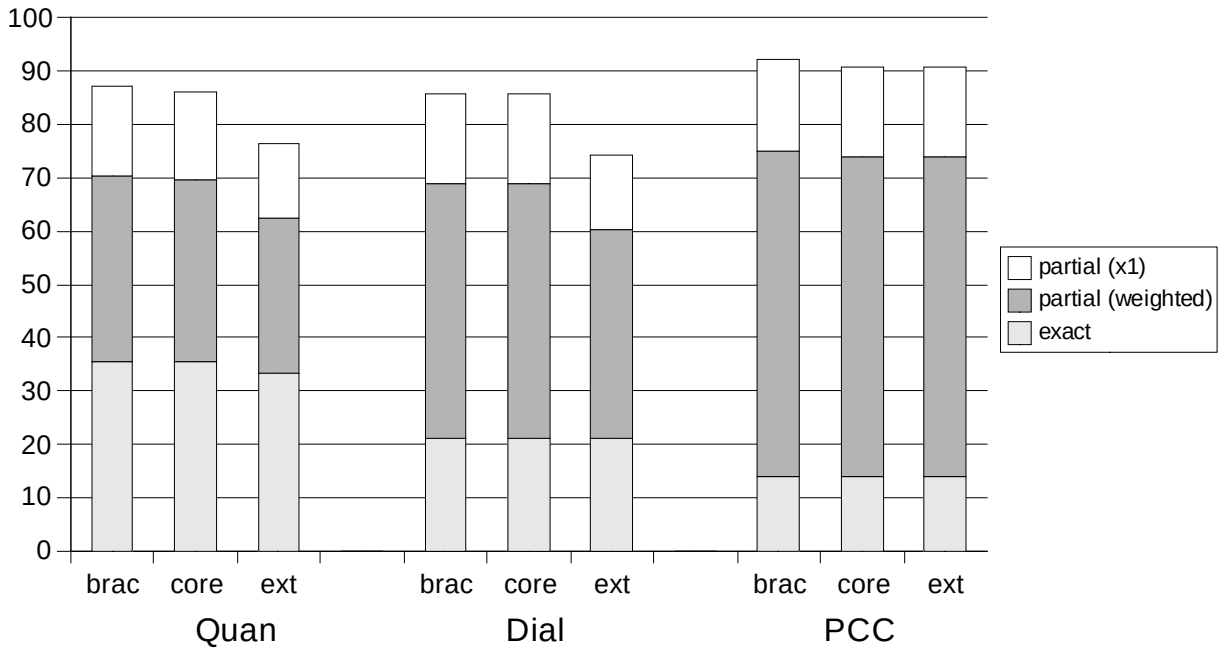


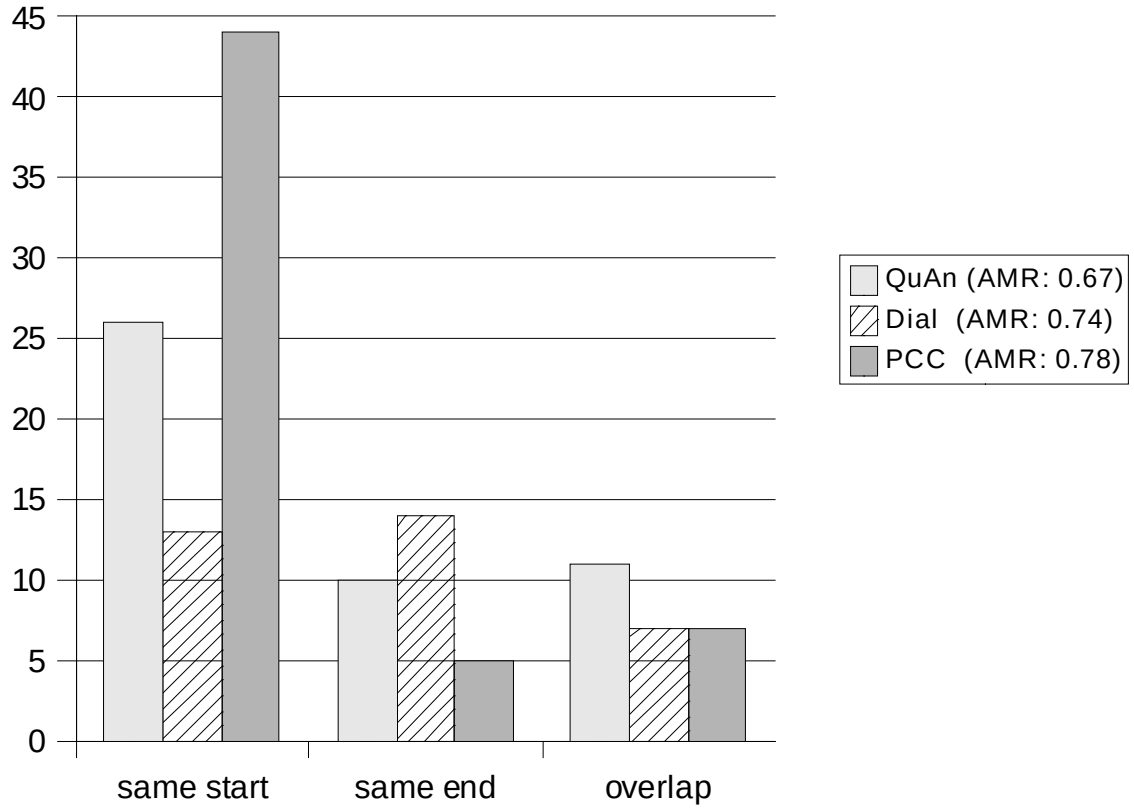
Figure 5 shows the F-scores of exact matches only (light-grey part), the F-scores when weighted partial matches are added (dark-grey part), and the F-scores that result if partial agreement is not weighted, i.e., not penalized at all (white part on top).[9]

We can see from Figure 5 that annotators disagree on the scope of focused segments more often than they agree, especially in the PCC data. The discrepancies are striking: exact agreement is at 13.99% across all three tasks, as opposed to 74.00%-75.09% agreement, when partial matches are also taken into account.

Figure 6 provides more detail about the partial matches. The annotators can agree with respect to the left boundary while disagreeing with respect to the right boundary ("same start"), or vice versa ("same end"), or else they disagree on both boundaries but mark some tokens within the same region ("overlap").

[9] The columns put in dark-grey encode the same information as the columns in Figure 4.

Figure 6: Focus annotation, details on partial matches



The figure shows that the annotators quite often agreed with regard to the starting point of a focused constituent. The average matching rate (AMR) of partial matches, which indicates to what extent the partially-matching segments overlap, is lowest for the QuAn data (0.67) and highest for the PCC data (0.78). Comparing these numbers with the results displayed in Figure 5, we see that among the different text types, the QuAn data yields the highest F-score of exact matches (cf. the light-grey parts in Figure 5), and, at the same time, the lowest AMR of partial matches. This suggests that in those cases where segmentation is not straightforward, (transcribed) spoken data is more difficult to segment than written data.

3.2 Calculating Kappa

A weak point of the F-score measure is the fact that it does not factor out *agreement by chance*. A measure like Kappa takes chance agreement into account, by subtracting chance agreement from the observed agreement. Kappa is computed as:

$$\kappa = \frac{P(O) - P(E)}{1 - P(E)} \quad (5)$$

where *P(O)* is the relative observed agreement among the annotators, and *P(E)* is the probability of agreement by chance. If the annotators' agreement is very high, κ approximates 1, if there is no agreement other than by chance, κ = 0.[10] A κ > 0.8 is usually considered as indicative of good reliability and .67 < κ < 0.8 allows for "tentative conclusions" to be drawn (Carletta 1996, Krippendorf 1980).[11]

For estimating chance agreement *P(E)* of some feature *F*, we have to know the probability of the annotators to annotate *F*. IS features, however, are annotated to segments, that is, we first have to estimate for each token the probability that the annotators mark a segment boundary at that place. To ease the evaluation, we therefore restrict ourselves to the NP segments of the syntax gold annotation, which was presented to the annotators in the IS test annotation. As a consequence, we do not evaluate the annotations of Focus, since Focus does not rely on the pre-defined NP segments.

The observed agreement $P_F(O)$ for some Feature *F* is then calculated as:

---

[10] Kappa is usually given as a number between 0 and 1 rather than as a percentage.

[11] For a critical assessment of the Kappa measure, see, e.g., Artstein & Poesio (2005). They found that "substantial, but by no means perfect, agreement among coders resulted in values of κ or α around the .7 level. But we also found that, in general, only values above .8 ensured a reasonable quality annotation [...] On the other hand, even the lower level .67 has often proved impossible to achieve in CL research, particularly on discourse".

(6) $$P_F(O) = \frac{\#match_F(A1,A2)}{\#NP}$$

where *A1* and *A2* are the annotators, *#match$_F$(A1,A2)* is the number of times the annotators agreed to mark *F* at some NP segment, and *#NP* is the total number of NP segments. The expected agreement *P_F(E)* is computed as:

(7) $$P_F(E) = P_{A1}(F) \times P_{A2}(F)$$

where *P_A(F)* is the probability of annotator *A* to annotate *F* to an NP segment.[12]

The Kappa measure diverges from F-score or percent agreement[13] in particular with features whose values do not occur uniformly distributed, i.e. each with the same frequency. For instance, assume that the feature F can have values V1 and V2. If the annotation F=V1 occurs very often in the data, but not F=V2, it is not surprising if both annotators agree on F=V1 quite often. This fact is taken into account by the Kappa measure.

Figures 7 and 8 illustrate this fact for the features InfStat and Topic. In the PCC data in Figure 7, the values for InfStat ("giv", "new", "acc", and "—"[14]) occur with similar frequencies, whereas for Topic, one of the values ("—") is highly prevalent. Accordingly, the difference between percent agreement and Kappa is greater in the Topic evaluation than with InfSta (see Figure 8). For instance, for Topic annotation in the Dial data, the value drops from 82.00% to a Kappa value of 0,50. The general picture, however, remains the same: QuAn data are easier to annotate than Dial or PCC data, and agreement with respect to Topic annotation varies considerably depending on the text type.

---

[12] For multi-valued features, *P_F(E)* is computed for each value and summed up.

[13] Percent (or percentage) agreement measures the percentage of agreement between both annotators, i.e., the number of segments that the annotators agreed on divided by the total number of segments (in our case: NP segments).

[14] "—" indicates that no value was annotated to the NP segment. With InfStat annotations, this may happen because none of the criteria applied. For Topic annotations, "—" indicates "Comment" segments.

Figure 7: IS evaluation, value distribution (PCC data)

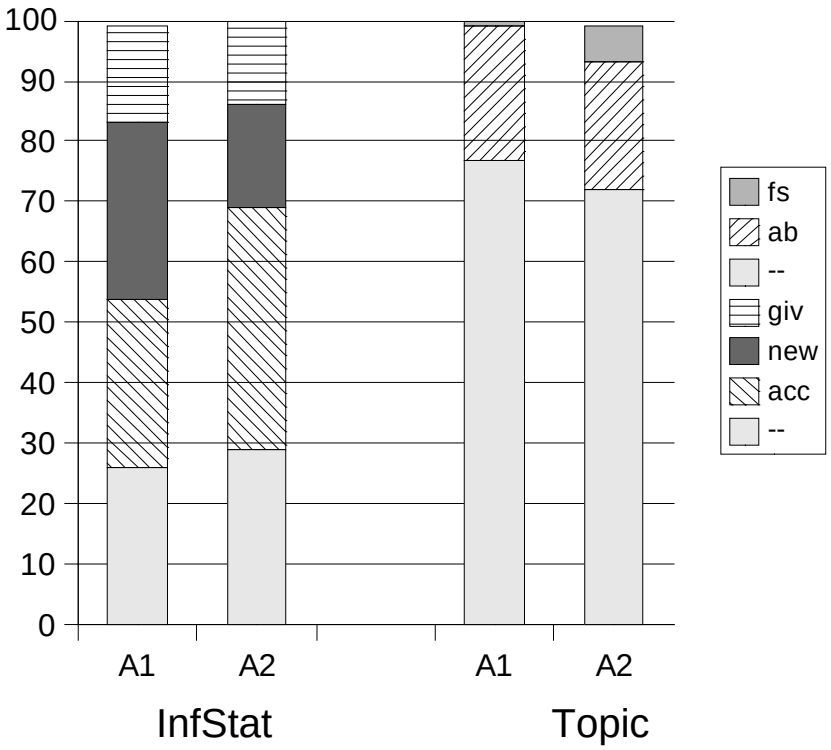
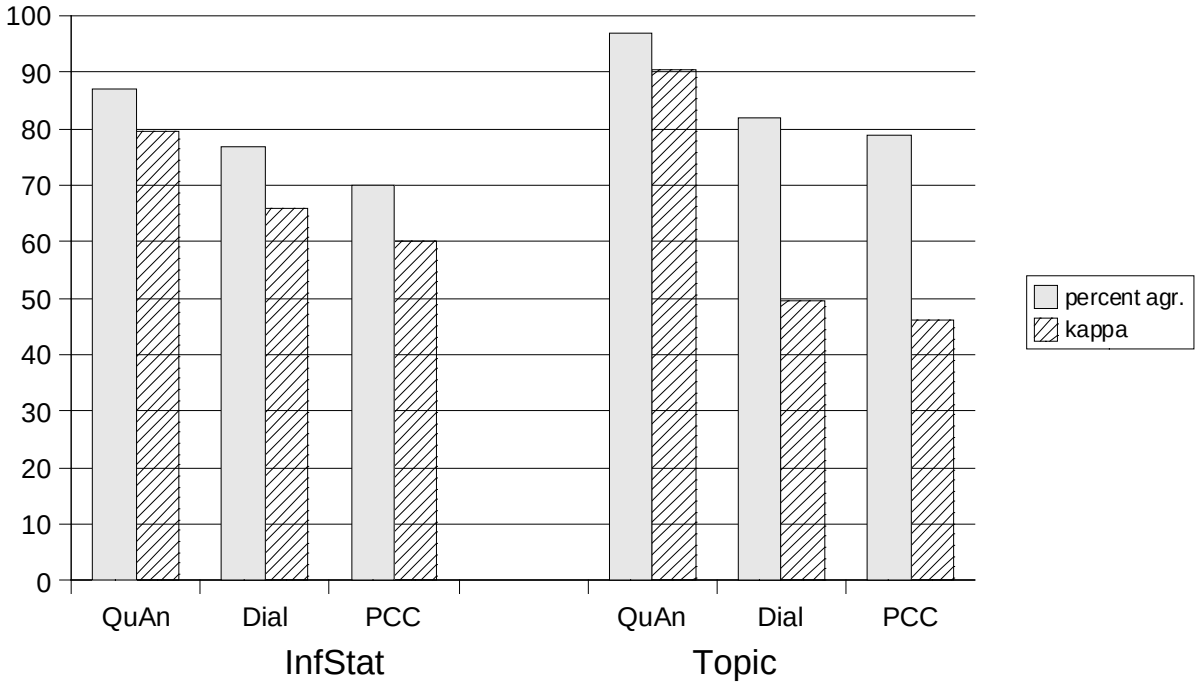


Figure 8: IS evaluation, percent agreement vs. kappa



3.3 Summary of the Evaluation

Syntax evaluation: The syntax evaluation shows that our (transcribed) spoken data is easier to annotate than the newspaper texts. The annotation of the dialogue data results in very high F-scores: 97.87% for unlabeled bracketing, 95.61% for labeled bracketing. Agreement in the PCC newspaper data is 90.04% (unlabeled) and 84.04% (labeled). The evaluation presented by Brants (2000) was also performed on German newspaper texts, and he reports an inter-annotator agreement of 93.72% (unlabeled F-score) and 92.43% (labeled F-score). However, the annotators in his evaluation were supported by a semi-automatic annotation tool, and the annotations consisted of syntax graphs rather than segments on tiers.

IS evaluation: The results obtained by the test IS annotation are more varied. The annotation of InfStat yields acceptable agreement, with F-scores of 87.90% (QuAn data), 70.50% (Dial), and 83.76% (PCC), and, for NPs, Kappa values of 0.80 (QuAn), 0.66 (Dial), and 0.60 (PCC). Topic annotation, in contrast, turned out to be a difficult task, resulting in high agreement only for the QuAn data: 91.14% F-score, 0.91 Kappa value; in contrast, for the Dial and PCC data, Topic annotation yielded rather poor agreement. The level of challenge of Focus annotation lies between that of InfStat and Topic.

We do not know of any comparable evaluation for German data. For English, inter-annotator agreement of annotation of *Information Status* has been evaluated: Nissim et al. (2004) report Kappa values of 0.845 (with four categories) and 0.788 (with a fine-grained tagset) for English dialogue data from

the Switchboard corpus.[15] Hempelmann et al. (2005) report Kappa values of 0.74 (with six categories) and 0.72 (seven categories) for English narrative and expository texts.

Postolache et al. (2005) and Vesela et al. (2004) present results for topic and focus annotations of the Prague Dependency Treebank, which consists of texts from Czech newspapers and a business weekly: percentage agreements of 86.24% (with a two-feature distinction, essentially encoding information about contextual boundedness) and 82.42% (with a three-feature distinction, including contrastiveness of bound elements). They did not compute Kappa values.

Training of the annotators has considerable impact on the results, as reported by Nissim et al. (2004) and Vesela et al. (2004). The annotators taking part in our three-days evaluation certainly did not have much time to absorb their training or to discuss the guidelines. Moreover, our test texts were highly heterogeneous.

Given the fact that annotating IS is an inherently-subjective task in many respects, e.g., due to differing world knowledge, inter-annotator consistency of IS annotation is hard to achieve. We think that further research should focus on the following aspects:

- Text-type-specific guidelines: e.g., the current methods for recognizing Focus in texts other than dialogues certainly leave room for improvement.
- Encoding of subjective knowledge: e.g., labels such as “acc-inf” (for inferable, accessible entities) or “acc-gen” (for general entities, accessible via word knowledge) could be accompanied by more detailed specifications of the accessibility of the entity. For example, annotators should specify whether they know the entity from personal experience,

[15] They provide a tag “not-understood” for the annotations. Segments annotated by this tag were excluded from the evaluation.

from the news, or due to their educational background. The specifications could also include the annotators' assumptions of the common ground.

- Encoding of subjective interpretations: as stated, e.g., by Reitter & Stede (2003) for the annotation of discourse structure, people perceive texts in different ways, and often, texts – and likewise sentences – can be assigned more than one interpretation. In this vein, an annotation encodes one possible interpretation, and strategies have to be developed as to how to classify and deal with competing annotations: disagreement might result either from (simple) annotation errors or from differences in interpretation.

We see the SFB annotation guidelines as a contribution to research on Information Structure, which has recently moved towards empirical and corpus-linguistic methods. The SFB corpora, which have been annotated according to the guidelines presented in this volume, offer an important resource for further research on IS.

4 References

Albert, Stefanie et al. 2003. *TIGER Annotationsschema*. Draft. Univerisities of Saarbrücken, Stuttgart, and Potsdam.

Artstein, Ron and Massimo Poesio. 2005. Kappa3 = Alpha (or Beta). Technical report CSM-437, University of Essex Department of Computer Science.

Bickel, Balthasar, Bernard Comrie, and Martin Haspelmath. 2002. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Leipzig: MPI for Evolutionary Anthropology & University of Leipzig (http://www.eva.mpg.de/lingua/files/morpheme.html)

Brants, Thorsten. 2000. Inter-annotator agreement for a German newspaper corpus. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000).

Carletta, Jean. 1996. Accessing Agreement on Classification Tasks: The Kappa Statistic, *Computational Linguistics* 249-54.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* 20: 37–46.

Hajičová, Eva, Jarmila Panevová, Petr Sgall, Alena Böhmová, Markéta Ceplová, Veronika Řezníčková. 2000. *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*. ÚFAL/CKL Technical Report TR-2000-09. Prague.

Hempelmann, C.F., Dufty, D., McCarthy, P., Graesser, A.C., Cai, Z., and McNamara, D.S. 2005. Using LSA to automatically identify givenness and newness of noun-phrases in written discourse. In B. Bara (Ed.), *Proceedings of the 27th Annual Meetings of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Jun, Sun-Ah (ed.). 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford, OUP.

König, Ekkehard (with Dik Bakker, Öesten Dahl, Martin Haspelmath, Maria Koptjevskaja-Tamm, Christian Lehmann, Anna Siewierska). 1993. *EUROTYP Guidelines*. European Science Foundation Programme in Language Typology.

Krippendorff, Klaus. 1980. Content analysis. An introduction to its methodology. Beverly Hills: Sage.

Ladd, D. Robert. 1996. *Intonational Phonology*. Cambridge, CUP.

Leech, G., A. Wilson. 1996. *Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Guidelines (EAG--TCWG--MAC/R)* (electronically available at: http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html)

Nissim, M., Dingare, S., Carletta, J., and Steedman, M. 2004. An Annotation Scheme for Information Structure in Dialogue. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal, May.

Poesio, Massimo. 2000. *The GNOME Annotation Scheme Manual*. http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm

Postolache, Oana, Ivana Kruijff-Korbayová, and Geert-Jan Kruijff. 2005. Data-driven Approaches for Information Structure Identification. In *Proceedings of HLT/EMNLP*, pp. 9-16. Vancouver, Canada.

Reitter, David and Manfred Stede. 2003. Step by step: underspecified markup in incremental rhetorical analysis In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, Budapest, 2003.

Santorini, Beatrice. 1990. *Annotation Manual for the Penn Treebank Project*. Technical Report, University. of Pennsylvania

Schiller, Anne, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Universität Stuttgart: Institut für maschinelle Sprachverarbeitung & Universität Tübingen: Seminar für Sprachwissenschaft.

Skopeteas, Stavros, Ines Fiedler, Sam Hellmuth, Anne Schwarz, Ruben Stoel, Gisbert Fanselow, and Manfred Krifka. 2006. *Questionnaire on Information Structure: Reference Manual*. Interdisciplinary Studies on Information Structure (ISIS) 4. Potsdam: Universitätsverlag Potsdam.

Stegmann, R., H.Telljohann, and E. W. Hinrichs. 2000. S*tylebook for the German Treebank in VERBMOBIL*. Technical Report 239. Verbmobil.

Vesela, Katerina, Jırı Havelka, and Eva Hajicova. 2004. Annotators' Agreement:. The Case of Topic-Focus Articulation. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2004)*.

Phonology and intonation

Caroline Féry, Sam Hellmuth, Frank Kügler, Jörg Mayer,
Ruben Stoel and Ruben van de Vijver

University of Potsdam

The encoding standards for phonology and intonation are designed to facilitate consistent annotation of the phonological and intonational aspects of information structure, in languages across a range of prosodic types. The guidelines are designed with the aim that a non-specialist in phonology can both implement and interpret the resulting annotation.

1 Preliminaries

This paper presents the conventions of the SFB 632 concerning the archiving of phonological and intonational information. Targets of the standardization are:

- to archive language data which are not understandable for every database user offhand in a comprehensive way;
- to enable phonological/intonation queries in the database, e.g., “search for a high pitch accent in a focused position”.

Some of the annotation conventions presented below are only comprehensible to phonologists. If you feel uncomfortable with some of the terms (‘mora’, ‘trochee’ could be some), ask for advice rather than using them arbitrarily.

Interdisciplinary Studies on Information Structure 07 (2007): 1–28
Dipper, S., M. Götze, and S. Skopeteas (eds.):
Information Structure in Cross-Linguistic Corpora

2 Levels declaration

The group “Phonology and Intonation” has prepared guidelines for the levels of phonology and intonation. The levels and the corresponding layers are declared in the following table:

Table 1: Levels and layers

Level	Layer	Name
Sound	General information	INFO
	Orthographic transcription	WORDS
	Phonemic transcription (syllables)	PHONES
Metrical Structure	Stress	STRESS
	Accent	ACCENT
Prosodic structure	Mora	MORA
	Foot	FT
	Phonological Words	PW
	Phonological Phrases	PHP
	Intonational Phrases	IP
Tones and Intonation	Underlying tones	TONES
	Surface tones	SURFACE
	Phonetic tones	PHONTONES

2.1 Related existing standards

Following standards have been considered for the development of the current guidelines.

Table 2: Existing standards

	ToBI	IViE	GAT
graphemic transcription	x	x	(x)
phonetic transcription	(c)	(x)	(x)
moraic layer	-	-	-
syllabic layer	-	-	-
accent domain / phonological phrase	-	-	-
intermediate layer	x	-	-
intonation phrase layer	x	x	(x)
utterance layer	(x)	(x)	(x)
underlying tonal layer	-	-	-
surface tonal layer	x	x	x
phonetic layer	-	x	-
prominence layer	-	x	(x)

There are significant differences between the first two and the last of the three standards: ToBI & IViE were developed for tonal transcription of intonation; GAT was developed for transcription of conversational speech data. Although the GAT system concerns a different purpose, it also contains a section about intonational transcription. For the purposes of the SFB 632, layers of prosodic information have to be included in the annotation system.

2.1.1 ToBI

ToBI stands for “Tone and Break Indices”. This standard is based on the theoretical groundwork of intonational phonology by Pierrehumbert (1980), and provides conventions for transcribing the intonation and prosodic structure of

spoken utterances in a language variety. Yet, a ToBI system represents an already conducted phonological analysis of the intonation of a certain language.

For more information of the general idea of ToBI see:

http://www.ling.ohio-state.edu/~tobi/

For labeling conventions of English ToBI (Beckman & Ayers 1997) see:

http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/

2.1.2 IViE

IViE stands for "Intonational Variation in English". This standard was developed on the basis of ToBI. The purpose is to ensure comparable transcription of several varieties of English using a single labeling system. The system has been developed within the research project "English Intonation in the British Isles" (Grabe & Nolan, see: http://www.phon.ox.ac.uk/~esther/ivyweb/). In addition to ToBI, IViE allows for transcription of prominence and phonetic variation. For more information on the labeling conventions see:

http://www.phon.ox.ac.uk/~esther/ivyweb/guide.html

2.1.3 GAT

GAT stands for "Gesprächsanalytisches Transkriptionssystem" (Selting et al. 1998). Its purpose is to provide a standard for transcription of conversational speech. The sequential structure of a conversation is iconically represented in the transcription, where each speaker's turn starts with a new transcription line. It is not clear whether a speaker's turn coincides with a prosodic unit, i.e. whether a one-to-one correspondence exists between a turn and an intonation phrase for instance. In addition to the segmental graphical level, GAT provides labels for prosodic annotation, which are however very impressionistic and rough.

For more information see:
http://www.fbls.uni-hannover.de/sdls/schlobi/schrift/GAT/ and
http://www.fbls.uni-hannover.de/sdls/schlobi/schrift/GAT/gat.pdf

3 Level I: Sound + Transcription

3.1 Declaration

The level consists of 3 layers: Orthographic transcription (WORDS), phonemic transcription (PHONES), and general information (INFO). Words can be annotated in transliteration or in the original orthography. The obligatory layer of broad phonemic transcription (PHONES) is done in either IPA or SAMPA. If you decide to use SAMPA and no version of SAMPA for your language is available, use X-SAMPA (for more information about SAMPA and X-SAMPA, see: http://www.phon.ucl.ac.uk/home/sampa/home.htm). Specify which phonetic alphabet you are using in the general information layer (INFO). You should also specify the sound file name in the INFO layer. Other optional entries to this layer are recording conditions, speaker characteristics, etc.

3.2 Objects of annotation

The annotated sentences must be available as sound files in formats that are readable by PRAAT, preferably in the .wav format. Other possible formats are: aiff, aifc (but not compressed aifc files), au files (NeXT/Sun) and NIST files: big-endian, little-endian, µ-law, A-law, Polyphone (NIST files compressed with **shorten** are not supported).

3.3 Tagset declaration

The WORDS and INFO layers are free text fields which require no tagset. The PHONES layer tagset comprises all IPA or (X-)SAMPA symbols, plus one tag

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC) WITH X-SAMPA EQUIVALENTS IN BLUE

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p p b b			t t d d		ʈ t` ɖ d`	c c ɟ J\	k k ɡ g	q q ɢ G\		ʔ ?
Nasal	m m	ɱ F		n n		ɳ n`	ɲ J	ŋ N	ɴ N\		
Trill	ʙ B\			r r					ʀ R\		
Tap or Flap				ɾ 4		ɽ r`					
Fricative	ɸ p\ β B	f f v v	θ T ð D	s s z z	ʃ S ʒ Z	ʂ s` ʐ z`	ç C ʝ j\	x x ɣ G	χ X ʁ R	ħ X\ ʕ ?\	h h ɦ h\
Lateral fricative				ɬ K ɮ K\							
Approximant		ʋ P (or v\)		ɹ r\		ɻ r\`	j j	ɰ M\			
Lateral approximant				l l		ɭ l`	ʎ L	ʟ L\			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ O\ Bilabial	ɓ Bilabial b_<	’ _> as in:
ǀ |\ Dental	ɗ Dental/alveolar d_<	pʼ Bilabial p_>
ǃ !\ (Post)alveolar	ʄ Palatal J_<	tʼ Dental/alveolar t_>
ǂ =\ Palatoalveolar	ɠ Velar g_<	kʼ Velar k_>
ǁ ||\ Alveolar lateral	ʛ Uvular G_<	sʼ Alveolar fricative s_>

SUPRASEGMENTALS

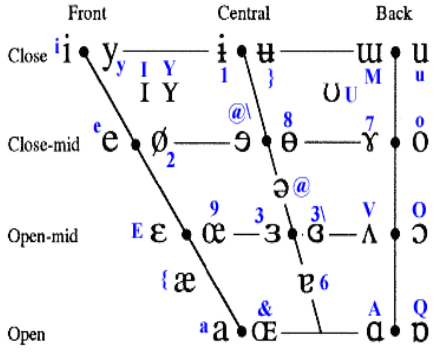
ˈ "	Primary stress	%foUn@"tIS@n
ˌ %	Secondary stress	ˌfoʊnəˈtɪʃən
ː :	Long	eː e:
ˑ :\	Half-long	eˑ e:\
˘ _X	Extra-short	ĕ e_X
. $	Syllable break	ɹi.ækt r\i.{kt
| |	Minor (foot) group	
‖ ||	Major (intonation) group	
‿ -\	Linking (absence of a break)	

TONES & WORD ACCENTS

LEVEL		CONTOUR	
˥ or 7	Extra high _T	˩˥ or /	Rising _R
˦ 1	High _H	˥˩ \	Falling _F
˧ -	Mid _M	˧˥ 7	High rising _H_T
˨ _	Low _L	˨˩ /	Low rising _B_L
˩ _	Extra low _B	˩˥ 7	Rising-falling _R_F
↓ !	Downstep	↗	Global rise <R>
↑ ^	Upstep	↘	Global fall <F>

etc.

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

DIACRITICS

X-SAMPA diacritics come after symbols, e.g. n_0

Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

_0 ◌̥	Voiceless	n̥ d̥	_t ◌̤	Breathy voiced	b̤ a̤	_d ◌̪	Dental	t̪ d̪
_v ◌̬	Voiced	s̬ t̬	_k ◌̰	Creaky voiced	b̰ a̰	_a ◌̺	Apical	t̺ d̺
h _h	Aspirated	tʰ dʰ	_N ◌̼	Linguolabial	t̼ d̼	_m ◌̻	Laminal	t̻ d̻
_O ◌̹	More rounded	ɔ̹	w _w	Labialized	tʷ dʷ	~ (or _~)	Nasalized	e~ or ẽ
_c ◌̜	Less rounded	ɔ̜	j (or _j)	Palatalized	tʲ dʲ	n _n	Nasal release	dⁿ
_+ ◌̟	Advanced	u̟	Y _G	Velarized	tˠ dˠ	l _l	Lateral release	dˡ
_- ◌̠	Retracted	i̠	ʕ _?\	Pharyngealized	tˤ dˤ	ˀ _}	No audible release	d̚
_" ◌̈	Centralized	ë	~ _e	Velarized or pharyngealized	ɫ (or velarized l: 5)			
_x ◌̽	Mid-centralized	e̽	_r ◌̝	Raised	e̝ (ɹ̝ = voiced alveolar fricative)			
= (or _=) ◌̩	Syllabic	ɹ̩	_o ◌̞	Lowered	e̞ (β̞ = voiced bilabial approximant)			
_^ ◌̯	Non-syllabic	e̯	_A ◌̘	Advanced Tongue Root	e̘			
˞ @`	Rhoticity	ɚ	_q ◌̙	Retracted Tongue Root	e̙			

OTHER SYMBOLS

W M	Voiceless labial-velar fricative	s\ z\ ɕ ʑ	Alveolo-palatal fricatives
w W	Voiced labial-velar approximant	ɺ l\	Alveolar lateral flap
H ɥ	Voiced labial-palatal approximant	ʃ x\ ʃ	Simultaneous ʃ and x
H\ H	Voiceless epiglottal fricative		
<\ ʢ	Voiced epiglottal fricative		
>\ ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

k͡p t͡s

k_p t_s

for pauses: <P> (you can use the <P>-tag also in the WORDS layer). The following table has all the phonetic symbols in IPA and in X-SAMPA. The table can be downloaded from the following site:

http://www.diku.dk/hjemmesider/studerende/thorinn/xsamchart.gif

In order to use phonetic symbols in PRAAT one has to have SIL Doulos IPA 1993 font installed. It can be downloaded from the PRAAT site (http://www.praat.org). A tutorial on the use of phonetic symbols in PRAAT can be found in the manual of the PRAAT program itself or here:

http://www.fon.hum.uva.nl/praat/manual/Phonetic_symbols.html.

Some illustrative examples are given below. Note the mismatch between word and syllable boundaries:

Dutch using IPA

(1)

<INFO>	Sound file name, IPA									
<WORDS>	Marie		geeft	Kees	een	klap	op	zijn	kop	.
<PHONES>										
<TRANS>	Marie hits Kees on the head.									

(2) Dutch using SAMPA

<INFO>	Sound file name, SAMPA								
<WORDS>	Marie	geeft	Kees	een	klap	op	zijn	kop	.
<PHONES>	ma: ri.	Ge:ft	ke:s	@n	klAp	Op	zEin	kOp	

(3) French

<INFO>	SAMPA							
<WORDS>	La	maison	s’	est	écroulée			.
<PHONES>	la	mE zo~	sE		te	kru	le	
<TRANS>	The house collapsed.							

3.4 Instructions

INFO

- Specify the sound file name and the phonetic alphabet used in the PHONES layer (IPA/SAMPA).
- If necessary add additional information.

WORDS

- Determine word boundaries and provide an orthographic transcription in the relevant interval.
- You can annotate significant pauses with the <P>-tag.

PHONES

- The transcription should be broad (phonemic). For example, in English there is no phonemic contrast between aspirated and non-aspirated stops, so it is not necessary to mark aspiration in the transcription.
- Obtain syllable boundaries and enter the appropriate phonetic symbols for each syllable either in IPA or in SAMPA. Though it is usually quite easy to decide on the number of syllables, the decision as to the exact location of their boundaries may be trickier. If there is consensus about syllabification, use the standards. Otherwise, use your own intuitions.
- Do not mark stress at this level; this should be done at the 'STRESS' level (see section 3.3).

4 Level II: Metrical structure

4.1 Objects of annotation

Lexical stresses are important to locate because they carry the pitch accents related to sentence accent. We use the term 'stress' to refer to the prominent syllable in a word, disregarding whether the syllable is realized as prominent or not, and the term 'accent' to refer to realization of stresses by means of a pitch accent. The phonetic and phonological properties of individual pitch accents are

the object of section 5. Here, we deal only with the abstract position of lexical stress and whether or not potential stresses are realized by means of pitch accents. Note that an accent may sometimes be realized on a syllable which does not carry lexical stress: e.g. ‘I said stalagMITE, not stalagTITE’. Some languages have no lexical stress, such as many tone languages, and investigators of these languages can omit the STRESS level; however there are languages which do not have lexical stress but which do have accents, such as French, in which case the ACCENT layer only should be used. Do not transcribe lexical tones at this level (see instead the Tones and Intonation level).

For languages with lexical stress, the STRESS layer should be ‘abstract’. It should consider words spoken in isolation, and not the phonetic realization in context, since this latter realization is subject to accent shifts, accent deletion and the like.

4.2 Tagset declaration

Table 3: Tagset declaration for metrical structure

tag	meaning	short description
1	Primary stress	Most prominent syllable of the language
2	Secondary stress	Second most prominent syllable of the language
3	Tertiary stress	Third most prominent syllable of the language

4.3 Instructions

STRESS

- If the object language has lexical stress, then annotate the most prominent syllables (or mora) of the language with 1.
- If the object language has lexical stress, then annotate the second most

prominent syllables (or mora) of the language with 2.

- Usually two layers of lexical stress are sufficient. In general, tertiary stress is not necessary. The third most prominent syllables (or mora) of the language can be annotated with 3 if you feel that it is important. All lower layers should be annotated with 3 as well.
- Unstressed syllables are not especially annotated.
- It is not necessary to annotate mono-syllabic function words.

ACCENT

- If the object language has lexical stress, annotate the stresses realized by means of a pitch accent with 1.
- If the object language does not have lexical stress but does have accents, annotate syllables realized with a pitch accent with 1.
- Unaccented syllables are not especially annotated.

(4) German, SAMPA

<WORDS>	Lena		verfolgt		den	Mann	mit	dem	Motorrad		
<PHONES>	le:	na	v6	fOlgt	d@n	man	mIt	d@m	mo	to:	Rat
<STRESS>	1	2		1		1				1	2
<ACCENT>	1					1				1	

5 Level III: Prosodic structure

5.1 Declaration

For the sake of information structure, it is crucial to annotate the higher levels of prosody which are relevant for information structure. In most cases, these are the Phonological Phrase (PHP) and the Intonation Phrase (IP). It is also obligatory to specify syllable boundaries (SYL). The other layers (Phonological Word, Foot, and Mora) are required only if the annotators are phonologically confident. Constituents larger than a sentence (intonation phrase) are not considered here.

5.2 Related standards

Terminological conventions differ in some respects. Phonological Words are sometimes called ‘Prosodic Words’. Phonological Phrases can be called ‘Accentual Phrases’. ‘Clitics Group’ is sometimes considered to be a necessary layer, but it is not part of the required annotations.

5.3 Tagset declaration

Table 4: Tagset declaration for metrical structure

Layer	Tags	Meaning	Short Description
IP	IP	Intonation Phrase	A phonological constituent roughly the size of a sentence
PHP	PP	Phonological Phrase	A phonological constituent roughly the size of a maximal projection
	aPP	Abstract PP	See below for a description
	rPP	Realized PP	
PW	PW	Prosodic Word	A phonological constituent roughly the size of a grammatical word
FT	F	Foot	A metric unit ideally consisting of two syllables, one of which is strong and the other weak (trochee or iamb)
MORA	M	Mora	A weight unit of the syllable (its nucleus and possibly the next consonant)

(5) English (adapted from Gussenhoven (2002: 271))

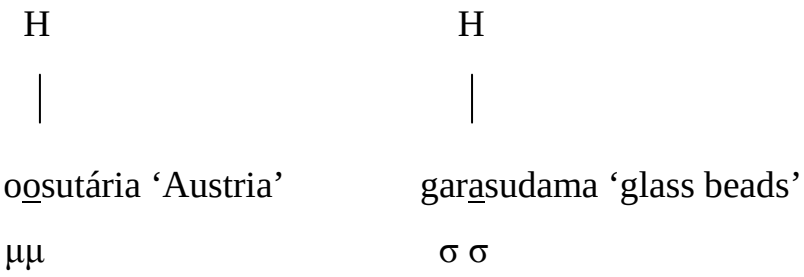
<WORDS>	Too	many		cooks	spoil	the	broth	.
<PHONES>	tu:	mE	ni	kUks	spOIl	D@	brOT	
<FT>	F	F		F	F		F	
<PW>	PW	PW		PW	PW	PW		
<PHP>	PP				PP	PP		
<IP>	IP							

(Note that, in this analysis, the word ‘the’ is not part of any foot.)

5.4 Instructions

5.4.1 Mora (MORA)

- Only some languages use the mora actively in the phonology and distinguish tones or accent placement in terms of moras. An example is Japanese in which the default high tone in unaccented words is on the second mora, which can be part of the first or the second syllable.



Syllable:

- Syllable boundaries should be annotated at the PHONES layer (see section 2.4).

Foot (FT):

- According to most researchers, trochees (strong - weak) and iambs (weak - strong) are the only universal metrical feet. Trochees may be based on moras or on syllables. This gives the following universal foot inventory (Hayes 1995):

a. Syllabic trochee: (canonical form: 'σσ)

b. Moraic trochee: (canonical form: 'μμ)

c. Iamb: (canonical form: σ'σ)

- If the metrical footing is described in the literature, use this standard, otherwise use your own.
- Feet are used for identification of lexical stress. It is often sufficient to locate primary (and secondary and tertiary) stress without any foot structure. If you do not feel confident, do not annotate this layer.

Phonological Word (PW):

- Prosodic Word is a constituent which is more or less isomorphic to the grammatical word, except for cliticization. In languages with lexical stress, one syllable has main prominence. PW are often domains for syllabification and can stand alone.
- The identification of this domain is not crucial for information structure. Use it only if you feel confident about it, or if it is important for the higher layers.

Phonological phrase (PHP):

- A phonological phrase is a domain for syntactic accents and for phrasing. Depending on the presence of accents, a PHP contains a main accent or an element which is more prominent than the other ones. A PHP is often delimited by boundaries, which can be tonal, segmental or durational.
- If you have very good knowledge or intuitions about this layer, you may distinguish between an abstract PHP and a realized PHP. An abstract PHP is defined by the syntactic structure.

Intonation phrase (IP):

- An intonation phrase is the domain of the tonal realization. In languages with lexical stress and/or sentence accent, it is the domain of the main

accent. In most cases, it follows the syntactic structure closely, and an IP is often isomorphic to a sentence. Embedded clauses usually form separate IPs, as do parenthetical clauses. A vocative or a tag (like ‘isn’t it’) also usually form separate phrases. Lists, elliptical constructions, inverted discontinuous constituents, may be in separate IPs as well as cleft constructions, topicalizations and extrapositions.

- In intonation and pitch accent languages always, and in tone languages most of the time, intonation phrases are delimited by a boundary tone, meaning that a word has a different intonation (falling or rising) when it is phrase-final than when it is phrase-medial.

6 Level IV: Tone and Intonation

6.1 Introduction

The intonation of a language can be transcribed using the ToBI framework. ToBI systems are available for English, German, Dutch, Greek, Japanese, Korean etc. See Jun (2005) for an overview of ToBi in different languages. We propose three layers of tonal transcription: Two of these layers (TONES and SURFACE) capture phonological descriptions of tones. If a ToBI system exists for the language, please use it. Otherwise follow the directions in the remainder of this document. See also Ladd (1996) and Gussenhoven (2004) for descriptions of intonation from a theoretical perspective. The third layer (PHONTONES) concerns a phonetic description in the style of the IviE system (see below).

6.2 Declaration

In tone languages, lexical items are associated with tones. The underlying tone of each syllable or word should be represented in the TONES layer. If words

associated with tones occur in combination, certain tones may change on the surface due to phonological process or because of constraints like the OCP (cf. 5.2.1.). In intonation languages, tonal processes such as tonal linking may also change the underlying tonal representation of pitch accents (cf. 5.2.2.). The output of these and similar processes is annotated in the SURFACE layer. If very little is known about a language one might at first describe surface pitch patterns around prominent syllables (in intonation languages) or tone-bearing units (in tone languages) in the PHONTONES layer. The phonetic layer may thus help capture surface variation which leads into phonological abstraction in further analysis.

6.2.1 Underlying tones (TONES)

The layer TONES comprises lexical tones (as in Chinese), lexical accents (as in Japanese or Swedish), intonational tones (as in English or German), and boundary tones. Tone bearing units (TBUs) can be syllables (most languages) or moras (Japanese for instance). Tonal layers should be indicated according to the following standards:

- use the labels H (high) and L (low); possibly also M (mid)
- alternatively, use tonal labels according to existing conventions, such as tone 1, 2, 3, 4 in Mandarin Chinese.

Please use the following conventions (see also ToBI conventions):

Table 5: Tagset declaration for metrical structure

H* / L*	high-tone / low-tone pitch accent
L*H	rising pitch accent
H*L	falling pitch accent
L*HL	rise-fall pitch accent (other combinations possible)
!H*	downstepped high-tone pitch accent

H- / L-	high/low boundary tone associated with Phonological Phrase
H% / L%	high/low boundary tone associated with Intonational Phrase right edge
%H / %L	high/low boundary tone associated with Intonational Phrase left edge
LH%	rising boundary tone (other combinations possible)
*?	uncertain if this is a pitch accent
X*	this is a pitch accent but I don't know which one
?- / ?%	uncertain if this is a boundary tone
X- / X%	this is a boundary tone but I don't know which one

Notes:

- A pitch accent is a tone that is associated with the stressed syllable of a word. A boundary tone marks the beginning or end of a prosodic domain, such as a Phonological Phrase or an Intonational Phrase.
- The transcription of both types of tones should be phonological rather than phonetic, thus do not include more detail than needed to make categorical distinctions.

(6) English (from Gussenhoven 2002: 271)

<WORDS>	Too	many		cooks	spoil	the	broth	.
<PHONES>	tu:	mE	ni	kUks	spOIl	D@	brOT	
<PHP>	PP				PP	PP		
<IP>	IP				IP			
<TONES>	H*			L*H	H*		H*L	L%

6.2.2 Surface tones (SURFACE)

The same conventions apply on the SURFACE layer as for the TONES layer (see above). The significant difference between the two layers is that the SURFACE layer gives information about tones that have undergone phonological adjustments, such as tone sandhi and tonal linking. If in doubt, assume that these two layers are identical.

(7) Mandarin (The underlying tone 1 of *yi* surfaces as tone 4, and the underlying tone 3 of *suo* surfaces as tone 2; the other tones have the same underlying and surface form)

<ORTHOGR>	一本□惹怒了所有的人.										
<TRANS>	A book annoyed everyone.										
<GLOSS>	one	CL	book	annoy		ASP	everyone				
<WORDS>	yi	ben	shu	re-nü		le	suo-you-de-ren				.
<PHONES>	i	pən	ʂu	rə	ny	lə	suə	iəu	tə	rən	.
<TONES>	1	3	1	3	4	0	3	3	0	2	
<SURFACE>	4						2				

6.2.3 Phonetic description (PHONTONES)

The PHONTONES layer provides space to label phonetic surface variation. This optional layer allows for a first analysis-free inspection of the data. If no intonational information of a language is available, this layer may be used to build up generalisations of surface tonal patterns. The systematic comparison of utterances in the same context provides insights in the phonological structure of a language. An additional effect is the outcome of phonology-phonetic mapping rules, and how they differ between languages (see IViE for the use of this layer). This layer comprises the pitch contour around prominent syllables, (Implementation domain: preaccented, accented plus following syllables up to the next accented one).

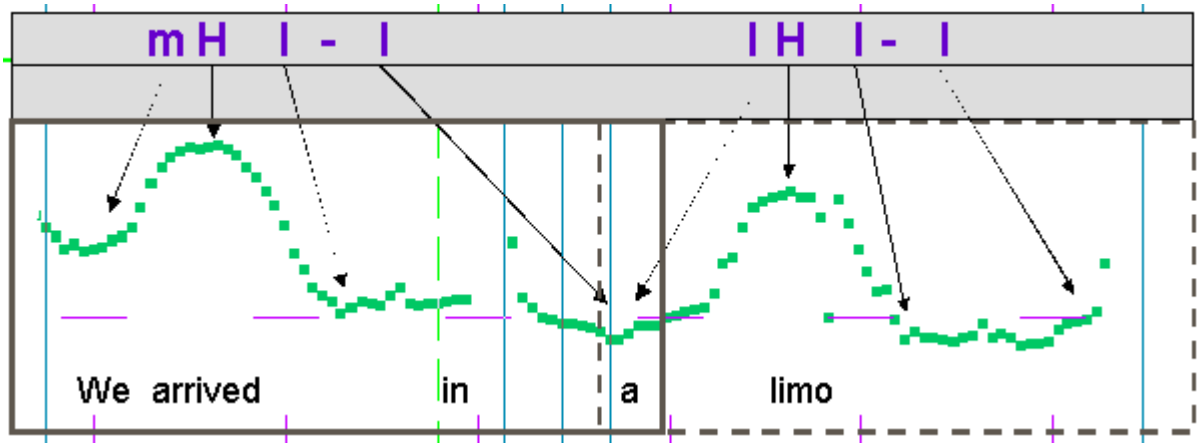
Table 6: Available labels from IViE

H / h	high pitch	Upper case is used for the accented syllable, lower case for preceding and following unaccented syllables.
L / l	low pitch	
M /m	mid pitch (optional)	
i	end of an implementation domain	
-	interpolation	

An implementation domain may contain maximally four labels (preaccentual, accented, postaccentual and final domain syllable), minimally two (accented and pre- *or* postaccentual syllable).

Figure 1: Illustrative examples of an implementation domain

(from http://www.phon.ox.ac.uk/~esther/ivyweb/guide.html)



7 Guidelines for phonological analysis

7.1 Checklist

To conduct a phonological analysis, you should try to answer the following questions:

- Does the language have lexical tones or pitch accents? A lexical tone is a specific melody (like a high tone, a low tone, a falling tone) associated with syllables or moras of words and contributing to the lexical meaning of the word. A pitch accent is found in some languages (for instance Japanese and Swedish). It has the same function as a lexical tone, but pitch-accent languages typically have only one pitch accent, whereas tone languages typically have several tones.
- What kind of tonal entities exist in the language? (Section 7.2)
- Are focus and topic expressed by intonation (i.e. an accented syllable)?

(Section 7.3)

- Is a narrow focus accent possible, even on an unstressed syllable (an English example: “I didn’t say coLA I said coDA”)? (Section 7.3)
- How is a given or ‘backgrounded’ element realized in the language? (Section 7.4)
- What is the relation between sentence type and intonation? In particular, can you say that a question is generally expressed with a rising intonation, and a declarative sentence with a falling intonation? (Section 7.5)

7.2 Tonal entities

7.2.1 Lexical tones and lexical pitch accents

Lexical tones and lexical pitch accents change the meaning of a word, and they are thus specified in the lexicon, contrary to intonational tones. A language has lexical tone if most syllables have a lexical tone associated with them. Examples include Chinese and many African and Amerindian languages (see the Chinese example above). A language has lexical pitch accent if even one syllable of a word (or phrase) has a tone associated with it. Examples include Japanese and Swedish.

Example Swedish:

(8a) Accent 1 word (HL*), from Bruce (1977), tonal labels adapted

<WORDS>	Man	vill	anamma		nåra	längre		[1]nummer	
<PHONES>	$	$	$	$	$	$	$	$	$	$	$
<TONES>				HL*			HL*		HL* H-	L%
<STRESS>				2			2		1	
<TRANS>	“One wants to receive some longer numbers”									

(8b) Accent 2 word (H*L), from Bruce (1977), tonal labels adapted

<WORDS>	Man	vill	lämna		nåra		långa		^{2}nunnor	
<PHONES>	$	$	$	$	$	$	$	$	$	$
<TONES>			H*L				H*L		H*L	H- L%
<STRESS>			2				2		1	
<TRANS>	“One wants to leave some long nuns.”									

Example: Mandarin Chinese

- lexical tone (1, 2, 3, 4); A single word (like *ma*) has completely different meanings according to the lexical tone associated with it.
- no pitch accent
- L and H boundary tones
- raised, expanded and compressed pitch range

Example: Japanese

- lexical pitch accent
- only one type of pitch accent, H*+L
- L and H boundary tones associated with the Accentual Phrase (AP)
- H, LH, and HL boundary tones associated with Intonational Phrase (IP)

H*L H*(L)

| | |

a. hána ‘a name’ b. haná ‘flower c. hana ‘nose

7.2.2 Intonational tones

Intonational tones are assigned at a ‘post-lexical’ level. Current theories of intonational phonology distinguish two types of intonational tones: (post-lexical) pitch accents and boundary tones. Both of these may be analyzed as sequences

of simple tones (H and L). A pitch accent is associated with the stressed syllable of a word, and a boundary tone with the beginning or end of a prosodic domain. Not all languages have pitch accents, but most languages appear to have boundary tones. Some (tone) languages use pitch range features instead of pitch accents.

There are typically many syllables that do not have a tone associated with them, and the pitch between two tones is filled in by interpolation, e.g. between a high tone and a low tone there will be a gradual decline in pitch, which does not need to be specified in the phonological transcription.

Example: German

- pitch accents[16]: L*, H*, LH*, L*H, H!H*, and H*L (or HL*)
- boundary tones: L and H

(9) (Example adapted from Féry 1993)

<WORDS>	Lena		verfolgt		den	Mann	mit	dem	Motorrad		
<PHONES>	le:	na	v6	fOlgt	d@n	man	mIt	d@m	mo	to:	Rat
<IP>	IP										
<TONES>						H*L				H*L	
<SURFACE>						H*				H*L	

7.3 Information structure

To study the relation between intonation and information structural categories such as topic or focus you might take a sentence and have it read with several different focus structures.

7.3.1 Focus structure

To establish whether intonation plays a role for focus assignment in a language, try sentences such as the following:

[16] A star refers to a pitch accent, and !H is downstep, i.e. the H tone is realized at a lower pitch than the preceding H tone.

a. Why is he so still? [The child is sleeping]$_{FOCUS}$

b. Who is sleeping? [The child]$_{FOCUS}$ is sleeping/It is the child….

c. What does the child do? The child [is sleeping]$_{FOCUS}$

d. What is happening with the mother and the child right now?/ What are the mother and child doing right now?

[The child]$_{TOPIC}$ [is sleeping]$_{FOCUS}$ but…

First check whether intonation is involved at all, or whether the language you are investigating only uses morphological markers or syntax to realize different focus structures. If intonation is involved, check if there is any difference between the intonation of these sentences. Is focus marked by an accent or phrasing? Try to answer the following questions:

- Does the topic in (d) get an accent? Is this accent different from the focus accent?
- Does the topic form a prosodic domain of its own (e.g. PHP or IP)?
- Does the language allow non-final focus as in (b)?
- What intonation pattern does the all-new sentence in (a) have?

See section 7.4 to answer these questions.

7.3.2 Narrow and contrastive focus

You should also check if narrow focus is marked by intonation. That is, is it possible to have a pitch accent (or expanded pitch in a tone language) on a non-final word in an XP?

In English, narrow focus can be marked by a pitch accent in all cases:

(10) (a) Question: Who loves Mary?
Answer: JOHN loves Mary.

(b) Question: How many apples do you need for your cake?
Answer: FOUR apples.

In certain cases, an accent signaling narrow focus is ambiguous with an accent signaling wide focus. This happens when the word which is narrowly focused is by accident identical to the word bearing default accent in a wide-focused sentences:

(11) Question: Who does Mary love?
Answer: Mary loves JOHN.

Narrow focus cannot always be realized intonationally, because in some languages, pitch accent is restricted to specific positions. In Manado Malay, for instance, an XP must have a final accent. For this reason, if focus is non-final in its phrase, it is not possible to mark it with a pitch accent. Compare the following sentences in English and in Manado Malay (accents are marked in upper case):

(12) (a) English:
How many kilos? $[\text{THREE}]_{\text{focus}}$ kilos.
(b) Manado Malay
Brapa kilo? $[\text{tiga}]_{\text{focus}}$ KILO.

It is important to distinguish narrow and contrastive focus. In a contrastive focus, another element is explicitly denied or contrasted.

(13) (a) Contrastive focus:
I don’t want a banana, I want an apple.
(b) Question: Is it John that Mary loves?
Answer: No, Mary loves BILL.

(14) Contrast
A RED flower is more attractive for birds than a WHITE one.

7.4 Deaccentuation, givenness, backgrounding

Sentences also contain parts which are not prominent. This happens because they are repeated from the sentence before, or simply because they are in the neighborhood of a focused constituent. In the following dialogue, the first part of the answer is backgrounded because it just repeats, albeit with different words, something which has been asked in the preceding question.

(15) Question: What do you want for dinner?

Answer: [I would like to eat] $_{\text{background}}$ [FISH & CHIPS] $_{\text{focus}}$

In the following example, 'children' may be new, and would thus deserve an accent, but because of the proximity of a contrastive accent, it is prone to deaccentuation.

(16) John has THREE and Bill has FOUR children.

7.5 Sentence type

The distinction between statement, yes-no question, and wh-question is typically expressed by intonation. In some languages, a statement and a yes-no question are distinguished only by intonation. If this is not the case, then try to make sentences of these types that are as similar as possible. In both cases see what the differences in intonation are. Typical differences include:

- the type of final boundary tone (statements often have L, and yes-no questions H);
- a different kind of pitch accent;
- the position of the pitch accent (e.g. does a wh-word get an accent?);
- the overall pitch level may be higher.

7.6 Practical considerations

If possible, the target words in your sentences (i.e. the words that you expect to get an accent or boundary tone) should not be too short. It is best if they have:

- nasals or voiced obstruents rather than voiceless sounds, since there is no pitch without voicing;
- penultimate or antepenultimate stress, so that no pitch accent and final boundary tone will occur in the same syllable.

8 References

Bruce, Gösta. 1977. *Swedish Word Accents in Sentence Perspective*. Lund: Gleerups.

Grabe, Ersther and Nolan, Francis. 1997. *Intonational Variation in the British Isles*. <http://www.phon.ox.ac.uk/~esther/ivyweb/>

Gussenhoven, C. 2002. Phonology of intonation. *Glot International* 6: 271–284.

Gussenhoven, C. 2004. *The phonology of tone and intonation*. Cambridge University Press.

Jun, S.A. (ed.) 2005. *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press.

Ladd, D.R. 1996. *Intonational phonology*. Cambridge University Press.

Selting, M. et al. 1998. Gesprächsanalytisches Transkriptionssystem. *Linguistische Berichte* 173, S. 91-122.

Morphology

Joanna Blaszczak[1], *Stefanie Dipper*[1], *Gisbert Fanselow*[1], *Shinishiro Ishihara*[1], *Svetlana Petrova*[2], *Stavros Skopeteas*[1], *Thomas Weskott*[1], *Malte Zimmermann*[1]

University of Potsdam ([1]) and Humboldt University Berlin ([2])

The guidelines for morphological annotation contain the layers that are necessary for understanding the structure of the words in the object language: morphological segmentation, glossing, and annotation of part-of-speech.

1 Preliminaries

The guidelines for these layers follow existing recommendations in language typology and norms for the creation of language corpora. The glossing guidelines belong to the paradigm of guidelines that has arisen on the basis of *Eurotyp* (König et al. 1993), being more closely related to the conventions of the *Leipzig Glossing Rules* (see Bickel et al. 2002). The guidelines for morphological categories combine the practices recommended in *Eurotyp* with norms that have been established for the morphological annotation of corpora such as *EAGLES* (Leech & Wilson 1996) and *STTS* (Schiller et al. 1999).

Interdisciplinary Studies on Information Structure 07 (2007): 55–94
Dipper, S., M. Götze, and S. Skopeteas (eds.):
Information Structure in Cross-Linguistic Corpora

2 Layer Declaration

Table 2: Layers

Layer	Abbreviation
morphemic segmentation	MORPH
morpheme-to-morpheme translation	GLOSS
part of speech	POS

3 Layer I: Morphemic Segmentation (MORPH)

3.1 Introduction

The layer of morphemic segmentation (sometimes referred to as morphemic transcription) indicates morpheme boundaries. It contains a copy of the original text and makes use of special characters like hyphens, dots, etc. to segment words into morphemes.

Instructions for the use of this layer:

(1) English

<WORDS>	The	wolf	jumps	out	of	the	building.
<MORPH>			jump-s				

The proposed guidelines are based on *Leipzig Glossing Rules* (see Bickel et al. 2002).

3.2 Tagset declaration

Table 3: Tagset declaration for morphemic segmentations

tag	meaning	see in:
<new cell>	word boundary	§3.3.1
-	morpheme boundary	§3.3.2
=	clitic boundary	§3.3.3
_	union of sublexical components	§0
0	zero affix	§3.3.6

3.3 Instructions

3.3.1 Word boundaries

Words are given in separate cells in Exmaralda (otherwise separated through spaces).

(2) English

<WORDS>	the	children	work
<MORPH>	the	children	work

Instructions for the identification of word boundaries:

- If the object language has an orthographical representation that indicates word boundaries, then annotate the word boundaries indicated in the local orthography.
- If the orthographical representation in the object language indicates sublexical units (usually syllables) instead of words, then see §0.

3.3.2 Morpheme boundaries

Morphemes are separated by a hyphen:

(3) English

<WORDS>	Peter	works
<MORPH>	Peter	work-s

Inflection

- If the morpheme boundaries in the object language are transparent, then they should be indicated in the morphemic transcription. This holds especially for agglutinative languages, but also for morphemes that may be easily distinguished in fusional languages.

(4) English

<WORDS>	Peter	works
<MORPH>	Peter	work-s

- If the morpheme boundaries in the object language are not transparent, then do not indicate boundaries in cases where it is not feasible to establish some uncontroversial conventions. This holds especially for fusional languages. In the morphemic translation, these cases must be treated as shown in §4.4.3.

(5) English

<WORDS>	children
<MORPH>	children

(6) German

<WORDS>	entbrannt
<MORPH>	entbrannt
<GLOSS>	conflagrant

Word formation

- If the stems of a compound can be easily separated and the semantics of the compound can be compositionally derived by the unification of the

semantics of the individual roots, then the analytical representation is preferred. Note that in contrast to some other current practices, the stems contained in the compound are separated by a hyphen (not by a plus sign):

(7) German

<WORDS>	Bürgersteig
<MORPH>	Bürger-steig
<GLOSS>	citizen-path

(8) Japanese

<WORDS>	gengogaku
<MORPH>	gengo-gaku
<GLOSS>	language-study

- Compositional morphemes are also separated by a hyphen and are indicated as such in the morphemic translation:

(9) German

<WORDS>	Legehenne
<MORPH>	Leg-e-henne
<GLOSS>	lay-0-hen(F)

- If the internal structure of compounds and derivatives displays difficulties in the object language (in terms of identification of the morpheme boundaries or in terms of semantic compositionality), then do not indicate the internal structure of the word.

(10) German

<WORDS>	Erdbeere
<MORPH>	Erdbeere
<GLOSS>	strawberry

3.3.3 Clitic boundaries

Clitic boundaries are indicated by an equal sign. They may be tokenized with their phonological target as in example (18). In other cases, it might be preferrable to tokenize the clitic separately, e.g. when the orthographical transcription in the <WORDS> layer requires separate tokens for the clitic and its target (see example (19) below):

(11) German

<WORDS>	wie	geht’s
<MORPH>	wie	geht=s
<GLOSS>	how	go:3.SG=it

Instructions for the identification of clitics: Clitics are phonologically weak (unstressed) elements that need a host in the form of a phonologically strong (stressed) element on which they (mostly in their reduced form) cliticize, e.g., *kommste* (= *kommst du*), *s’Fenster* (= *das Fenster*)

- For elements like *zum*, *am*, *ins*, *vom* (German), *au*, *des*, *aux* (French), see §4.4.4.
- In languages which provide an opposition between clitic and emphatic (personal, relative, etc.) pronouns or auxiliaries, clitics are identified through the use of the clitic boundary “=”:

(12) Greek

<WORDS>	to	thélo
<MORPH>	to=	thél-o
<GLOSS>	3.SG=	want-1.SG

(13) Greek

<WORDS>	aftó	thélo
<MORPH>	aftó	thél-o
<GLOSS>	3.SG	want-1.SG

(14) English

<WORDS>	he	‘s	leaving
<MORPH>	he	=s	leav-ing

(15) English

<WORDS>	he	is	leaving
<MORPH>	he	is	leav-ing

3.3.4 Union of sublexical components

This rule applies especially in languages in which blank spaces in the orthography do not always indicate word boundaries. Sublexical components of one word are put in one cell and are connected by an underscore:

(16) Vietnamese

<WORDS>	tiểu thuyê↙t
<MORPH>	tiểu_thuyê↙t
<GLOSS>	roman

The original form is one orthographical form in Vietnamese. Blank spaces in Vietnamese are orthographically ambiguous: they denote both word boundaries and syllable boundaries. Many words contain more than one syllable, which may be assigned only a common translation (a syllable-by-syllable translation is not possible). In morphemic segmentation, syllable boundary is represented by blank space.

3.3.5 Special characters

Special characters, i.e. non-alpha-numerical characters, such as -, %, ‘, “,), etc., that are used in orthographic representations (that may be used in WORDS) are left out at the layer of morphemic segmentation, see examples (17)-(18).

(17) German

<WORDS>	das	“Pünktchen”
<MORPH>	das	Pünkt-chen
<GLOSS>	DEF:N.SG.NOM	point-DIM

Note that the hyphen has different meaning in the two layers of example (18): at the layer WORD it is an orthographic symbol, and at the layer MORPH it encodes morpheme boundaries.

(18) German

<WORDS>	die	“Pünktchen”-Partei
<MORPH>	die	Pünkt-chen-Partei
<GLOSS>	DEF:F.SG.NOM	point-DIM-party

3.3.6 Zero morphemes

The indication of zero morphemes is sometimes part of the morphemic segmentation. Since a morphemic analysis in terms of zero morphemes is not theory neutral, we recommend avoiding the use of zeroes in the database. If a project needs this kind of information for its data, the standard symbol ‘0’ is recommended (note that ‘0’ is also used in glossing, compare (57)).

(19) German

<WORDS>	die	Lehrer
<MORPH>	die	Lehrer-0
<GLOSS>	DEF:NOM.PL	teacher-PL

4 Layer II: Morphemic Translation (GLOSS)

4.1 Introduction

The layer of morphemic translation identifies the lexical meaning or grammatical function of individual morphemes as they are segmented at the layer of morphemic transcription. This section includes:

- rules for morpheme-to-morpheme translation;
- the list of tags for the recommended glosses.

4.2 Related standards

The proposed guidelines are based on *Leipzig Glossing Rules* (see Bickel et al. 2002) and Eurotyp (see König et al. 1993). In particular, a basic list of abbreviations is adopted from LGR – and if not available in this standard from Eurotyp (see König et al. 1993); further tags for terms that are not available in these standards and are needed for our corpus have been introduced in our document.

4.3 Tagset declaration

The symbols used at the MORPH layer are replicated at the GLOSS layer. In addition to these symbols (see §3.2), some symbols are only used in the GLOSS:

Table 4: Conventions for morphemic translation

tag	meaning	see in:
x:y	x and y are different morphemes with non-segmentable boundaries	§4.4.4; 4.4.5
x.y	x and y are semantic components of the same morpheme	§4.4.4; 4.4.5
x_*n*	all x_*n* are parts of the same discontinuous morpheme	§4.4.3
x/y	x and y are alternating meanings/meaning components	§4.4.6
{x}	x is a feature not realized in this context	§4.4.6
[x]	x is non-overtly encoded	§4.4.6; 0
XXX	grammatical meaning	§4.4.8

4.4 Instructions

4.4.1 Isomorphism between GLOSS and MORPH

Symbols introduced at the layer of morphemic segmentation for the indication of boundaries (§3.2) are also used obligatorily in morpheme translations in a one-to-one relation. For exceptions to the general principle of isomorphism see §4.4.2-0.

- word boundaries

(20) German

<WORDS>	heute	morgen
<MORPH>	heute	morgen
<GLOSS>	today	morning

- morpheme boundaries

(20) English

<WORDS>	works
<MORPH>	work-s
<GLOSS>	work-3.SG

- clitic boundaries

(21) German

<WORDS>	wie	geht’s
<MORPH>	wie	geht=s
<GLOSS>	how	go:3.SG=3.SG.NOM

4.4.2 Non-Isomorphism: Sublexical components

In case the morphemic transcription contains more than one sublexical components (indicated by an underscore; see §0), they correspond to one unit at the GLOSS layer.

(22) Vietnamese

<WORDS>	tiểu thuyê↙t
<MORPH>	tiểu_thuyê↙t
<GLOSS>	roman

4.4.3 Non-Isomorphism: Discontinuity

Discontinuous morphemes are indicated by repeating the gloss in each part of the morpheme. The parts of the discontinuous morpheme are indicated through the index ‘*_n*’. In infixation, the discontinuous morpheme is the root:

(23) Tagalog

<WORDS>	bili
<MORPH>	bili
<GLOSS>	buy

<WORDS>	bumili
<MORPH>	b-um-ili
<GLOSS>	buy_1-A.FOC-_1

In circumfixation, the discontinuous morpheme is the affix:

(24) Tuwali Ifugao, Philippines

<WORDS>	baddang
<MORPH>	baddang
<GLOSS>	help

<WORDS>	kabaddangan
<MORPH>	ka-baddang-an
<GLOSS>	NMLZ_1-help-_1

The same logic applies to cases like the particle verbs in German, where the particle can be separated from the verb and can occur like an independent word:

(25) German

<WORDS>	ich	fange	mit	dem	Studium	an
<MORPH>	ich	fange	mit	dem	Studium	an
<GLOSS>	1.SG	start:1.SG_1	with_1	DEF:DAT.N	study[DAT.N]	_1

<WORDS>	weil	ich	mit	dem	Studium	anfange
<MORPH>	weil	ich	mit	dem	Studium	anfange
<GLOSS>	because	1.SG	with	DEF:DAT.N	study[DAT.N]	start:1.SG

4.4.4 Non-Isomorphism: Non-indicated boundaries

If the original form contains different morphemes that are not segmented (at the MORPH layer), then a colon is used in the gloss:

(26) German

<WORDS>	geht
<MORPH>	geht
<GLOSS>	go:3.SG

Special instructions for non-indicated boundaries:

- Morpheme boundaries that may not be easily identified in a theory neutral way, are not indicated (see §3.3.2):

(27) German

<WORDS>	ging
<MORPH>	ging
<GLOSS>	go:PAST:1.SG

- In the case of portmanteau morphemes (i.e. morphemes that fuse more than one grammatical functions), it usually makes no sense to indicate boundaries in the morphemic transcription; however, the different grammatical functions can be read off the GLOSS layer:

(28) French

<WORDS>	au
<MORPH>	au
<GLOSS>	to.DEF.SG.M

4.4.5 Non-Isomorphism: Complex glosses

If the morphemic translation contains more than one gloss, the glosses are separated by periods:

(29) Polish

<WORDS>	ciastko
<MORPH>	ciastko
<GLOSS>	cake:SG.NOM.N

Special instructions for complex glosses:

- Amalgamated grammatical information in fusional languages is translated through complex glosses:

(30) Polish

<WORDS>	ciastko
<MORPH>	ciastko
<GLOSS>	cake:SG.NOM.N

- Person and number combinations are treated as complex glosses:

(31) German

<WORDS>	geht
<MORPH>	geht
<GLOSS>	go:3.SG

- Lexical information that may not be translated by a single element in the translation language is treated as a complex gloss:

(32) Hawaian

<WORDS>	ulua
<MORPH>	ulua
<GLOSS>	old.man

- In complex glosses conveying grammatical information the following orders are used:

  NOMINAL INFLECTION

  {gender}.{number}.{case} (for nouns, adjectives, and determiners)

  The order of these categories corresponds to the cross-linguistically preferred order for the realization of the corresponding morphemes.

(33) Polish

<WORDS>	ciastko
<MORPH>	ciastko
<GLOSS>	cake:N.SG.NOM

(34) Spanish

<WORDS>	mojigata
<MORPH>	mojigata
<GLOSS>	prude:F.SG.NOM

(35) Spanish

<WORDS>	una
<MORPH>	una
<GLOSS>	INDEF:F.SG.NOM

PRONOMINAL INFLECTION

{person}.{number}.{gender}.{case}

The idea of this order is to start the GLOSS with the information which identifies the paradigms as they are commonly presented in grammars, e.g. "2nd singular", "3rd singular masculine"; the relational information, i.e. case, comes at the end of the GLOSS.

(36) German

<WORDS>	du
<MORPH>	du
<GLOSS>	2.SG.NOM

(37) German

<WORDS>	ihm
<MORPH>	ihm
<GLOSS>	3.SG.M.DAT

(38) German

<WORDS>	wir
<MORPH>	wir
<GLOSS>	1.PL.NOM

- Elements denoting person/number are decomposed into their semantic

features if they are personal pronouns (i.e., if they belong to a syntactically identifiable paradigm that structures person/number oppositions in the object language):

(39) German

<WORDS>	sie
<MORPH>	sie
<GLOSS>	3.SG.NOM.F

<WORDS>	mir
<MORPH>	mir
<GLOSS>	1.SG.DAT

<WORDS>	wir
<MORPH>	wir
<GLOSS>	1.PL.NOM

- If the categorial status of these elements is not different from simple nouns, then their meaning is rendered by the English translation:

(40) Japanese

<WORDS>	kanojo
<MORPH>	kanojo
<GLOSS>	she

VERB INFLECTION

{aspect}.{voice}.{finiteness}.{tense}.{mood}.{person}.{gender}.{number}

(41) Ancient Greek

<WORDS>	lusaímēn
<MORPH>	lusaímēn
<GLOSS>	unbind:PFV.MID.PST.OPT.1.SG

The conventions for the order of morphological categories only hold for complex morpheme glosses, which contain more than one piece of grammatical information. Otherwise, the GLOSS corresponds to the actual order of morphemes.

(42) Turkish

<WORDS>	bilmiyorum
<MORPH>	bil-m-iyor-um
<GLOSS>	know-NEG-PROG-1.SG

4.4.6 Non-isomorphism: Alternative meanings

If a given grammatical or lexical morpheme has different meanings (that are activated in different contexts; in cases of either polysemy or homonymy), we recommend that only the context-relevant meaning is given:

(43) German

<WORDS>	vom	Jahr
<MORPH>	vom	Jahr
<GLOSS>	from:DEF.SG.DAT.N	year[DAT.SG]

(44) German

<WORDS>	das	Band
<MORPH>	das	Band
<GLOSS>	DEF:N.SG.NOM	tape[NOM.SG]

<WORDS>	der	Band
<MORPH>	der	Band
<GLOSS>	DEF:M.SG.NOM	volume[NOM.SG]

If in particular parts of the corpus you wish to indicate the ambiguity of particular morphemes which is resolved in syntactic context, then you may set the further alternatives in curly brackets:

(45) German

<WORDS>	vom	Jahr
<MORPH>	vom	Jahr
<GLOSS>	from:DEF.SG.DAT.N	year[DAT]{/NOM/ACC}

(46) German

<WORDS>	das	Band
<MORPH>	das	Band
<GLOSS>	DEF:N.SG.NOM	tape[DAT]{/volume[DAT]}

Complex examples of homonymy of case morphemes:

(47) Greek

<WORDS>	kaló
<MORPH>	kaló
<GLOSS>	good{N.{NOM/ACC}.SG/M.ACC.SG}

4.4.7 Non-isomorphism: Non-overtly encoded meaning

The German word *Frau* ‘woman’ consists of only one lexical morpheme, but it also contains information about grammatical number. Thus, the glossing:

(48) German

<WORDS>	Frau
<MORPH>	Frau
<GLOSS>	woman

is incomplete, because the word *Frau* ‘woman’ in contrast to *Frauen* ‘women’ also includes the information ‘singular’. If non-overtly encoded information should be stored, use square brackets:

(49) German

<WORDS>	Frau
<MORPH>	Frau
<GLOSS>	woman[SG]

Instructions for the annotation of non-overtly encoded information:

- If the non-overtly encoded category is the unmarked category, then our recommendation is to not indicate it in the gloss. The following rules may be postulated as default:

(50) Lack of voice in the gloss for a verb implies “active”.
Lack of number in the gloss for a noun implies “singular”.
Lack of tense in the gloss for a verb implies “present”.
Lack of case in the gloss for a noun implies “absolutive” in an ergative system.

These rules are language-specific: Lack of number morpheme indicates ‘singular’ in some languages, whereas in other languages it shows ‘general number’, lack of tense/aspect morpheme indicates ‘present’ in some languages, whereas in other languages it indicates ‘imperfective’, lack of case morpheme indicates absolutive in some languages, in some languages accusative, in some languages nominative, etc. That means the rules under (50) should be respectively postulated for every language.

- If a category which is treated cross-linguistically as unmarked is encoded through paradigmatic opposition and not through the lack of a morpheme, then this category is given in the gloss:

(51) Modern Greek

<WORDS>	neró
<MORPH>	neró
<GLOSS>	water:SG.NOM.N

<WORDS>	near
<MORPH>	near
<GLOSS>	water:PL.NOM.N

(52) Modern Greek

<WORDS>	gráfo
<MORPH>	gráfo
<GLOSS>	write:ACT.PRS.IND.1.SG

4.4.8 Tags

Table 4: Tags for glosses

tag	term
0	Element without semantic content or syntactic function
1	First person
2	Second person
3	Third person
A	Agent-like argument of canonical transitive verb
ABL	Ablative
ABS	Absolutive
ACC	Accusative
ALL	Allative
ANTIP	Antipassive
APPL	Applicative
ART	Article
BEN	Benefactive
CAUS	Causative
CLF	Classifier
COMPR	Comparative

tag	term
COM	Comitative
COMP	Complementizer
COMPL	Completive
COND	Conditional
COP	Copula
DAT	Dative
DECL	Declarative
DEF	Definite
DEM	Demonstrative
DIM	Diminutive
DIREV	Direct evidential marker
DIST	Distal (long distance from deictic center)
DISTR	Distributive
DU	Dual
DUR	Durative
ERG	Ergative
EXCL	Exclusive
EXPEV	Evidential marker for personal experience
F	Feminine
FILL	Break filler
FOC	Focus
FUT	Future
GEN	Genitive
HAB	Habitual
IMP	Imperative
INCL	Inclusive

tag	term
IND	Indicative
INDF	Indefinite
INF	Infinitive
INS	Instrumental
INTR	Intransitivizer
IPFV	Imperfective
IRR	Irrealis
ITER	Iterative
LOC	Locative
M	Masculine
MED	Medial (medial distance from deictic center)
MID	Middle (voice which excludes passive voice)
N	Neuter
NEG	Negative
NMLZ	Nominalizer
NOM	Nominative
NON	Negativelly defined categories
OBJ	Object
OBL	Oblique
P	Patient-like argument of canonical transitive verb
PASS	Passive
PFV	Perfective
PL	Plural
POSS	Possessive
POT	Potential
PRF	Perfect

tag	term
PRS	Present
PROG	Progressive
PROH	Prohibitive
PROX	Proximal (short distance from deictic center)
PST	Past
PTCP	Participle
PURP	Purposive
Q	Question particle/marker
QUOT	Quotative
RECP	Reciprocal
REFL	Reflexive
REL	Relative
REP	Reportative evidential marker
RES	Resultative
S	Single argument of canonical intransitive verb
SBJ	Subject
SBJV	Subjunctive
SG	Singular
SUPERL	Superlative
TOP	Topic
TR	Transitivizer

4.4.9 Special instructions

- Negative defined categories may be rendered with the abbreviation NON, e.g. NON.SG non-singular, NON.FUT non-future).

(53) Dyirbal

<WORDS>	balgan
<MORPH>	balgan
<GLOSS>	hit.NON.FUT

- It is recommended to use negatively defined categories as non-compositional categories of particular languages, and to use the negator compositionally as in the following example:

(54) English

<WORDS>	drink
<MORPH>	drink
<GLOSS>	drink.NON(3.SG)

- This tag is only used if the language possesses a category, which is negatively defined. Negatively defined terms are not used for the indication of polysemy. Thus:

(55) Modern Greek

<WORDS>	neró
<MORPH>	neró
<GLOSS>	water:SG.{NOM/ACC}

may not be rendered as in (56):

(56) Modern Greek

<WORDS>	neró
<MORPH>	neró
<GLOSS>	water:NON.PL.NON.GEN

- The tag ‘0’ is used for elements that lack semantic content. Note that the layer “morphemic translation (GLOSS)” contains the meaning or syntactic function of the elements of the layer “morphemic segmentation”. Elements that do not have such a function are rendered as ‘0’s. E.g. in

French questions, there is a liaison particule as in *que se passe-t-il?*. The *t* in this example has no semantic value, it is only there as liaison between a vowel ending verb and a vowel initial pronoun. The gloss of this element looks as follows:

(57) French

<WORDS>	que	se	passe-t-il
<MORPH>	que	se	passe-t-il
<GLOSS>	what	REFL.3.SG	happen:3.SG-0-3.SG.M

- The use of lexical verbs as auxiliaries for the formation of inflectional forms is not indicated in gloss. The gloss contains the lexical meaning of the verb. The special use of the verb in this case is indicated at the POS layer.

(58) French

<WORDS>	ai	aimé
<MORPH>	ai	aimé
<GLOSS>	have:1.SG	love:PTCP.PRF
<POS>	VAUX	VLEX

- Complex verbal aspects like ‘aorist’ should be decomposed, e.g. Modern Greek aorist is glossed as ‘PFV.PAST’ in indicative mood and as ‘PFV’ in non-indicative moods.

(59) Modern Greek

<WORDS>	fáe
<MORPH>	fáe
<GLOSS>	eat:IMPR.PFV.2.SG
<TRANS>	Eat!

(60) Modern Greek

<WORDS>	éfaje
<MORPH>	éfaj-e
<GLOSS>	eat:PFV.PAST-3.SG
<TRANS>	he/she/it has eaten

- Break fillers are elements like “hmmm…”, “äh…”, etc. These elements are glossed as ‘FILL’.

(61) German

<WORDS>	ich	gehe	...hmm...	ins	Kino	.
<MORPH>	ich	gehe	hmm	in=s	Kino	
<GLOSS>	1.SG	go:1.SG	FILL	in:DEF:ACC.SG.N	cinema [ACC.SG.N]	
<TRANS>	I am going to the cinema.					

5 Layer III: Part of Speech (POS)

5.1 Introduction

The layer “part of speech” indicates the grammatical categories of words. The general principle behind part of speech categorization in these guidelines is syntax-oriented. The idea is not to establish language specific categories, but to provide categorial information which is relevant for syntax. For instance, the word *walk* in English may be used as a noun or a verb. Rather than establishing a new category which captures all possible functions, e.g., “V/N” for *walk*, we recommend specifying the categorial information which is relevant in that context:

(62) English

<WORDS>	the	walk
<POS>	DET	N

(63) English

<WORDS>	to	walk
<POS>	PTC	VLEX

5.2 Tagset declaration

Similar to STTS, tag names for parts of speech are organized in a hierarchical manner: The first letter(s) indicate the superordinate category, e.g. N for ‘noun’, and subsequent letters denote subclasses, e.g. NCOM for ‘common noun’.

Table 5: List of tags for part of speech

tag	term
A	adjective
ADV	adverb
AT	attributive
CLF	classifier
COOR	coordinating conjunction
DET	determiner
N	noun
NCOM	common noun
NPRP	proper noun
P	preposition/postposition
PRON	pronoun
PRONDEM	demonstrative pronoun
PRONEXPL	expletive pronoun
PRONINT	interrogative pronoun
PRONPOS	possessive pronoun
PRONPRS	personal pronoun
PRONQUANT	quantifier

PRONREL	relative pronoun
PRONRFL	reflexive pronoun
PTC	particle
SU	substantive
SUB	subordinating conjunction
SUBADV	adverbial subordinating conjunction
SUBCOM	complementizer
V	verb
VAUX	auxiliary verb
VCOP	copula verbs
VDITR	ditransitive verb
VINTR	intransitive verb
VLEX	lexical verb
VMOD	modal verb
VN	verbal noun
VTR	transitive verb
CLIT	clitic form
FULL	full form

If a part of speech has some subclasses, as, e.g., in the case of ‘nouns’ which may be further divided into ‘common nouns’ and ‘proper nouns’, then it is recommended to choose one level of categorization, i.e. either annotate every noun just as ‘N’, or make the distinction between ‘NCOM’ and ‘NPRP’ every time. The same also holds for verbs, pronouns, etc.

(64) English, annotation of supercategories

<WORDS>	Peter	bicycle
<POS>	N	N

(65) English, annotation of subcategories

<WORDS>	Peter	bicycle
<POS>	NPRP	NCOM

5.3 Specific instructions

5.3.1 Nouns

General case

(66) English

<WORDS>	water
<POS>	N

Subclasses

- proper nouns:

(67) English

<WORDS>	Peter
<POS>	NPRP

- common nouns:

(68) English

<WORDS>	house
<POS>	NCOM

5.3.2 Verbs

General case

(69) English

<WORDS>	sleep
<POS>	V

Subclasses

The following subclasses of verbs may be used according to the function of the verb in certain contexts, i.e. the verb be would be annotated as VCOP in *be happy* and VAUX in *be destroyed.* Similarly, the German verb *wollen* ‘want’ would be annotated as VMOD in *ich will gehen* ‘I want to go’ and as VLEX in *ich will ein Eis* ‘I want ice-cream’.

- modal verbs:

(70) English

<WORDS>	can
<POS>	VMOD

- auxiliary verbs:

(71) English

<WORDS>	have
<POS>	VAUX

- copula verbs:

(72) English

<WORDS>	be
<POS>	VCOP

- lexical verbs:

(73) English

<WORDS>	walk
<POS>	VLEX

The annotation of part of speech follows the syntactic function of the verb. I. e., the verb *haben* in German may be a transitive verb if it is used with a direct object, or an auxiliary verb when it is used for the formation of perfect tenses.

(74) German

<WORDS>	Hunger	haben
<MORPH>	hunger	have:INF
<GLOSS>	NCOM	VLEX

(75) German

<WORDS>	gegessen	haben
<GLOSS>	eat:PRF.PTCP	have:INF
<POS>	VLEX	VAUX

- transitivity

  It is possible to distinguish between intransitive, transitive, and ditransitive verbs by using the following glosses:

(76) English

<WORDS>	sleep
<POS>	VINTR

(77) English

<WORDS>	buy
<POS>	VTR

(78) English

<WORDS>	give
<POS>	VDITR

5.3.3 Adjectives

(79) Spanish

<WORDS>	aburrido
<GLOSS>	boring
<POS>	A

5.3.4 Adverbs

(80) English

<WORDS>	soon
<POS>	ADV

(81) English

<WORDS>	where
<POS>	ADV

So called pronominal adverbs in German are also annotated as ADV:

(82) German

<WORDS>	darüber
<GLOSS>	there:over
<POS>	ADV

(83) German

<WORDS>	hierüber
<GLOSS>	here:over
<POS>	ADV

(84) German

<WORDS>	worüber
<GLOSS>	where:over
<POS>	ADV

(85) German

<WORDS>	dessentwegen
<GLOSS>	DEM:M.GEN.SG:because.of
<POS>	ADV

(86) German

<WORDS>	meinetwegen
<GLOSS>	1.SG.GEN:because.of
<POS>	ADV

5.3.5 Adpositions

Including all types of X-positions:

(87) English

<WORDS>	behind	the	house
<POS>	P	DET	NCOM

(88) English

<WORDS>	two	years	ago
<POS>	DET	NCOM	P

5.3.6 Determiners

Determiners include articles and numerals used as determiners (see §0; §5.3.8). They do not include demonstratives or quantifiers (cf. 5.3.8).

(89) English

<WORDS>	the
<POS>	DET

5.3.7 Conjunctions

All types of subordinators are annotated as SUB:

(90) English

<WORDS>	if
<POS>	SUB

<WORDS>	that
<POS>	SUB

<WORDS>	when
<POS>	SUB

If you need to indicate complementizers or adverbial subordinating conjunctions separately, then use the corresponding tags:

(91) English

<WORDS>	when
<POS>	SUBADV

(92) English

<WORDS>	that
<POS>	SUBCOM

Coordinating conjunctions are annotated as COOR:

(93) English

<WORDS>	and
<POS>	COOR

5.3.8 Pronouns

- personal pronouns:

(94) English

<WORDS>	you
<POS>	PRONPRS

- interrogative pronouns:

(95) English

<WORDS>	who
<POS>	PRONINT

- demonstrative pronouns:

(96) English

<WORDS>	this
<POS>	PRONDEM

Notice that German displays a demonstrative pronoun that is in most cases homonymous to the definite article.

(97) German

<WORDS>	Das	ist	es	.
<GLOSS>	this:N.SG.NOM	be:3.SG	3.SG.NOM	
<POS>	PRONDEM	VCOP	PRONPERS	

- reflexive pronouns:

  This category should be used only if the language possesses pronouns which are always used as reflexives, e.g. the English reflexive pronouns (not the German pronouns of the type *ich schäme mich*, where the ambiguity personal/reflexive is resolved in the argument structure of the given verb).

(98) English

<WORDS>	myself
<POS>	PRONRFL

- possessive pronouns:

(99) English

<WORDS>	your
<POS>	PRONPOS

- relative pronouns:

(100) English

<WORDS>	which
<POS>	PRONREL

- expletive pronouns:

  Expletive pronouns (also called “impersonal pronouns”, “pleonastic pronouns”) are pronouns which do not have any meaning but are syntactically required, as for instance:

(101) English

<WORDS>	there	is	a	man	.
<POS>	PRONEXPL	V	DET	N	

(102) German

<WORDS>	es	riecht	nach	Erdbeeren	.
<GLOSS>	3.SG	smell:3.SG	to	strawberry:DAT.PL	
<POS>	PRONEXPL	V	P	N	

(103) German

<WORDS>	es	regnet	.
<GLOSS>	3.SG	rain:3.SG	
<POS>	PRONEXPL	V	

We also use PRONEXPL for pre-field *es* in German. The difference between *es* in (101)-(103) and *es* in (104) is encoded at the syntactic layer:

(104) German

<WORDS>	es	kamen	drei	Sportler	.
<GLOSS>	3.SG	come:3.PL	three	sportsman[PL]	
<POS>	PRONEXPL	V	DET	N	

- quantifiers:

  The properties of quantifiers are described in detail in the semantics guidelines.

(105) German

<WORDS>	jeder
<GLOSS>	every.one:M.SG.NOM
<POS>	PRONQUANT

(106) German

<WORDS>	jeder	Mann
<GLOSS>	every:M.SG.NOM	man
<POS>	PRONQUANT	NCOM

(107) German

<WORDS>	alle
<GLOSS>	all:PL.NOM
<POS>	PRONQUANT

If you need to differentiate between substantive and attributive paradigms of pronouns, then use the following tags (append SU and AT respectively). Substantive pronouns replace the whole NP, attributive ones function as a determiner:

(108) English

<WORDS>	yours
<POS>	PRONPOSSU

(109) English

<WORDS>	your
<POS>	PRONPOSAT

5.3.9 Particles

(110) German

<WORDS>	ja
<GLOSS>	yes
< POS >	PTC

Interjections are also annotated as particles:

(111) German

<WORDS>	oh
<GLOSS>	oh
<POS>	PTC

5.3.10 Special instructions

Clitic vs. full forms

If a language makes a difference between clitic and full forms in a given category, then append the tags ‘FULL’ and ‘CLIT’. E.g.,

(112) Croatian

<WORDS>	jesam	sam
<MORPH>	be:1.SG	be:1.SG
<GLOSS>	VAUXFULL	VAUXCLIT

(113) Modern Greek

<WORDS>	eména
<GLOSS>	1.SG.ACC
<POS>	PRONPRSFULL

<WORDS>	me
<GLOSS>	1.SG.ACC
<POS>	PRONPRSCLIT

Numerals

Numerals are treated as members of broader syntactic categories (for the explicit marking of numerals, use the Semantic Annotation Layer QuP):

- cardinal numerals in English are treated as determiners;
- ordinal numerals in English are treated as adjectives;
- adverbial numerals in English are treated as adverbs.

(114) English

<WORDS>	two
<POS>	DET

<WORDS>	second
<POS>	A

<WORDS>	twice
<POS>	ADV

Discontinuity

Similar to discontinuous morphemes (see §4.4.3), discontinuous elements are indicated by indices also in the POS layer:

(115) English

<WORDS>	either	John	or	Mary
<POS>	COOR_1	NPRP	_1	NPRP

(116) German

<WORDS>	ich	fange	jetzt	an
<MORPH>	ich	fange	jetzt	an
<GLOSS>	1.SG	start:1.SG_1	now	_1
<POS>	PRONPRS	VLEX_1	ADV	_1

(117) German

<WORDS>	um	unseres	Vaters	willen
<POS>	P_1	PRONPOS	NCOM	_1

6 References

Bickel, Balthasar, Bernard Comrie, and Martin Haspelmath. 2002. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Leipzig: MPI for Evolutionary Anthropology & University of Leipzig (http://www.eva.mpg.de/lingua/files/morpheme.html).

König, Ekkehard (with Dik Bakker, Öesten Dahl, Martin Haspelmath, Maria Koptjevskaja-Tamm, Christian Lehmann, Anna Siewierska). 1993, *EUROTYP Guidelines*. European Science Foundation Programme in Language Typology.

Syntax

Joanna Blaszczak[1], *Stefanie Dipper*[1], *Gisbert Fanselow*[1], *Shinishiro Ishihara*[1], *Svetlana Petrova*[2], *Stavros Skopeteas*[1], *Thomas Weskott*[1], *Malte Zimmermann*[1]

University of Potsdam ([1]) and Humboldt University Berlin ([2])

The guidelines for syntactic annotation contain the layers that are especially relevant for queries related to the interaction of information structure with syntax. The layers of this level are constituent structure, grammatical functions, and semantic roles.

1 Preliminaries

The following guidelines is the original product of the collaboration among different projects within the SFB 632. They are only partially related to other syntactic annotation standards (e.g. Penn Treebank (Santorini 1990), GNOME (Poesio 2000), TIGER corpus (Albert et al. 2003), Verbmobil (Stegmann et al. 2000)). Our main goal is to annotate the most important syntactic information in a theory neutral way. In this sense, the guidelines contain a systematic list of linguistic categories that allow for retrieval of syntactic information from a cross-linguistic corpus. Users who need more fine-grained distinctions may declare further categories as long as this corresponds to the general logic of the guidelines.

Interdisciplinary Studies on Information Structure 07 (2007): 95–132
Dipper, S., M. Götze, and S. Skopeteas (eds.):
Information Structure in Cross-Linguistic Corpora

2 Layer declaration

Table 1: Layers

Layer	Name
Constituent structure	CONST
Grammatical functions	FUNCTION
Semantic roles	ROLE

Constituents, their functions, and their roles are annotated within single cells in the hierarchical tiers CS1... CS*n*. They are given in the order: constituent categorial label (e.g. NP) – grammatical function (e.g. SUBJ) – semantic role (e.g. AG).

(1) English

<WORDS>	I	saw	the	boy	who	ate	the	mango	.
<CS1>					NP-SUBJ-AG	V	NP-OBJ-THE		
<CS2>					S-ATTR				
<CS3>	NP-SUBJ-EXP	V	NP-OBJ-TH						
<CS4>	S-MAIN								

3 Layer I: Constituent structure (CS1… CS*n*)

3.1 Introduction

Since labeling of constituent structures always involves embedding, we will use multiple layers on EXMARaLDA for constituent structures. In principle, there is no limit for the number of constituent structure (i.e., one can create as many layers as he/she needs). Each layer will be named ‘CS1’, ‘CS2’, ..., ‘CS*n*’. The ordering of the numbers of annotation layers proceeds from the embedded layer ‘CS*j*’ to the embedding layer ‘CS*j*+1’ (see examples in §3.3). If one needs a

relatively complex and deeply embedded structure, it would be better to annotate the constituent structure using tools like ANNOTATE.

The table form of the constituent structure annotation looks like a reversed syntax tree. Sister constituents are annotated in the same table-line. Daughter constituents are annotated in the higher line.

3.2 Tagset declaration

Our annotation system defines a restricted number of phrasal constituents as declared in Table 2, focusing on the most important syntactic components.

Table 2: Tagset declaration for constituent structure

tag	meaning
AP	adjectival phrase
NP	noun phrase
PP	prepositional phrase
V	verbal head
S	sentence/clause

3.3 Instructions and illustrative examples

There are four obligatory labels for the annotation of constituent structure: NP, PP, V, and S. Verbs and arguments are directly dominated by the S node, i.e. there is no explicit VP node. The difference between internal and external arguments can be retrieved through the layers FUNCTION/ROLE. Only lexical verbs and copular verbs (i.e. units annotated as VLEX or VCOP at the POS layer) are annotated as V; modals and auxiliaries are not marked at the syntactic layer (they can be retrieved through the layer POS; see below). Final punctuation marks are part of the matrix S:

(2) English

<WORDS>	Peter	bought	apples	.
<CS1>	NP	V	NP	
<CS2>	S			

3.3.1 Noun phrase (NP)

An NP consists of a head noun plus any modifying or determining material, i.e. adjectives, relative clauses, determiners, demonstratives, etc.

- NPs typically occur as complements to verbs or prepositions/postpositions:

(3) English

<WORDS>	Peter	followed	the	elephant	.
<CS1>	NP	V	NP		
<CS2>	S				

(4) English

<WORDS>	the	ball	is	on	the	ground	.
<CS1>					NP		
<CS2>	NP		V	PP			
<CS3>	S						

(5) Japanese

<WORDS>	Taro-ga	hana-o	kat-ta	.
<GLOSS>	Taro-NOM	flowers-ACC	buy-PST	
<CS1>	NP	NP	V	
<CS2>	S			

- Substantive pronouns (*he*, *she*, *it*, *this*, *that*, *someone*, *anyone*, etc.) are NPs:

(6) English

<WORDS>	He	knows	that	.
<CS1>	NP	V	NP	
<CS2>	S			

(7) German

<WORDS>	Alles	klar
<GLOSS>	all:NOM.SG.N	clear
<CS1>	NP	
<CS2>	S	

- NPs can be embedded within another NP; note that the part of the NP *das Buch* is not annotated as such:

(8) German

<WORDS>	das	Buch	des	Lehrers
<GLOSS>	DEF: NOM.SG.N	book [NOM.SG.N]	DEF: GEN.SG.M	teacher: GEN.SG.M
<CS1>			NP	
<CS2>	NP			

- The syntactic structure of complex names may be ignored.

(9) English

<WORDS>	Noam	Chomsky
<CS1>	NP	

- In the case of discontinuous constituents, such as Split NPs or extraposed relative clauses, label both parts of the NP with an index number :

(10) German

<WORDS>	Autos	kaufte	Hans	blaue	.
<GLOSS>	car: N.ACC.PL	buy: PRT.3.SG	Hans: M.NOM.SG	blue: N.ACC.PL	
<CS1>	NP_1	V	NP	_1	
<CS2>	S				

(11) German

<WORDS>	A	boy	came	yesterday	who	ate	the	mango	.
<POS>			VLEX			VLEX			
<CS1>					NP	V	NP		
<CS2>					S				
<CS3>	NP_1		V		_1				
<CS4>	S								

- Expletive subjects are annotated as NPs:

(12) English

<WORDS>	It	is	raining	.
<POS>	PRONEXPL	VAUX	VLEX	
<CS1>	NP		V	
<CS2>	S			

- In languages like German and Dutch, expletives can occupy the first position in the sentence (so called pre-field) without being the subject. These expletive elements are not annotated at the syntactic layer.

(13) German

<WORDS>	Es	hat	ein	Mann	angerufen	.
<GLOSS>	3.SG.N	have: 3.SG	INDEF: M.NOM.SG	man: M.NOM.SG	call: PRF.PTCP	
<POS>	PRONEXPL	VAUX	DET	NCOM	VLEX	
<CS1>			NP		V	
<CS2>	S					

IMPORTANT NOTE: All kinds of pronouns are annotated as NP! Exceptions: expletives in the pre-field in German; non-substantive pronouns, e.g., possessive pronouns (which do not substitute for a complete NP but for a determiner only).

3.3.2 Prepositional phrase (PP)

A PP consists of a prepositional/postpositional head and its NP-complement (see (3) above), plus optional modifiers.

(14) German

<WORDS>	exactly	in	the	middle
<CS1>			NP	
<CS2>	PP			

- In the case of Preposition stranding, label both parts of the PP (preposed NP and the head P) with the same index number (just like the Split NP case).

(15) English

<WORDS>	Who	did	you	give	the	book	to	?
<CS1>	NP							
<CS2>	PP_1		NP	V	NP		_1	
<CS3>	S							

- Pronominal adverbs are also PP constituents.

(16) German

<WORDS>	Ich	warte	darauf	.
<GLOSS>	1.SG:NOM	wait:1.SG	there.on	
<POS>	P	VLEX	ADV	
<CS1>	NP	V	PP	
<CS2>	S			

- Constituents containing a subordinating conjunction such as “as”

governing an NP are annotated as PPs.

(17) German

<WORDS>	wie	Hans
<GLOSS>	as	Hans.M[SG.NOM]
<POS>	SUB	NPRP
<CS1>		NP
<CS2>	PP	

(18) German

<WORDS>	als	erstes
<GLOSS>	as	first.N[SG.NOM]
<POS>	SUB	NPRP
<CS1>		NP
<CS2>	PP	

- PPs may be embedded within higher PPs.

(19) German

<WORDS>	bis	zum	Auto
<MORPH>	bis	zu-m	Auto
<GLOSS>	1.SG:NOM	to-DEF:[N]SG.DAT	car.N[SG.DAT]
<POS>	P	P-DET	NCOM
<CS1>			NP
<CS2>		PP	
<CS3>	PP		

- In the following cases, the annotation just marks a flat PP.

(20) German

<WORDS>	von	hinten
<GLOSS>	from	back
<POS>	P	ADV
<CS1>	PP	

(21) German

<WORDS>	zum	Pferd	zurück
<MORPH>	zu-m	Pferd	zurück
<GLOSS>	to-[N]SG.DAT	horse:[N]SG.DAT	back
<POS>	P-DET	NCOM	ADV
<CS1>		NP	
<CS2>	PP		

3.3.3 Verb (V)

A V at the syntax layer is either a lexical (VLEX) or a copula verb (VCOP) at POS layer. Modal verbs and auxiliaries are not annotated in the constituent structure. The verb and its arguments are placed at the same CS*n* layer.

(22) English

<WORDS>	Peter	enthusiastically	sang	a	song	to	Mary	.
<POS>			VLEX					
<CS1>							NP	
<CS2>	NP		V	NP		PP		
<CS3>	S							

(23) English

<WORDS>	I	saw	the	boy	who	ate	the	mango	.
<POS>		VLEX				VLEX			
<CS1>					NP	V	NP		
<CS2>					S				
<CS3>	NP	V	NP						
<CS4>	S								

(24) English

<WORDS>	There	is	a	man	in	the	garden	.
<POS>		VCOP						
<CS1>						NP		
<CS2>		V	NP		PP			
<CS3>	S							

In examples with modal verbs and auxiliaries, the V node is only assigned to the lexical verb.

(25) English

<WORDS>	He	must	go	to	Paris	.
<POS>		VMOD	VLEX			
<CS1>					NP	
<CS2>	NP		V	PP		
<CS3>	S					

Raising and control verbs are treated like ordinary verbs. They subcategorize for a sentential complement as shown in (26) and (27) below. Compare the annotation of the control verb *intend* and the raising verb *seem* below with the

annotation of the modal verb *must* in (25): Note that at the layer CS3 of (26) and (27), infinitival phrases are annotated as S.

(26) English

<WORDS>	He	intended	to	go	to	Paris	.
<POS>		VLEX		VLEX			
<CS1>						NP	
<CS2>				V	PP		
<CS3>	NP	V	S				
<CS4>	S						

(27) English

<WORDS>	He	seems	to	go	to	Paris	.
<POS>		VLEX		VLEX			
<CS1>						NP	
<CS2>				V	PP		
<CS3>	NP	V	S				
<CS4>	S						

3.3.4 Adjectival Phrase (AP)

In general, adjectives are not annotated at the syntactic layer. However, there are two exceptions: adjectives (or APs) that function as nominal predicates are annotated with AP. The head of the AP is not labeled; this information can be retrieved from the POS layer.

(28) English

<WORDS>	He	seems	to	be	thick	.
<POS>		VLEX		VLEX	ADJ	
<CS1>				V	AP	
<CS2>	NP	V	S			
<CS3>	S					

Similarly, APs that have arguments are also annotated.

(29) English

<WORDS>	Der	auf	Maria	stolze	Mann	lacht	.
<GLOSS>	DEF: M.NOM.SG	on	Maria: F.SG	proud:M.NOM.SG	man:M.NOM.SG	laugh:3.SG	
<CS1>			NP				
<CS2>		PP					
<CS3>		AP					
<CS4>	NP					V	
<CS5>	S						

3.3.5 Clause (S)

'S' stands for clauses. It marks both main clauses and subordinate clauses. The root S symbol also covers the final punctuation mark.

Here are some examples:

- Relative clause; note that the part without the relative clause (*the boy*) is not annotated as NP.

(30) English

<WORDS>	the	boy	who	ate	the	mango	...
<CS1>			NP	V	NP		
<CS2>			S				
<CS3>	NP						

- Clausal complement (embedded clause)

(31) English

<WORDS>	I	thought	that	John	loves	Mary	.
<CS1>				NP	V	NP	
<CS2>	NP	V	S				
<CS3>	S						

Dependent verb forms (infinitives, gerunds, participles, etc.) are labeled as S:

- Infinitival complements of lexical verbs are annotated as S:

(32) English

<WORDS>	I	intend	to	go	to	Paris	.
<CS1>						NP	
<CS2>				V	PP		
<CS3>	NP	V	S				
<CS4>	S						

- The same holds for nominalized verb forms:

(33) English

<WORDS>	He	likes	eating	apples	.
<CS1>			V	NP	
<CS2>	NP	V	S		
<CS3>	S				

- Verbs used in attributive constructions are annotated as S if they contain arguments or PP adjuncts (compare the examples below):

(34) German

<WORDS>	Der	lachende	Mann	schläft	.
<CS1>	NP			V	
<CS2>	S				

(35) German

<WORDS>	Der	auf	Maria	wartende	Mann	lacht	.
<GLOSS>	DEF: M.NOM.SG	on	Maria: F.SG	wait: PTC:M.NOM.SG	man: M.NOM.SG	laugh: 3.SG	
<CS1>			NP				
<CS2>		PP					
<CS3>		S					
<CS4>	NP					V	
<CS5>	S						

- Like attributive verbs, adverbial forms are annotated as S if they contain

arguments or PP adjuncts:

(36) English

<WORDS>	Er	fliegt	auf	einem	Teppich	sitzend	.
<GLOSS>	3.SG: M.NOM	fly: 3.SG	on	INDEF: M.DAT.SG	carpet: M.SG	sit:PTC	
<CS1>				NP			
<CS2>			PP				
<CS3>	NP	V	S				
<CS4>	S						

- Examples for correlative expressions:

(37) German

<WORDS>	Ich	warte	darauf	,	dass	er	kommt	.
<GLOSS>	1.SG:NOM	wait:1.SG	there.on		that	3.SG: NOM.M	come: 3.SG	
<POS>	P	VLEX	ADV		SUBCOM	P	VLEX	
<CS1>						NP	V	
<CS2>					S			
<CS3>	NP	V	PP					
<CS4>	S							

(38) German

<WORDS>	Er	weiß	es	,	dass	sie	kommt	.
<GLOSS>	3.SG: NOM.M	know: 3.SG	3.SG: ACC.N		that	3.SG: NOM.F	come: 3.SG	
<POS>	P	VLEX	P		SUBCOM	P	VLEX	
<CS1>						NP	V	
<CS2>					S			
<CS3>	NP	V	NP					
<CS4>	S							

3.3.6 Coordination

Coordinated constituents are annotated as S if they contain a verb.

(39) English

<WORDS>	John	eats	apples	and	Mary	eats	oranges	.
<CS1>	NP	V	NP		NP	V	NP	
<CS2>	S				S			
<CS3>	S							

(40) English

<WORDS>	John	eats	apples	and	drinks	water	.
<CS1>	NP	V	NP		V	NP	
<CS2>	S				S		
<CS3>	S						

(41) English

<WORDS>	John	eats	apples	and	water	.
<CS1>	NP	V	NP		NP	
<CS2>	S					

If the coordinated constituents belong to different categories, their union is annotated as S.

(42) English

<WORDS>	wo	und	wer	bist	Du	?
<GLOSS>	where	and	who:M.NOM.SG	be:2.SG	2.SG.NOM	?
<CS1>	PP		NP	VCOP	NP	
<CS2>	S		S			
<CS3>	S					

3.3.7 Punctuation marks

In general, punctuation marks are not included in the constituent structure. The only exception exception is the sentence final punctuation (‘.’ or ‘?’, etc.) which

is dominated by root S. This allows for easy retrieval of sentence type (declarative, interrogative, imperative).

(43) English

<WORDS>	He	met	Peter	,	who	read	a	book	.
<CS1>					NP	V	NP		
<CS2>			NP		S				
<CS3>	NP	V	NP						
<CS4>	S								

3.3.8 Ellipsis, traces, etc.

The current guidelines only support annotation for overt information. Elided elements are not annotated as such.

(44) English

<WORDS>	Peter	bought	apples	and	Mary	oranges	.
<CS1>	NP	V	NP		NP	NP	
<CS2>	S				S		
<CS3>	S						

(45) German

<WORDS>	Peter	möchte	ein	rotes	.
<GLOSS>	Peter:M.NOM.SG	want:3.SG	INDEF:N.ACC.SG	red:N.ACC.SG	
<POS>	N	VLEX	DET	ADJ	
<CS1>	NP	V	NP		
<CS2>	S				

(46) German

<WORDS>	Zu	der	Post	?
<GLOSS>	to	DEF:F.DAT.SG	post.office:F.SG[DAT]	
<POS>	P	DET	N	
<CS1>		NP		
<CS2>	PP			
<CS3>	S			

4 Layer II: Grammatical functions (FUNCTION)

4.1 Introduction

This layer encodes the syntactic relations that various syntactic constituents in a clause (NP, PP, AP, S) entertain with respect to the main verb of that clause. Relevant information at this layer relates to the questions of (i) whether a syntactic constituent is an obligatory addition to the verb (*argument*), or whether it is an optional addition that could be easily left out (*adjunct*), (ii) whether the relative status of the different arguments differs and – if so – which of the arguments of a verb (if any) has a prominent status with respect to grammatical processes such as agreement, binding, focus marking etc.

Note that only constituents that are annotated at the CS layers may be labeled for grammatical function.

4.2 Tagset declaration

Table 3: Core annotation scheme

tag	meaning
ADJ	adjunct
ADV	adverbial subordinate clause
ARG	argument

ATTR	relative clause
MAIN	main clause
PRDNOM	predicate nominal

Table 4: Extended annotation scheme

tag	meaning
SUBJ	Subject
OBJ	unspecified object
DO	direct object
IO	indirect object

4.3 Instructions

4.3.1 General

The tags for grammatical functions are given within the layers of constituent structure after the constituent labels.

(47) English

<WORDS>	He	met	Peter	,	who	read	a	book	.
<CS1>					NP-SUBJ	V	NP-DO		
<CS2>	NP-SUBJ	V	NP-DO		S-ATTR				
<CS3>	S-MAIN								

4.3.2 Core vs. Extended Annotation scheme

Not all projects will need equally fine-grained distinctions between the various grammatical functions. For instance, while for some projects it may be sufficient to mark the difference between a syntactic *argument* (ARG) and a syntactic *adjunct* (ADJ), others may want to mark differences between different kinds of arguments, say subject (SUBJ) vs (direct) object (DO). In the absence of further

guidelines from the individual project, the annotators are recommended to restrict their annotation to the core scheme.

4.3.3 Core scheme

ARG.

The category ARG is assigned to those syntactic constituents that appear as obligatory complements to the main verb. This means that they CANNOT be left out without a change in grammaticality or a significant change in meaning. (Notice that the obligatory appearance of an element in some syntactic position, does not necessarily mean, that this element is an argument. It can be that such elements appear in specific syntactic position due to some special syntactic requirements of a given language, e.g., in V2 sentences in German, some element (sometimes an expletive one) must obligatorily appear in the first position, e.g. *danach kommt ein Einhörnchen, ein Einhörnchen kommt danach, *kommt ein Einhörnchen danach, es kommt ein Einhörnchen danach.*)

ARGs mostly (but not always!) refer to (groups) of individuals and are assigned structural case (NOM, ACC, PAR) in case-assigning languages.

As the classical terminology suggests, intransitive verbs such as *sleep* in *John sleeps* only take one argument, namely the NP *John* (note the ungrammaticality of **sleeps*). Transitive verbs such as *criticizes* in *John criticizes a book* take two arguments, namely *John* and *a book* (note that omission of *a book* induces an unspecific generic meaning, along the lines of 'John generally criticizes something or other'). A ditransitive verb such as *give* in *John gave Mary a book* takes three arguments, namely *John*, *Mary*, and *a book*. Again, omission of one or more of the arguments either leads to ungrammaticality (**gave Mary*) or to a change in (verb) meaning (*John gave a book* = 'John donated a book').

(48) English

<WORDS>	John	gave	Mary	a	book	.
<CS1>	NP-ARG	V	NP-ARG	NP-ARG		
<CS2>	S-MAIN					

ADJ.

The category ADJ (=adjunct) is assigned to those constituents that appear as optional additions, be it to the main verb or to a given noun.

This means that they CAN be left out freely without a change in grammaticality or a significant change in meaning. In *John called Mary (from school) (with his cell phone)* the optional additions *from school* and *with his cell phone* are such optional additions that can be left out freely.

Adjuncts are generally used to convey additional information about the time, place, manner, or cause of the event or situation described by the clause (see below). That is, they restrict the class of events/situations described by the clause to a subset. If required the category ADJ can be split up into semantic sub-categories, that are annotated in layer semantic roles (time, location, etc.).

(49) English

<WORDS>	Today	John	came	to	school	.
<CS1>					NP-ARG	
<CS2>		NP-ARG	V	PP-ARG		
<CS3>	S-MAIN					

PPs may be either arguments or adjuncts. They are annotated as arguments when they are governed by the verb. Some identifying properties of arguments PPs are that (a) the semantics is not compositional and (b) the choice of preposition depends on the verb totally. The prototypical category are verbs that govern certain prepositions:

(50) English

<WORDS>	I	am	waiting	for	Mary	.
<CS1>					NP-ARG	
<CS2>	NP-SUBJ		V	PP-ARG		
<CS3>	S-MAIN					

PPs that are not governed by the verb are adjuncts.

(51) English

<WORDS>	I	am	sleeping	in	bed	.
<CS1>					NP-ARG	
<CS2>	NP-SUBJ		V	PP-ADJ		
<CS3>	S-MAIN					

Note that NPs may be adjuncts too:

(52) English

<WORDS>	The	other	day	John	came	to	school	.
<CS1>							NP-ARG	
<CS2>	NP-ADJ			NP-ARG	V	PP-ARG		
<CS3>	S-MAIN							

4.3.4 Extended scheme

SUBJ.

The category SUBJ is assigned to a designated argument that is prominent with respect to a number of grammatical relations such (i.) as constituency with the verb, (ii.) agreement, (iii.) and binding, etc. This prominence is often taken to correspond to a prominent position in the syntactic structure of the clause.

(i) Unlike direct objects, subjects do not seem to form a constituent with the verb as shown by the fact that the two cannot be topicalised together in **[**Johann** gesehen] hat den Mann nicht* vs *[Den Mann gesehen] hat Johann nicht*.

(ii) In agreement languages, the subject is that argument that the verb always agrees with (in some languages the verb additionally agrees with the object as well): ***Johann*** *(sg.) sleeps (sg.)* vs. **The boys (pl.) sleeps (sg.)*

(iii) subjects can bind reflexive pronomina: ***Peter*** *blamed himself* vs. **Heself blamed Peter*.

Other properties that may help to identify SUBJs:

- In NOM-ACC-languages, the default case of nominal subjects is the Nominative: ***Der Mann*** *ist gekommen* vs **Den Mann ist gekommen*. This can be formalized in a rule:

  If there is only one nominative constituent in a clause, mark this constituent as SUBJ.

- Subjects are most often expressed by nominal constituents (NPs), but sentential subjects as in *[**That Peter won the race**] surprised me* are also possible with certain verbs.

- In languages that do not mark case morphologically, the subject status is coded by word order, i.e. the subject usually occupies a designated (linear) position relative to verb and direct object (if present). E.g., in English or French, the subject precedes the verb (and the direct object) in the default case: ***Peter*** *saw her* vs. **Her saw Peter*.

  Note that this test has to be applied with care. It seems to work fine with transitive SVO-structures, but in intransitive or existential structures the subject may also follow the object: e.g. *There came Peter out of the hall*.

- In languages that mark NOM only sporadically (e.g. on pronouns and full NPs, but not on CPs (German), or only on pronouns, but not on full NPs and CPs), a substitution test combined with considerations of linear order may help in some cases:

If a constituent α is not morphologically marked for case, but if α is in the default position for subjects (this must be independently established on the base of reference grammars) and if α can be replaced with a NOM-marked constituent β, mark α with SUBJ:

a. [$_{\alpha}$ Peter] saw Mary → He$_{nom}$ saw Mary.
   substitution possible → mark α = Peter with SUBJ

b. [$_{\alpha}$ That Peter came] surpised us. → He$_{nom}$ surprised us.
   substitution possible → mark α = [that Peter came] with SUBJ

Warning: It does not follow automatically from the impossibility of substitution that α is NOT a subject. In English, case-marked pronouns cannot be substituted for subject NPs in existential sentences because of their definiteness: *There came **Peter** down the Hill* //→ **There came **he** down the hill.*

- Often the subject has the semantic role of AGENT (see 5 below), but this is not a 1:1-correspondence. E.g., in passive structures, non-agentive constituents function as subjects syntactically: ***He** was beaten*. Likewise, the subject of transitive psych-verbs such as *to like* in ***He** likes dogs* does not refer to the agent of a particular event, but rather to the experiencer (EXP) of a particular psychological disposition.

DO.

The category DO is assigned to the second argument of a transitive verb, which is not designated in the sense that it is less prominent than the subject. This rule-of-thumb makes the NP *Bill* in *The boys like **Bill*** the DO, since it does not agree with the main verb in number.

Like subjects, DOs are assigned structural case (ACC/PAR or ABS) in case-assigning languages. Like subjects, DOs have a default base position

relative to verb and subject in languages that do not assign case: In English and French, the DO follows the main verb (and the subject). DOs are generally taken to stand in close syntactic relation with the main verb, which is reflected by the fact that they can be displaced together: *[**Den Mann** gerufen] haben wir*.

Apart from this, DOs are often only identifiable based of the absence of properties typical for subjects. E.g. a DO cannot bind a reflexive in a subject position (see above), and it cannot agree with the verb in the absence of subject-verb-agreement. Other Properties that may help to identify DOs:

- There is a tendency for DOs to express the semantic role of PATIENT/THEME. However, even if all PATIENT/THEME -expressions are DOs the reverse does not hold completely. Consider e.g. *The news surpised John*, where the DO *John* expresses the semantic role of experiencer.
- As with subjects, DOs are most often expressed by nominal constituents (NPs), but sentential DOs are also possible, especially with attitude verbs (*to think, to belief*) or verbs of saying: *John said [that Maria had come late again]*.

IO.

The category IO is assigned to that argument of a (ditransitive) verb that is not assigned the status of SUBJ nor DO. In case-languages, IOs are often assigned the Dative. Semantically, the IO is often used to express the receiver or beneficient/maleficient of an event, such as the NP *John* in *Mary gave **John** a book/ kiss*.

Unlike SUBJs and DOs, IOs seem to always refer to individuals and must be expressed by a nominal constituent.

(53) English

<WORDS>	John	gave	Mary	a	book	.
<CS1>	NP-SUBJ	V	NP-ARG	NP-ARG		
<CS2>	S-MAIN					

Prepositional objects are annotated with the generic label OBJ:

(54) German

<WORDS>	Ich	warte	auf	Hans	.
<GLOSS>	1.SG:NOM	wait:1.SG	on	Hans:ACC.SG.M	
<CS1>				NP-ARG	
<CS2>	NP-SUBJ	V	PP-OBJ		
<CS3>	S				

4.3.5 Nominal Predicates

PRDNOM: a nominal predicate (noun or adjective), either with or without copula.

(55) English

<WORDS>	He	is	thick	.
<CS1>	NP-SUBJ	V	AP-PRDNOM	
<CS2>	S-MAIN			

(56) English

<WORDS>	He	is	the boss	.
<CS1>	NP-SUBJ	V	NP-PRDNOM	
<CS2>	S-MAIN			

(57) Russian

<WORDS>	Ona	studentka	.
<CS1>	NP-SUBJ	NP-PRDNOM	
<CS2>	S-MAIN		

The term nominal predicate may be used for the complements of further copulative verbs (cf. small clauses), e.g. *consider, call,* etc.

(58) English

<WORDS>	He	considers	him	a	thief	.
<CS1>	NP-SUBJ	V	NP-OBJ	NP-PRDNOM		
<CS2>	S-MAIN					

4.3.6 Sentences and clauses

Sentences and clauses are annotated in four categories:

- The tag MAIN is used for main clauses.

(59) English

<WORDS>	John	sleeps	.
<CS1>	NP-SUBJ	V	
<CS2>	S-MAIN		

- Relative clauses are annotated as ATTR.

(60) English

<WORDS>	I	saw	the	boy	who	ate	the	mango	.
<POS>		VLEX				VLEX			
<CS1>					NP-SUBJ	V	NP-OBJ		
<CS2>					S-ATTR				
<CS3>	NP-ARG	V	NP-OBJ						
<CS4>	S-MAIN								

- Subordinate clauses with the function of an argument (subject or object) are annotated as ARG.

(61) English

<WORDS>	Mary	thinks	that	he	came	.
<CS1>				NP-SUBJ	V	
<CS2>	NP-SUBJ	V	S-ARG			
<CS3>	S-MAIN					

- Subordinate clauses with adverbial function are annotated as ADV.

(62) English

<WORDS>	Tom	sleeps	when	the	sun	rises	.
<CS1>				NP-SUBJ		V	
<CS2>	NP-SUBJ	V	S-ADV				
<CS3>	S-MAIN						

4.3.7 Non-annotated syntactic functions

The following elements are not annotated for grammatical function:

- particles, e.g., German *ja, jawohl, doch*, etc.: these elements do not have a grammatical function, but rather they express speaker's attitudes towards the proposition.
- conjunctions, e.g., *and, but, because,* etc.
- adjectives in attributive use, e.g. *a nice boy*: the attributive function may be inferred by the fact that the adjective is part of the entire NP.

5 Layer III: Semantic roles (ROLE)

5.1 Introduction

Lexical heads not only require a certain number of arguments but also determine the semantic properties of these arguments depending on how these are involved in the state of affairs described by the lexical head. This means that the syntactic arguments enter certain *semantic* (also called *thematic* or *theta*-) roles, which are

pre-established by the selecting properties of the lexical head. The relationship between a lexical head and its arguments can be explained by the use of a small finite set of universally applicable notions which indicate whether a certain argument is the performer of an action, just undergoes an action etc. Note that only constituents that are annotated at the CS and FUNCTION layers may be labeled for semantic role.

5.2 Tagset declaration

The tags of semantic roles are not given in separate layers. They are inserted in the layers of constituent structure after the labels of grammatical functions.

Table 5: Core annotation scheme

tag	meaning
AG	Agent
CAUSE	Cause
COM	Comitative
EXPER	Experiencer
GOAL	Goal
INSTR	Instrument
LOC	Location
MAN	Manner
POSS	Possessor
THEME	Theme
TIME	Time

5.3 Instructions

5.3.1 General

The tags for semantic roles are given within the layers of constituent structure after the grammatical functions.

(63) English

<WORDS>	He	met	Peter	,	who	read	a	book	.
<CS1>					NP-SUBJ-AG	V	NP-DO-THEME		
<CS2>					S-ATTR				
<CS3>	NP-SUBJ-AG	V	NP-DO-THEME						
<CS4>	S-MAIN								

The tags for semantic roles are used with NPs, PPs, or S-ARGs that function either as arguments of verbs (*John sleeps*), or as adjuncts (*in Athens…*), or as dependents of NPs (*the house on the hill*). Not all constituents are annotated for semantic role, e.g. NP arguments of prepositions, relative clauses, etc. are not labeled for this layer.

5.3.2 Agent

NPs that refer to the entities that cause actions, either animates or inanimates, are annotated as agents.

(64) English

<WORDS>	The	boy	opens	the	window	.
<CS1>	NP-SUBJ-AG		V	NP-OBJ-THEME		
<CS2>	S-MAIN					

(65) English

<WORDS>	The	wind	opens	the	window	.
<CS1>	NP-SUBJ-AG		V	NP-OBJ-THEME		
<CS2>	S-MAIN					

5.3.3 Theme

Theme is a general term covering the notions of:

- *Patient*: an entity affected by the action

(66) English

<WORDS>	The	girl	paints	the	fence	.
<CS1>	NP-SUBJ-AG		V	NP-OBJ-THEME		
<CS2>	S-MAIN					

- *Result*: an entity effected by the action, i.e. which emerges out of the action:

(67) English

<WORDS>	The	woman	built	a	house	.
<CS1>	NP-SUBJ-AG		V	NP-OBJ-THEME		
<CS2>	S-MAIN					

- *Theme*: an entity effected by the action, i.e. which emerges out of the action:

(68) English

<WORDS>	Akropolis	is	in	Athens	.
<CS1>				NP-ARG	
<CS2>	NP-SUBJ-THEME	V	PP-ARG-LOC		
<CS3>	S-MAIN				

5.3.4 Experiencer

Experiencer is the sentient being that participates in a state/event of emotion (*love, hate,* etc.), volition (*wish, want, etc.*), cognition (*think, remember*, etc.), perception (*see*, *hear*, etc.) or bodily sensation (*feel cold*, *feel hungry*, etc.).

(69) English

<WORDS>	Mary	enjoys	algebra	.
<CS1>	NP-SUBJ-EXPER	V	NP-OBJ-THEME	
<CS2>	S-MAIN			

(70) English

<WORDS>	Algebra	interests	John	.
<CS1>	NP-SUBJ-THEME	V	NP-OBJ-EXPER	
<CS2>	S-MAIN			

5.3.5 Goal

Goal is a general term covering the notions of:

- *Recipient*: an entity which receives something:

(71) English

<WORDS>	John	gave	Mary	a	book	.
<CS1>	NP-SUBJ-AG	V	NP-IO-GOAL	NP-ARG-THEME		
<CS2>	S-MAIN					

- *Benefactive*: an entity to whose advantage an action is performed (or *malefactive*: an entity to whose disadvantage an action is performed):

(72) English

<WORDS>	John	bought	flowers	for	Mary	.
<CS1>					NP-ARG	
<CS2>	NP-SUBJ-AG	V	NP-OBJ-THEME	PP-ADJ-GOAL		
<CS3>	S-MAIN					

- *Purpose*: the intension for which an action is performed:

(73) English

<WORDS>	John	said	it	for	more	clarity	.
<CS1>					NP-ARG		
<CS2>	NP-SUBJ-AG	V	NP-OBJ-THEME	PP-ADJ-GOAL			
<CS3>	S-MAIN						

5.3.6 Instrument

Instruments are means with the help of which the action is carried out.

(74) English

<WORDS>	John	opened	the	door	with	the	keys	.
<CS1>						NP-ARG		
<CS2>	NP-SUBJ-AG	V	NP-OBJ-THEME		PP-ADJ-INSTR			
<CS3>	S-MAIN							

5.3.7 Possessor

Possessor is the entity who owns something.

(75) English

<WORDS>	Bill	has	a	new	car	.
<CS1>	NP-SUBJ-POSS	V	NP-OBJ-THEME			
<CS2>	S-MAIN					

(76) English

<WORDS>	Bill’s	car
<CS1>	NP-POSS	NP
<CS2>	NP	

5.3.8 Location

Location covers the spatial relations of:

- static spatial location:

(77) English

<WORDS>	Mary	is	in	New	York	.
<CS1>				NP-ARG		
<CS2>	NP-SUBJ-THEME	V	PP-ARG-LOC			
<CS3>	S-MAIN					

- direction of movement (do not mistake *direction* with *goal*, the latter being preserved for the intended target of an action not necessarily connected with spatial movement, see 5.3.5):

(78) English

<WORDS>	He	rushed	to	the	street	.
<CS1>				NP-ARG		
<CS2>	NP-SUBJ-AG	V	PP-ARG-LOC			
<CS3>	S-MAIN					

(79) English

<WORDS>	He	put	the	money	into	his	pocket	.
<CS1>						NP-ARG		
<CS2>	NP-SUBJ-AG	V	NP-OBJ-THEME		PP-ARG-LOC			
<CS3>	S-MAIN							

- source: indicating the origin of movement

(80) English

<WORDS>	The	gold	falls	from	the	sky	.
<CS1>					NP-ARG		
<CS2>	NP-SUBJ-THEME		V	PP-ARG-LOC			
<CS3>	S-MAIN						

- path: indicating a place through which the movement takes place.

(81) English

<WORDS>	He	ran	through	the	door	.
<CS1>				NP-ARG		
<CS2>	NP-SUBJ-THEME	V	PP-ARG-LOC			
<CS3>	S-MAIN					

5.3.9 Time

Time covers a point or an interval of time at which the action takes place.

(82) English

<WORDS>	He	came	at	noon	.
<CS1>				NP-ARG	
<CS2>	NP-SUBJ-AG	V	PP-ADJ-TIME		
<CS3>	S-MAIN				

(83) English

<WORDS>	He	worked	all	night	long	.
<CS1>	NP-SUBJ-AG	V	NP-ADJ-TIME			
<CS2>	S-MAIN					

5.3.10 Cause

Cause indicates the reason why something happens and is often expressed by a PP (*because of, with, through* etc.). Sometimes this role is close to the role of Instrument. The criterion for the choice of tag CAUSE is if the expression can be paraphrased through a clausal subordinate clause:

(84) He convinced me <u>with his honesty.</u> ↔ He convinced me because he was honest.

(85) He climbed <u>with a hammer.</u> ≠ He climbed because he had a hammer.

(86) English

<WORDS>	He	stroke	me	with	his	originality	.
<CS1>					NP-ARG		
<CS2>	NP-SUBJ-THEME	V	NP-OBJ-EXPER	PP-ADJ-CAUSE			
<CS3>	S-MAIN						

(87) English

<WORDS>	I	worked	because	he	liked	it	.
<CS1>				NP-SUBJ-EXP	V	NP-OBJ-THEME	
<CS2>	NP-SUBJ-AG	V	S-ADJ-CAUSE				
<CS3>	S-MAIN						

(88) English

<WORDS>	Why	did	it	happen	?
<CS1>	ADJ-CAUSE		NP-SBJ-THEME	V	
<CS2>	S-MAIN				

5.3.11 Manner

Manner applies to constituents that denote how something is carried out.

(89) English

<WORDS>	Handle	with	care	!
<CS1>			NP-ARG	
<CS2>	V	PP-ADJ-MAN		
<CS3>	S-MAIN			

Adverbs may also denote manner, however, they are not annotated at any of the syntactic layers.

(90) English

<WORDS>	Ann	drove	quickly	.
<CS1>	NP-SUBJ-AG	V		
<CS2>	S-MAIN			

5.3.12 Comitative

Comitative applies to an animate entity that accompanies a participant of the action.

(91) English

<WORDS>	Peter	walked	with	Bill	.
<CS1>				NP	
<CS2>	NP-SUBJ-AG	V	PP-ADJ-COM		
<CS3>	S-MAIN				

6 Problematic cases

6.1 Sentence fragments

As a rule of the thumb: Provide the maximum information for what you see. For instance, in case of fragmentary answers to yes/no questions (compare (92)), annotate *yes* or *no* as S. In case of fragmentary answers to constituent questions (compare (93)), annotate the fragment according to its syntactic category and function; note that the fragment is also annotated as S.

(92) English

<WORDS1>	Are	you	hungry	?		
<WORDS2>					yes	.
<CS1>	V	NP-SUBJ-THEME	PRDNOM		S-MAIN	
<CS2>	S-MAIN					

(93) English

<WORDS1>	Who	ate	beef	?		
<WORDS2>					John	.
<CS1>	NP-SUBJ-AG	V	NP-OBJ-THEME		NP-SUBJ	
<CS2>	S-MAIN				S-MAIN	

6.2 Correction and breaks

Corrections by the speaker, i.e., words or sequences of words that serve to correct erroneous utterances, are marked with the symbol "!". This indicates that a constituent which has already been introduced is updated/corrected. Breaks and break fillers are not annotated in the constituent structure.

(94) English

<WORDS>	John	...	eh	...	no	Peter	laughed	.
<CS1>	NP-SUBJ-AG					!NP-SUBJ-AG	V	
<CS2>	S-MAIN							

(95) English

<WORDS>	Mary	saw	John	...	no	Peter	.
<CS1>	NP-SUBJ-EXP	V	NP-OBJ-THEME			!NP-OBJ-THEME	
<CS2>	S-MAIN						

In case only parts of constituents are corrected, only the corrected version (the complete constituent) is annotated in the constituent structure layer.

(96) English

<WORDS>	A	...	eh	...	the	woman	comes	.
<CS1>					NP-SUBJ-AG		V	
<CS2>	S-MAIN							

6.3 Non-grammatical sequences

In the case of ungrammatical information, if it is obvious to the annotator what the speaker actually wanted to say, the ungrammatical feature is marked with the symbol “#”. The symbol is annotated at the layer at which the error arises, e.g. with incorrect case, at the morphological transcription, as in (75), or with incorrect word order, at the constituent structure, as in (76), (77). In case of errors in the constituent structure, the error should be marked as locally as possible, i.e., at the smallest erroneous constituent, compare (76) and (77).

(75) German

<WORDS>	Hans	sah	mir	.
<GLOSS>	Hans:NOM.SG.M	see:PAST.3.SG	1.SG:#DAT	
<CS1>	NP-SUBJ-EXP	V	NP-OBJ-THEME	
<CS2>	S-MAIN			

(76) German

<WORDS>	Ich	will	essen	Nudeln	.
<GLOSS>	1.SG.NOM	want:3.SG	eat:INF	spaghetti-ACC.PL	
<CS1>	NP-SUBJ-EXP		V	NP-OBJ-THEME	
<CS2>	#S-MAIN				

(77) German

<WORDS>	Er	trinkt	Bier	ein	.
<GLOSS>	3.SG.NOM	drink:3.SG	beer[ACC.SG.N]	DEF:ACC.SG.N	
<CS1>	NP-SUBJ-EXP	V	#NP-OBJ-THEME		
<CS2>	S-MAIN				

Often it might be difficult to know for sure what the intended utterance would have been. If it is not obvious to the annotator how to reconstruct the grammatical, intended utterance, only grammatical fragments of the sentence are annotated as usually, whereas questionable fragments are marked by “#” , to

mark their ungrammaticality. Note that no constituents dominating such questionable fragments are annotated, i.e., there is no “S” annotation in (78).

(78) German

<WORDS>	Er	denkt	Bier	ein	.
<GLOSS>	3.SG.NOM	thinks:3.SG	beer[ACC.SG.N]	DEF:ACC.SG.N	
<CS1>	NP-SUBJ-EXP	V	#	#	

7 References

Albert, Stefanie et al. 2003. *TIGER Annotationsschema*. Draft. Univerisities of Saarbrücken, Stuttgart, and Potsdam.

Poesio, Massimo. 2000. T*he GNOME Annotation Scheme Manual*. http://cswww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm.

Santorini, Beatrice 1990. *Annotation Manual for the Penn Treebank Project*. Technical Report, University of Pennsylvania.

Stegmann, R., H.Telljohann, and E. W. Hinrichs. 2000. *Stylebook for the German Treebank in VERBMOBIL*. Technical Report 239. Verbmobil.

Semantics

Cornelia Endriss[1], *Stefan Hinterwimmer*[2], *Stavros Skopeteas*[1]
University of Potsdam ([1]) and Humboldt University Berlin ([2])

The guidelines for semantics comprise a number of layers related to quantificational structures as well as some crucial semantic properties of NPs with respect to information structure: definiteness, countability, and animacy.

1 Preliminaries

Those features that are decisive for the semantic interpretation of a sentence have to be represented. We assume that syntactic annotation has already taken place so that all relevant syntactic features which are also interesting for the semantic level are explicit already.

The present guidelines were developed for annotating elements that occur in a corpus text. Elements that do not form part of the archived data, but arise from the analysis of the data (as in the case of ellipsis, traces, etc.) are not supported in the current version.

2 Layer Declaration

Table 1: Layers

Layer	Name
Quantificational properties	QuP
Interpretation of adverbially quantified structures	IN_adv

Interdisciplinary Studies on Information Structure 07 (2007): 133–143
Dipper, S., M. Götze, and S. Skopeteas (eds.):
Information Structure in Cross-Linguistic Corpora

Interpretation of possibly ambiguous quantified structures	IN_quant
Definiteness properties	DefP
Countability	C
Animacy	A

Table 2: Tagset declaration

Layer	Tags	Short Description
Quantificational properties (QuP)	ALL	universal quantifier
	EXIST	existential quantifier
	GEN	generic quantifier
	NUM	numerals
	Q	other quantifier
Interpretation of adverbially quantified structures (IN_adv)	N	nucleus
	QADV	quantificational adverbial
	R	restrictor
Interpretation of possibly ambiguous quantified structures (IN_scope)	ALL	universal quantifier
	EXIST	existential quantifier
Definiteness properties (DefP)	GEN	generic
	SP	specific
	U	unique
	USP	unspecific
Countability (C)	C	count
	M	mass
Animacy (A)	A	animate, non-human
	H	animate, human
	I	inanimate
	IA	unclear if animate or inanimate

3 Layer I: Quantificational properties (QuP)

3.1 Preliminaries

This layer deals with the annotation of quantificational elements and the resulting interpretation of those parts or the sentences containing those parts.

3.2 Tagset declaration

The semantic annotation has to enable queries combining quantificational and syntactic properties, e.g. “search for quantificational DPs”; “search for existential adverbs”, etc.

Assuming that syntactic information is provided by the layers that describe constituent structure, the semantic annotation contains the following labels:

Table 2: Tags for quantificational properties

tag	description	markable
ALL	universal quantifier	quantificational NPs/adverbials
EXIST	existential quantifier	quantificational NPs/adverbials
GEN	generic quantifier	covert operator
Q	other quantifiers	quantificational NPs/adverbials

3.3 Illustrative examples

If we assume an annotation layer ‘CS’ that displays constituent structures:

(1) English

<WORDS>	every	girl	likes	some	horse
<CS>	NP			NP	
<QuP>	ALL			EXIST	

(2) English

<WORDS>	dogs	always	have	green	eyes
<CS>	NP	ADV	VP		
<QuP>		ALL			

(3) English

<WORDS>	no	one	saw	three	horses
<CS>	NP			NP	
<QuP>	Q			Q	

A covert generic operator should be annotated whenever a sentence gets a generic interpretation. This can be tested in following way: Whenever *always/generally* can be inserted without changing the intended interpretation, a generic covert operator can be assumed:

(4) English

<WORDS>	a	dog		has	a	tail
<CS>	NP		ADV	VP		
<QuP>			GEN			

(5) English

<WORDS>	dogs		have	tails
<CS>	NP	ADV		
<QuP>		GEN		

This should not be confused with the existential interpretation that bare plurals often get:

(6) English

<WORDS>	dogs	were	sleeping
<CS>	NP	VP	
<QuP>	EXIST		

4 Layer II: Interpretation of adverbially quantified structures (IN_ADV)

4.1 Preliminaries

This layer deals with the annotation of the relation of restrictor and nucleus in sentences with quantificational adverbials.

4.2 Tagset declaration

Table 3: Tags for interpretation adverbially quantified structures

tag	description	markable
N	nucleus	part of sentences with Q-Adverbs
QADV	quantificational adv	adverbial
R	restrictor	part of sentences with Q-Adverbs

4.3 Illustrative example: Adverbially quantified structures interpretation

(7) English

<WORDS>	dogs	always	have	green	eyes
<IN_adv>	R	QADV	N		

5 Layer III: Interpretation of possibly ambiguous quantified structures (IN_scope)

5.1 Preliminaries

This layer deals with the annotation of the interpretation of quantificational elements, i.e. the scope of DP-quantifiers.

5.2 Tagset declaration

The units to be annotated are possibly ambiguous sentences that contain quantificational elements. The possible reading(s) of these sentences should be marked in the <IN_scope> field.

Table 4: Tags for interpretation of possibly ambiguous quantified structures

tag	description	markable
ALL	universal quantifier	quantificational NPs/adverbials
EXIST	existential quantifier	quantificational NPs/adverbials
GEN	generic quantifier	covert operator
Q	other quantifiers	quantificational NPs/adverbials
>	has scope over	sentences

5.3 Illustrative example: Scope interpretation

(8) English

<WORDS>	every	girl	likes	some	horse
<IN_scope>	ALL> EXIST; EXIST>ALL				

6 Layer IV: Definiteness properties (DefP)

6.1 Preliminaries

This layer contains information about definiteness. Definiteness encoded (e.g. through articles) is given in the morphemic translation (of the article).

6.2 Tagset Declaration

Table 5: Tags for definiteness properties

tag	description	markable
GEN	generic	NP (Indefinites/Definites)
SP	specific	NP (Indefinites)
U	unique	NP (Definites)
USP	unspecific	NP (Indefinites)

6.3 Instructions and illustrative examples

Annotate as definite:

- definite articles: *the*
- demonstratives: *this*
- possessives: *your horse, his book*

Annotate as indefinite:

- indefinite articles: *a*

(9) English

<WORDS>	Peter	is	looking	for	some	horse
<DefP>						USP

Test for unspecificity: The respective sentence could e.g. be followed by *And it does not matter which one.*

(10) English

<WORDS>	Peter	is	looking	for	some	horse
<DefP>					SP	

Test for specificity: The respective sentence could e.g. be followed by *But he has not found it yet.*

Kind Interpretations (Note the difference to generic interpretations that should be annotated as originating from a covert operator, cf. Section 2):

(11) English

<WORDS>	der Dinosaurier / Dinosaurier	ist / sind	ausgestorben
<DefP>	GEN		

The markables are DPs (and not single definite/indefinite markers): if more definite and indefinite markers occur in the same DP, only the resulting definiteness is annotated:

(12) English

<WORDS>	the	three	cowboys
<DefP>	U		

DPs can, of course, be stacked and should be annotated as such. (In Exmeralda, this can only be done by supplying for multiple DefP layers):

(13) English

<WORDS>	the	mother	of	the	boys
<DefP1>	U			U	
<DefP2>	U				

The respective DP should only be annotated as unique if the text allows us to conclude that the object denoted by the DP is the only object for which the property described by the corresponding NP holds. (In the literature it is sometimes claimed that non-unique definites exist.)

7 Layer V: Countability (C)

At this layer we encode information concerning the entity type (count/mass). The markables for this information are nouns. Nouns/DPs that turn up as part of sayings do not have to be annotated. (Looking at German, there are many

sayings or phrases that combine with bare singulars such as “in Frage stellen”, which regularly do not exist in German. These should not be annotated.)

7.1 Tagset Declaration

The following abbreviations are used for the annotation of the count/mass property of the noun/NP.

Table 6: Tags for countability

tag	description	markable
C	count	noun/DP
M	mass	noun/DP

7.2 Instructions and illustrative examples

(14) English

<WORDS>	cat
<C>	C

(15) English

<WORDS>	milk
<C>	M

8 Layer VI: Animacy (A)

8.1 Preliminaries

At this layer we encode information concerning the animacy. Since this annotation layer will be especially interesting for corpus studies concerning the impact of animacy on word order, topicality, and related issues, we adopt a rather detailed classification, so that users of the database are able to specify in

their queries which kind of NPs they want to count as animates or inanimates. The markables for animacy are nouns (both proper or common nouns).

8.2 Tagset Declaration

The following abbreviations are used for the annotation of the animacy property of the noun.

Table 7: Tags for animacy

tag	description	markable
A	animate, non-human	noun
H	animate, human	noun
I	inanimate	noun
IA	inanimate/animate	noun

8.3 Instructions and illustrative examples

Clear instances of human beings are annotated as ‘h’:

(16) English

<WORDS>	woman
<A>	H

Clear instances of non-human animates are annotated as ‘a’:

(17) English

<WORDS>	cat
<A>	A

Clear instances of inanimates are annotated as ‘i’:

(18) English

<WORDS>	milk
<A>	I

The following categories concern types of entities that are not clear instances of the above categories. Since it depends on the criteria of a certain study whether each of these categories should be treated as animate or inanimate or if it should simply be excluded from the query, we recommend grouping these in the remaining category IA:

- body parts:

(19) English

<WORDS>	hand
<A>	IA

- non-humans with human-like properties. These referents are not humans, but they may have similar properties to humans in several respects (agenthood, shape, motion) and may be treated like humans in certain languages.

(20) English

<WORDS>	robot
<A>	IA

Information structure

Michael Götze[1], *Thomas Weskott*[1], *Cornelia Endriss*[1], *Ines Fiedler*[2], *Stefan Hinterwimmer*[2], *Svetlana Petrova*[2], *Anne Schwarz*[2], *Stavros Skopeteas*[1], *Ruben Stoel*[1]
Second Edition: Julia Ritz[1], *Philippa Cook*[3], *Felix Bildhauer*[3], *Malte Zimmermann*[1], *Stefan Hinterwimmer*[2], *Sophie Repp*[2], *Amir Zeldes*[2], *Stefan Müller*[3], *Manfred Stede*[1], *Sören Schalowski*[1], *Marc Reznicek*[2]

University of Potsdam([1]) and Humboldt University Berlin([2]), Freie Universität Berlin ([3])

1 Preliminaries

These guidelines are designed for the annotation of information structural features in typologically diverse languages. The main objectives of these guidelines are i) language independence, ii) openness towards different theories, and iii) reliability of annotation.

These objectives resulted in a number of decisions that were implicitly made in the guidelines, the most relevant being the following:

- Annotation instructions rely mainly on functional tests, rather than tests involving linguistic form.
- Possibly different dimensions of information structure are annotated independently from each other, postulating no relation between these different features (as one could do e.g. for topic and focus).
- Most tagsets offer an obligatory tagset (or 'Core Annotation Scheme') and a tagset with optional tags (or 'Extended Annotation Scheme'), where the Core Annotation Scheme enables a more reliable and quick annotation and the Extended Annotation Scheme offers more detailed descriptions of

the data.

The guidelines are structured as follows: in the next sections, annotation instructions for three different dimensions of information structure, Information Status (Section 3), Topic (Section 4), Focus (Section 5) and Contrast (Section 6) are provided. In Section 7, an annotation procedure is proposed and described.

2 Tagset Declaration

2.1 Core Annotation Scheme for Information Structure

Table 1: Tags of the Core Annotation Scheme for Information Structure

Layer	Tags	Short description
Information Status	giv	given
	acc	accessible
	new	new
	cat	cataphor
	nil	non-referential
Topic	ab	aboutness topic
	fs	frame setting topic
Focus	nf	new-information-focus
Contrast	ctr	contrast (linking of contrastive pair where applicable)

2.2 Extended Annotation Scheme for Information Structure

Table 2: Tags of the Extended Annotation Scheme for Information Structure

Layer	Tags	Short description

Information Status	giv	given (underspecified)
	giv-active	active
	giv-inactive	inactive
	acc	accessible (underspecified)
	acc-sit	situationally accessible
	acc-aggr	aggregation
	acc-abs	abstract resumption (event/situation/proposition)
	acc-inf	inferable
	acc-gen	general
	new	new
	cat	cataphor
	nil	non-referential (underspecified)
	nil-expl	expletive
	nil-idiom	idiom/metaphor
	nil-pred	predication/attribution
Topic	ab	aboutness topic
	fs	frame setting topic
Focus	nf	new-information-focus (underspecified)
	nf-sol	Solicited new-information focus
	nf-unsol	unsolicited new-information focus
Contrast	ctr	contrastive focus (underspecified)
	ctr-repl	replacing
	ctr-sel	selection
	ctr-part	partiality
	ctr-impl	implication

	ctr-ver	truth-value (verum)
Note:	…+op	All kinds of foci given above can occur as bound by focus operators like the particles *only*, *even*, *also etc.* as well as negation operators. In this case, the tags are supplied with the additional marking *+op* (cf. 4.5).

3 Layer I: Information Status

3.1 Introduction

The objective of this annotation layer is to annotate discourse referents for their information status in the discourse, cf. Prince (1992). The figure below indicates how the concepts of information status and discourse status are related:

information status	giv	acc	new
discourse status	discourse-old	discourse-new	discourse-new
hearer status	hearer-old	hearer-old	hearer-new

“Discourse referents” are meant to comprise entities of many different types, that is individuals, places, times, events and situations, and sometimes even propositions. All these can be picked up by anaphoric expressions.

Their information status[17] reflects their "retrievability", which is meant to be understood as the difficulty of accessing the antecedent referent: a referent mentioned in the last sentence is easily accessible or "given", whereas one that has to be inferred from world knowledge is only "accessible" to the degree that the inference relation is shared between speaker and hearer. A discourse referent which lacks an antecedent in the previous discourse, isn't part of the discourse situation, nor is accessible via some relational reasoning has to be assumed to be "new".

The annotation scheme for information status proposed here consists of 1) a core annotation scheme for the obligatory tags ('giv', 'acc', 'new', 'cat', 'nil'), 2) an extended annotation scheme for optional tags ('giv', 'giv-active', 'giv-inactive', 'acc', 'acc-sit', 'acc-aggr', 'acc-abs', 'acc-inf', 'acc-gen', 'new', 'cat', 'nil', 'nil-expl', 'nil-idiom', 'nil-pred'), and 3) a recommended annotation procedure.[18]

This section is structured as follows: after the tagset declaration, instructions for annotating information structure are provided. In the last section, a procedure for applying these instructions is recommended.

[17] Related and widely used terms are 'activation', 'retrievability', 'cognitive status', 'givenness', etc.

[18] Many principles of this annotation scheme are closely related to Nissim et al. 2004. A more detailed discussion of the annotation scheme will follow.

3.1.1 Tagset Declaration

Table 3: Information status tags

Annotation layer:	Information Status	
Description:	Information status (“activation”) of the discourse referents	
Unit:	Any NP or pronoun (for exceptions, see Section 3.2.1 Units of interest).	
Core Annotation Scheme		
Tags:	giv	given
	acc	accessible
	new	new
	cat	cataphor
	nil	non-referential

Extended Annotation Scheme		
Tags:	giv	given (underspecified)
	giv-active	active
	giv-inactive	inactive
	acc	accessible (underspecified)
	acc-sit	situationally accessible
	acc-aggr	aggregation
	acc-abs	abstract resumption
	acc-inf	inferable
	acc-gen	general
	new	new
	cat	cataphor
	nil	non-referential (undersp.)
	nil-expl	expletive
	nil-idiom	idiom/metaphor
	nil-pred	predication/attribution

3.2 Instructions for Annotating Information Status

In this section, instructions for annotating information status are provided. A procedure for applying these instructions can be found in the next section.

For annotating according to the ‘Core Annotation Scheme’, the sections 3.2.1 Units of interest (markables), 3.2.2 Referring vs. non-referential (nil) expressions, 3.2.4 Given (giv), 3.2.6 Acc (acc), 3.2.8 New (new), and 3.2.9 Cataphor (cat) are relevant. However, the examples in the remaining sections might be helpful as well. For annotating according to the ‘Extended Annotation Scheme’, all sections have to be considered.

3.2.1 Units of interest (markables)

This chapter defines the units that are to be annotated (=‘markables’); in the following, they will be denoted by square brackets: [my markable].

Markables are defined by NPs and pronouns. Reflexive pronouns constitute markables, relative pronouns[19] do not.

Names also form markables. A name includes first, last, and middle names, initials, titles, and name affixes (“[Dr. Martin Luther King, Jr.]”, “[Louis ‘Satchmo’ Armstrong]”, “[A.B. Smith, M.D.]”, “[Chocolate Ltd.]”). Should the name include other names, then these also form markables (e.g. “[Cadbury Schweppes [Canada]]”).

In general, complex NPs form more than one markable:

- embedded NPs and appositions introduce extra markables (“[[Picasso’s] [picture of [a bowl of [fruit]]]]”, “[[Angela Merkel], [the German chancellor]] , …”,).
- Relative clauses are considered to belong to the same markable as the head they refer to, irrespective of the restrictive/non-restrictive distinction or possible extraposition.
- NPs with more than one head, e.g. conjunctions, do not form a common markable. Instead, the individual elements are annotated, e.g. “[the boys] and [the girls]”. The only exception is when the annotation would separate the head from determiners or modifiers (correct annotation: “[the young boys and girls]”).

Fragmentary material should also be considered for annotation, in particular truncated and elliptic constructions (“They cleaned [**the in-**] and painted [the outside].”, “I planned for three hours but finished in [**two**].”).

[19] Especially in German, relative pronouns are not to be confused with pronouns in sentences with left dislocation e.g. “Peter, [der] wohnt hier.” (‘Peter, [he] lives here.’).

Finally, there are some language specific adaptions: If a language has contracted forms, e.g. preposition+definite determiner in French and German, the markable must include the definiteness information: “[Zur Entscheidung] sind sie auch heute nicht gekommen” (‚Even today, they haven’t come [to.DEF decision].’), “Elle l’a présenté au festival.” (‘She presented it [at.DEF festival].’). Analogously, in the case of a contraction of preposition+pronoun (e.g. German *darin* ‘therein’), these complex pronouns form markables. (Non-contracted forms and discourse anaphors such as *there*, *then*, do not form markables.)
If a language allows for discontinuous constituents, a markable should be created that contains all parts of this constituent.

3.2.2 Referring vs. non-referential (nil) expressions

- Linguistic expressions can refer to e.g. objects, sets, activities, properties, types, concepts, states, events, situations, or propositions.Expressions that don’t refer to discourse referents are assigned the label *nil*.Examples for expressions that don’t refer in this sense are(not a markable)
- expletives (“it”, as in “**It** always rains on Sundays.”)
- interrogative elements (e.g. “who”, “which train”, etc.)
- or (parts of) idiomatic phrases such as “on (**the other hand**)”, “for (**some reason**)”, “as (**a result**)” or of metaphors (“language is **a virus**”). Further examples are given in (1) and (2).
- predications or appositions (“Barack Obama is **the president of the United States**”, “He is considered **a moderate**”, or “Elisabeth Blackwell, **the first woman physician,** …”, “Paediatricians, **specialists in child medicine**, …”)

Negatively quantified expressions (‘no man’, ‘neither of them’) and expressions in negated sentences are referential only if they can be resumed using a pronoun

(referring examples: “I’ve never seen **snow**, but I like it.”, “**No cellphones** on the ice, keep them in your bag.”, “**Unicorns** do not exist, but they have been sought for.”, non-referring example: “He didn’t wear **a hat**, *because it was too big for him.”).

(1)

<WORDS>	Peter	kicked	the	bucket	.
<CS>	NP		NP		
<INFOSTAT>	new		nil		

(2)

<WORDS>	Hans	warf	die	Flinte	ins	Korn	.
<CS>	NP		NP			NP	
<INFOSTAT>	new		nil		nil		
<TRANS>	Hans threw the rifle into the cornfield. (= Hans threw in the towel.)						

Note that abstract nouns (e.g. “quality”, “trust”, “danger”) and bare plurals (“creative ideas”, “price trends”) refer, if it is possible to resume them using a pronoun.

3.2.3 Extended Annotation Scheme: Subcategories of non-referential (*nil*)

The expression does not discribe a referent (object, set, activity, property, type, concept, state, event, situation, or proposition).

- *Note:* If you annotate tags at this layer, be as specific as possible. Only if you are not sure about which sub-tag (‘nil-expl’, ‘nil-idiom’ or ‘nil-pred’) to choose, choose the less specific tag, i.e. ‘nil’.

Expletives (*nil-expl*)

Expletives fill a syntactic role without contributing to the meaning of the sentence, as in “I have potted some lavender and [**it**] seems like it is dying.” (see also Example (3)).

Idiom (*nil-idiom*)

NPs/PPs in idioms are characterised by the fact that they lose most of their original meaning and usually cannot be referred back to (Examples (1) to (3)).
Typical tests for idioms are

- the translation test: If a phrase cannot be translated by its literal meaning, it should be annotated as ‘nil-idiom’. Some idioms analogously exist in several languages, e.g. ‘to sit in the same boat’; nevertheless, they are listed as a whole phrase in a dictionary, and thus count as idioms.
- The modification test: If a phrase cannot be felicitously modified (e.g. by an adjective or PP), it is likely to be idiomatic. (“Peter threw in the new towel”).

The label ‘nil-idiom’ is also used for metaphoric uses of NPs/PPs (“People died like **flies**.”).
If an idiomatic or metaphoric expression is referred back to, however, its first mention should be labelled ‘new’ (e.g. “Hit [**the books**] hard, and keep hitting them for another year.”).

Predication / Apposition (*nil-pred*)

Predications ascribe a certain property to a referent (“Peter is **an expert**.”). They can occur e.g. as appositions, or with verbs like ‘be’, ‘become’, ‘seem’, etc., in passive constructions like ‘was elected/announced’, or in constructions with ‘as’ (Example (3)). Predications can be made across different tenses or modalities, and might make use of names (“Lady Gaga has been called [a warmed-over 80’s wannabe], but in fact, she might be [the new Madonna].”).

(3)

<WORDS>	It	was	Tom's	first	day	as	a	teacher	.
<CS>	NP		NP	NP			NP		
				NP					
			NP						
<INFOSTAT>	nil-expl		new	new			nil-pred		
				new					
			new						

<WORDS>	It	was	a	monday	but	he	was	at	his	best	.
<CS>	NP		NP								
<INFOSTAT>	giv		nil-pred			giv			nil-idiom		

3.2.4 Given (*giv*)

The expression has an explicitly mentioned antecedent in the previous discourse: the referent has already been mentioned and is picked up again. In most cases, it is sufficient to check the preceding 5 sentences for an antecedent, but sometimes, anaphoric relations may stretch even across paragraphs.

- *IMPORTANT:* The referent must be referred to *explicitly* in the preceding discourse! That means that there must be expressions that refer to this discourse referent.
- bound pronouns are considered 'given' as well (e.g. "If [some patient] called for help, [**they**] would usually get it within 8 to 10 minutes.")

Note that referents of propositional type as in example (4) are considered accessible.

(3)

<WORDS>	Peter	went	into	the	garden	.	He	was	happy	.
<CS>	NP			NP			NP			
<INFOSTAT>	new			new			**giv**			

(4) context: ‘this cat’ has been mentioned before.

<WORDS>	Peter	liked	Tom	.	But	this	cat	wouldn’t	believe	that	.
<CS>	NP		NP			NP				NP	
<INFOSTAT>	new		new			**giv**				**acc-abs**	

3.2.5 Extended Annotation Scheme: Subcategories of given (*giv*)

We differentiate two subcategories of ‘giv’, ‘giv-active’ and ‘giv-inactive’.

Note: If you annotate tags at this layer, be as specific as possible. Only if you are not sure about which sub-tag (either ‘giv-active’ or ‘giv-inactive’) to choose, choose the less specific tag, i.e. ‘giv’.

Active (*giv-active*)

The referent was referred to *within the last or in the current sentence.*

(5)

<WORDS>	Peter	went	into	the	garden	.	It	was	blooming	.
<CS1>	NP			NP			NP			
<CS2>	S						S			
<INFOSTAT>	new			new			giv-active			

(6)

<WORDS>	Peter	liked	Tom	.	But	Maria	wouldn’t
<CS>	NP		NP			NP	
<INFOSTAT>	new		new			new	

<WORDS>	believe	that	.
<CS>		NP	
<INFOSTAT>		acc-abs	

(7)

<WORDS>	...	They	laughed	.
<CS>	...	NP		
<INFOSTAT>	...	giv-active		

<WORDS>	And	then	they	fought	each	other	again	.
<CS>			NP		NP			
<INFOSTAT>			giv-active		giv-active			

Inactive (*giv-inactive*)

The referent was referred to *before the last sentence.*

(8)

<WORDS>	Peter	went	into	the	garden	.
<CS>	NP			NP		
<INFOSTAT>	new			new		

<WORDS>	It	was	blooming.	Peter	was	happy.
<CS>	NP					
<INFOSTAT>	giv-active			giv-inactive		

(9)

<WORDS>	Peter	went	into	the	garden	.
<CS>	NP			NP		
<INFOSTAT>	new			new		

<WORDS>	It	was	blooming	.	He	was	happy	.
<CS>	NP							
<INFOSTAT>	giv-active				giv-inactive			

3.2.6 Accessible (*acc*)

The referent of the expression has not been mentioned, but is accessible via some kind of relation to a referent in the previous discourse, in the situative context, or the assumed world knowledge of the hearer, or a combination thereof. In particular, the referent should fulfil one of the criteria in the next section (Section 2.2.5).

E.g. in the example below, the NP “the flowers” refers to a part of the previously introduced discourse referent “the garden”.

(10)

<WORDS>	Peter	went	into	the garden	.	The flowers	blossomed	.
<CS>	NP			NP		NP		
<INFOSTAT>	new			new		**acc**		

(11)

<WORDS>	Could	you	pass	the sugar	,	please	?
<CS>		NP		NP			
<INFOSTAT>		**acc**		**acc**			
	(situative context)						

In example (11), both the adressed person and the sugar are part of the situative context of the communication.

(12)

<WORDS>	Peter	loves	violets	,	above	all	.
<CS>			NP				
<INFOSTAT>	giv		**acc**				
	(world knowledge)						

3.2.7 Extended Annotation Scheme: Subcategories of Accessible (*acc*)

The referent of the expression has not been mentioned, but is accessible via some kind of relational information, the situative context, or the assumed world knowledge of the hearer.

- *Note:* If you annotate tags at this layer, be as specific as possible. Only if you are not sure about which sub-tag (either ‘acc-sit’, ‘acc-aggr’, ‘acc-inf’ or ‘acc-gen’) to choose, choose the less specific tag, i.e. ‘acc’.

Situative (*acc-sit*)

The referent is part of the discourse situation.

(13)

<WORDS>	Could	you	pass	the sugar	,	please	?
<CS>		NP		NP			
<INFOSTAT>		acc-sit		acc-sit			
	(in dialogue during breakfast)						

(14)

<WORDS>	The	kid	hits	the	cow	.
<CS>	NP			NP		
<INFOSTAT>	acc-sit			acc-sit		
	(pointing with the finger at the figures in the book)					

Aggregation (*acc-aggr*)

The referring expression denotes a group consisting of accessible or given discourse referents.

(15)

<WORDS>	Peter	went	shopping	with	Maria	.	They	bought	many	flowers	.
<CS>	NP				NP		NP		NP		
<GIVEN>	new				new		acc-aggr		new		

(16) [otherwise identical to Example (15)]

<WORDS>	I	went	shopping	with	Maria	.
<CS>	NP				NP	
<GIVEN>	acc-sit				new	

<WORDS>	We	bought	many	flowers	.
<CS>	NP		NP		
<GIVEN>	acc-aggr		new		

Inferable (*acc-inf*)

Since reliably distinguishing various types of inferables[20] appears to be difficult (cf. Nissim et al. 2004), we restrict ourselves to identifying inferables as such and don't annotate their subtypes. However, we provide some types here as a help for recognizing various instances of inferables.

Assign 'acc-inf', if the referent is part of one of the following bridging relations:

- part-whole: The referent is in a part-whole relation to a referent in the preceding discourse.

(17) Context: 'The garden' has been mentioned before.

[20] or Bridging expressions.

<WORDS>	The	garden	is	beautiful	.	Its	entrance	is	just	across	this	river	.
<CS>	NP					NP					NP		
<GIVEN>	giv-act					acc-inf					acc-sit		

- set-rel: The referent is part of a set relation (i.e. subset, superset, member-of-the-same-set) to a referent in the preceding discourse.

(18) Context: ‘the flowers’ and ‘the garden’ have been mentioned at some distance

<WORDS>	The	flowers	in	the	garden	blossom	.
<CS>	NP			NP			
	NP						
<GIVEN>	giv-inactive			giv-inactive			
	giv-inactive						

<WORDS>	The	flowers	near	the	gate	blossom	violet	.
<CS>	NP			NP				
	NP							
<GIVEN>	acc-inf			acc-inf				
	acc-inf							

(19) ‘the children’ have been mentioned at some distance, the lake is inferred from the place of utterance.

<WORDS>	The	children	swam	in	the	lake	.
<CS>	NP				NP		
<GIVEN>	giv-inactive				acc-sit		

<WORDS>	The	familiy	experienced	a	beautiful	day	.
<CS>	NP			NP			
<GIVEN>	acc-inf			new			

- entity-attribute: The referent is constitutes an attribute of a referent in the preceding discourse.

(20) There are flowers on the table.

<WORDS>	The	flowers	enchanted	Peter	.	Their	scent	was	wonderful	.
<CS>	NP					NP				
<GIVEN>	acc-sit			giv-inactive		acc-inf				

NPs/PPs with a possessive determiner are always to be labelled ‘acc-inf’ (except for idiomatic expressions).

General (*acc-gen*)

The speaker can assume that the hearer knows the referent from his or her world knowledge. Note that the expression can take on different forms (i.e. indefinite, definite, or bare NP).

- gentype: The referent of the expression is a set or kind of objects. In a generic sentence, the other referents may as well be generic, like ‘children’ in example (21).

(21)

<WORDS>	The	lion	is	dangerous	,	when	she	has	children	.
<CS>	NP						NP		NP	
<GIVEN>	acc-gen						giv-active		acc-gen	

- gentoke: The referent of the expression is a unique object which is assumed to be part of world knowledge.

(22)

<WORDS>	The	sun	set	.	Pele	scored	his	second	goal	.
<CS>	NP				NP		NP			
<GIVEN>	acc-gen				acc-gen		acc-inf			

Abstract resumption (*acc-abs*)

A discourse referent is introduced to resume an event, state, situation, abstract object, or proposition described in the context.

An example of an event resumed is given in the following: “Weatherford International Inc. said it **canceled** plans for a prefered-stock swap. Weatherford said market conditions led to [**the cancellation**] of the planned exchange.”
An example of a proposition resumed is given in (4) and (6) above.

3.2.8 New (*new*)

The referent is new to the hearer and to the discourse.
(23)

<WORDS>	Peter	went	into	the	garden.	Another	man	appeared.
<CS>	NP			NP		NP		
<INFOSTAT>	new			new		new		

3.2.9 Cataphor (*cat*)

Cataphors are pronouns that need the subsequent context for their interpretation, e.g. “When [**he**] is in a hurry, James always forgets his watch.”.
NPs/PPs containing cataphoric possessives are NOT considered cataphors (“When [**his** beeper]$_{new}$ goes off, James has to run.”).

3.3 Annotation Procedure

Please follow the following steps for every NP or PP (or according pronouns and adverbials) in the discourse:
Q1: Is the expression referential?

- yes: go to *Q2*!
- no: label expression as **nil**!

If you annotate with the Extended Annotation Scheme:
Q1.1: Is the expression expletive?

- yes: label expression as **nil-expl**!
- no: go to *Q1.2*!

Q1.2: Is the expression idiomatic/metaphoric?

- yes: label expression as **nil-idiom**!
- no: go to *Q1.3*!

Q1.3 Is the expression a predication or attribution?

- yes: label expression as **nil-pred**!
- no: use the underspecified label **nil**!

Q2: Is the expression a cataphor?

- yes: label the expression as **cat**!
- no: go to *Q3*!

Q3: Has the referent been mentioned in the previous discourse?

- yes: label expression as **giv**!

  If you annotate with the Extended Annotation Scheme:

  Q3.1: Was the referent referred to within the last sentence?

  yes: label expression as **giv-active**

  no: label expression as **giv-inactive**
- no: go to *Q4*!

Q4: Is the referent a physical part of the utterance situation?

- yes: label expression as **acc**!

  If you annotate with the Extended Annotation Scheme:

  Label the expression as **acc-sit**!
- no: go to *Q5*!

Q5: Is the referent accessible (1) via some kind of relation to other referents in the previous discourse, (2) from assumed world knowledge, (3) by denoting a group consisting of accessible or given discourse referents, or (4) in that it resumes a matter described in the context?

- yes: go to *Q6*!
- no: label expression as **new**!

Q6: Does the referring expression denote a group consisting of accessible or given discourse referents?

- yes: label element as **acc**!

  If you annotate with the Extended Annotation Scheme:

  Label the expression as **acc-aggr**!
- no: go to *Q7*!

Q7: Is the referent inferable from a referent in the previous discourse by some relation as specified in Section 3.2.7 under 'Inferable (acc-inf)'?

- yes: label element as **acc**!

  If you annotate with the Extended Annotation Scheme:

  Label the expression as **acc-inf**!
- no: go to *Q8*!

Q8: Is the referent assumed to be inferable from assumed world knowledge?

- yes: label element as **acc**!

  If you annotate with the Extended Annotation Scheme:

  Label the expression as **acc-gen**!
- no: go to *Q9*!

Q9: Is the referent an abstract matter (event, state, situation, proposition, etc.) described in the context?

- yes: label element as **acc-abs**!
- no: go back to *Q1* and start all over again! You must have missed something.

4 Layer II: Topic

4.1 Introduction

In its current version, the annotation scheme for Topic consists solely of the Core Annotation Scheme.

4.1.1 Tagset Declaration

Table 3: Topic tags

Annotation Layer:	Topic	
Description:	Sentence or Clause topics	
Unit:	XP	
Core Annotation Scheme		
Tags:	**ab**	aboutness topic: > what the sentence is about
	fs	frame-setting topic > frame within which the main predication holds
Note:		Topics may be nested within a focus.

4.2 Core Annotation Scheme for Topic

Topics come in two varieties: aboutness topics and frame setting topics. The two categories are not exclusive, i. e. a sentence can have an aboutness topic as well as one or several frame setting topics.

Note that not all sentences have topics (see 4.2.1 below). In some languages topics are marked overtly (either by a morphological marker or by a designated position in the syntax), while in others, topics can be identified only indirectly, i. e. via clause-internal or contextual information.

Concerning complex sentences, choose the following strategy: check whether the whole sentence has an aboutness and/or a frame setting topic. Then check for each single finite clause contained within the complex sentence – with

the exception of restrictive relative clauses – whether it has an aboutness or a frame setting topic.

4.2.1 Topicless sentences

All-new or event sentences do not have a topic. They are answers to the questions *What happened? What's new?*

(The informant is shown a picture of a burning house, and is asked: What happens?)

(24)

<WORDS>	A	house	is	on	fire	.
<TOPIC>						

4.2.2 Aboutness Topic (*ab*)

The aboutness topic is the entity about which the sentence under discussion makes a predication. Our concept of topic is defined on sentence level, and differs from the concept of text topics.The only expressions that can denote aboutness topics are:

(i.) referential NPs (i. e. definite descriptions and proper names),

(ii.) indefinite NPs with specific and generic interpretations, and indefinites in adverbially quantified sentences that show Quantificational Variability Effects,

(iii.) bare plurals with generic interpretations, and bare plurals in adverbially quantified sentences that show Quantificational Variability Effects, and

(iv.) finite clauses denoting concrete facts about which the subsequent clause predicates (see below).

Note 1 (Specificity)

- Specificity can be tested as follows: If the respective indefinite can be

preceded by “a certain …” without forcing a different interpretation, it gets interpreted as a specific indefinite.

Note 2 (Genericity)

- Genericity can be tested as follows: If a sentence containing an indefinite or a bare plural is roughly equivalent to a universal quantification over the set of individuals that satisfy the respective NP-predicate, it is a generic sentence. Examples: (25a) below is roughly equivalent to (25b) and (26a) is roughly equivalent to (26b).

(25) a. *A dog* is smart.
b. *All dogs* are smart.

(26) a. *Cats* are snooty.
b. *All cats* are snooty.

Note 3 (Quantificational Variability Effects)

- Quantificational Variability Effects can be defined as follows: An adverbially quantified sentence that contains an indefinite NP or a bare plural is roughly equivalent to a sentence where the combination Q-adverb + indefinite NP/bare plural has been replaced by a quantificational NP with corresponding quantificational force. Examples: (27a) is roughly equivalent to (27b), and (28a) is roughly equivalent to (28b).

(27) a. *A dog* is *often* smart.
b. *Many dogs* are smart.

(28) a. *Cats* are *usually* snooty.
b. *Most cats* are snooty.

Quantificational NPs other than indefinites and other kinds of XPs can *never* be aboutness topics.

Whether an NP (with the exception of specifically interpreted indefinites) should be marked as the aboutness topic of a sentence can be tested in the following way:

Test for Aboutness Topics

An NP *X* is the aboutness topic of a sentence *S* containing *X* if

- *S* would be a natural continuation to the announcement *Let me tell you something about X*
- *S* would be a good answer to the question
  What about X?
- *S* could be naturally transformed into the sentence
  *Concerning X, S´*where *S´* differs from *S* only insofar as *X* has been replaced by a suitable pronoun.

Note that in the case of generic sentences and adverbially quantified sentences that contain singular indefinites, the first occurrence of *X* in the tests above must be replaced by a corresponding bare plural.
(See the examples below.)

Whether a specific indefinite should be marked as the aboutness topic of a sentence can be tested in the following way:

Test for Aboutness Topics for Specific Indefinites

A specific indefinite *X* is the aboutness topic of a sentence *S* containing *X* if the following transformation of *S* sounds natural:

- Within *S*, replace the indefinite article in *X* by *this* or *that*
- Transform the resulting sentence *S´* into *Concerning X, S´*.

(See example 33 below.)

(29) {The informant is shown a picture of a burning house, and is asked: What about the house?}

<WORDS>	The	house	is	on	fire	.
<TOPIC>	ab					

(30) {Yesterday I met Peter and Anne in London.}

<WORDS>	Peter	was	wearing	red	socks	.
<TOPIC>	ab					

Transforming S into “Concerning Peter, he is wearing red socks” or testing the sentence in the context “Let me tell you something about Peter” sounds natural.

(31) {A dog is often smart.}

<WORDS>	A	dog	is	often	smart	.
<TOPIC>	**ab**					

Transforming S into “Concerning dogs, a dog is often smart” or preposing “Let me tell you something about dogs” sounds natural.

(32) {Cats are snooty.}

<WORDS>	Cats	are	snooty	.
<TOPIC>	**ab**			

Transforming S into “Concerning cats, cats are snooty” or preposing “Let me tell you something about cats” sounds natural.

(33) German

<WORDS>	Einen	Hund	mag	Peter	wirklich	.
<GLOSS>	A/One-ACC	dog	likes	Peter	really	
<TOPIC>	**ab**					
<TRANS>	Peter really likes one/a certain dog.					

Specificity: “A dog” can be replaced by “A certain dog”. (Aboutness-) Topicality: S can be transformed into “Concerning a certain dog, Peter really likes that dog”.

(34)

<WORDS>	That	Maria	is	still	alive	is	pleasing	.
<TOPIC>		ab						
<TOPIC>	ab							

Transforming the matrix sentence S into “Concerning the fact that Maria is still alive, S” is possible. Concerning the subordinate clause S´, the proper name “Maria” is the aboutness topic of this clause, as this clause can be transformed into the sentence “Concerning Maria, she is still alive”.

Note that aboutness topics may be embedded: “Bereits 50 mg [der Substanz] können tödlich sein.” (’50 mg [of this substance] alone can be lethal.’).

4.2.3 Frame Setting (*fs*)

Frame setting topics constitute the frame within which the main predication of the respective sentence has to be interpreted. They often specify the time or the location at which the event/state denoted by the rest of the clause takes

place/holds. Temporal or locative PPs, adverbial phrases and subordinate clauses denoting (sets of) spatial or temporal locations are therefore typical frame setting topics crosslinguistically.

Note, however, that not every such phrase is a frame setting topic: Frame-setting topics are typically fronted, and the spatial or temporal locations denoted by them are often already part of the shared background of the discourse participants, or can at least be inferred easily.

Furthermore, fronted adverbials denoting domains against which the subsequently reported fact is to be evaluated can be frame setting topics, too (Typical examples are adverbs like *physically*, *mentally* etc. in sentences like *Physically, Peter is doing fine*).

In some languages (e.g. Chinese, Vietnamese) the choice is even wider: There, for example, constituents denoting supersets of the entities of which something is predicated in the subsequent clause can also be frame setting topics (see the Chinese example below). In languages like German and English, on the other hand, the same meaning can only be expressed by employing special constructions like *Concerning X, S,* or *As for X, S* (where X is the frame setting topic, and S the subsequent clause).

Note: In contrast to aboutness topics, with frame setting topics there is never a direct predication relation between the frame setting topic and the subsequent clause.

(35) Vietnamese

<WORDS>	Đi	chợ	Mỗi	Tuần	Tôi	đi	ba	lần	.
<GLOSS>	Go	market	Every	week	1.SG	go	three	time	
<TOPIC>	fs		fs		ab				
<TRANS>	As for going to the market, every week I go three times.								

(36) Manado Malay: {They told me she was waiting for me at my home.}

<WORDS>	Kita	pe	pulang	dia	so	Pigi	.
<GLOSS>	1.SG	POSS	come home	she			
<TOPIC>	fs			ab			
<TRANS>	When I came home, she had already left. (My coming home ...)						

(37) German

<WORDS>	Gestern	abend	haben	wir	Skat	gespielt	.
<GLOSS>	Yesterday	evening	have	we	Skat	played	
<TOPIC>	fs			ab			
<TRANS>	Yesterday evening, we played Skat.						

(38) German

<WORDS>	Körperlich	geht	es	Peter	sehr	gut	.
<GLOSS>	Physically	goes	it	Peter	very	well	
<TOPIC>	fs			ab			
<TRANS>	Physically, Peter is doing very well.						

(39) Chinese

<WORDS>	Yie.sheng	Dong.wu	Wo	zui	xi.huan	Shi	zi	.
<GLOSS>	Wild	animal	I	very	like	lion	Suffix	
<TOPIC>	fs		ab					
<TRANS>	Concerning wild animals, I really like lions.							

(40)

<WORDS>	In	Berlin	haben	die	Verhandlungspartner	…
<GLOSS>	In	Berlin	have	the	negotiating partners	
<TOPIC>	fs			ab		
<TRANS>	In Berlin, the negotiating partners did not pay attention to one rule.					

<WORDS>	…	eine	Regel	nicht	beachtet	.
<GLOSS>	…	one	rule	not	paid-attention-to	
<TOPIC>	…					
<TRANS>	In Berlin, the negotiating partners did not pay attention to one rule.					

4.3 Annotation Procedure

For every XP, procede as follows:

Q1: Is the current sentence an all-new/thetic/event-presenting sentence, i.e. a felicitous answer to the question *What has happened? What's new?*

- yes: label expression as **thetic** and go to *Q2*!
- no: label expression as **non-thetic** and go to *Q2*!

Q2: Perform the aboutness-topic test for all the XPs in the current sentence (*Concerning X, S'*, including adaptations for specific indefinites) or recall the results. How many XPs are aboutness-topic candidates in this sentence?

- 0: go on to next XP.
- 1 or more: label the XP as **ab** and go to *Q3*.

Q3: Assign the aboutness-topic candidate a rank within the sentence. 0 is reserved for XPs in non-thetic sentences with no other aboutness-topic candidates (the 'one-and-only' aboutness topic). Lower numbers mean preferred candidates, higher numbers mean less preferred candidates. In sentences that are possibly thetic (but nevertheless have constituents that pass the aboutness test), ranking starts with 1, as there is no candidate at the highest confidence level (which would be assigned 0).

In any case, go to *Q4*.

Q4: Is the XP a framesetting topic?

- yes: label the XP as fs and go to *Q5*.
- no: go on to next XP.

5 *Q5*: Assign the framesetting topic a rank. Assign 0 if the XP is the only framesetting topic in any analysis of the sentence. Otherwise assign 1.Layer III: Focus

5.1 Introduction

The annotation guidelines for Focus consist of a *Core Annotation Scheme* and an *Extended Annotation Scheme* which differ with respect to size and detailedness.

5.1.1 Tagset Declaration

Table 4: Focus tags

Annotation Layer:	Focus	
Definition:	That part of an expression which provides the most relevant information in a particular context as opposed to the (not so relevant) rest of information making up the *background* of the utterance. Typically, focus on a subexpression indicates that it is selected from possible alternatives that are either implicit or given explicitly, whereas the background can be derived from the context of the utterance.	
Unit:	Focus can extend over different domains in the utterance (like affixes, words, clause constituents, whole clause) and can be discontinuous as well. One expression can contain more than one focus.	
Core Annotation Scheme		
Tags:	nf	new-information focus
Extended Annotation Scheme		
Tags:	nf	new-information focus

	nf-sol	solicited new-information focus
	nf-unsol	unsolicited new-information focus
Note:	…+op	All kinds of foci given above can occur as bound by focus operators like the particles *only*, *even*, *also etc.* as well as negation operators. In this case, the tags are supplied with the additional marking +*op* (cf. 4.5).

5.1.2 Some preliminaries

The Core Annotation Scheme is designed for basic annotation of focus phenomena in large amounts of language data. It aims at high inter-annotator agreement.

We define the focus as that part of an utterance which provides new information and/or carries the discourse forward.

-

On the basis of the Core Annotation Scheme, further sub-types of focus can be distinguished as shown in the Extended Annotation Scheme.

5.2 New-information focus (*nf*)

5.2.1 Core Annotation Scheme

New-information focus (*nf*) is that part of the utterance providing the new and missing information which serves to develop the discourse.

(41)

<WORDS>	Who	is	reading	a	book	?
<NFocus>	nf					
<CFocus>						

<WORDS>	Mary	is	reading	a	book	.
<NFocus>	nf					
<CFocus>						

5.2.2 With the exception of single-word focus, annotate whole phrases (NPs etc.) instead of phrasal heads only. Auxiliaries and modals are only in the focus domain if they contribute to the focus, e.g. tense, aspect, or mood. In the example „Peter ist zu seiner Mutter gefahren. Er hat [die Blumen gegossen].“ (‘Peter has gone to his mother’s. He has [watered the flowers].’), the second sentence continues in the perfect. In „Peter ist zu seiner Mutter gefahren. Er [hatte die Blumen gegossen].“ (‘Peter has gone to his mother’s. He [had watered the flowers].’), the second sentence is in the past perfect.Extended Annotation Scheme: Subcategories of new-information focus (nf)

In defining the new-information focus domain of a sentence, we propose two strategies according to the major distinction between question-answer sequences and running texts. For these two cases, we use *nf-sol* and *nf-unsol* in the Extended Annotation Scheme, respectively.

Note: If you annotate tags at this layer, be as specific as possible. Only if you are not sure about which sub-tag (either *nf-sol* or *nf-unsol*) to choose, choose the less specific tag, i.e. *nf*.

Solicited new-information focus (*nf-sol*)

The *solicited new-information focus* is that part of a sentence that carries information explicitly requested by another discourse participant.

Comment: Note that the focus domain in the answer differs according to the information already presupposed by the question. The following examples illustrate this test for various focus domains.

- all-focus sentences: answers to questions like *"What's new?"*, *"What's going on?"*

(42)

<WORDS>	What	's	that	smell	?
<NFocus>	nf				
<CFocus>					

<WORDS>	The	kitchen	is	burning	.
<NFocus>	nf-sol				
<CFocus>					

Non-biased yes-no questions (also known as polar questions) and their answers are also cases of all-focus sentences since they are expressed to identify the truth-value of the entire proposition.

(43)

<WORDS>	Is	this	book	in	German	?
<NFocus>	nf					
<CFocus>						

<WORDS>	Yes	,	it	is	.
<NFocus>	nf-sol				
<CFocus>					

(44)

<WORDS>	Is	this	book	in	German	?
<NFocus>	nf					
<CFocus>						

<WORDS>	No	,	it	is	not	.
<NFocus>	nf-sol					
<CFocus>						

- VP-focus: extended over the whole VP of the answer:

(45)

<WORDS>	What	is	Mary	doing	?
<NFocus>	nf				
<CFocus>					

<WORDS>	She	is	reading	a	book	.
<NFocus>		nf-sol				
<CFocus>						

- narrow (XP-) focus: extended over one constituent or on a part of a constituent only

(46)

<WORDS>	Who	is	reading	a	book	?
<NFocus>	nf					
<CFocus>						

<WORDS>	Mary	is	reading	a	book	.
<NFocus>	nf-sol					
<CFocus>						

(47)

<WORDS>	What	is	Mary	reading	?
<NFocus>	nf				
<CFocus>					

<WORDS>	She	is	reading	a	book	.
<NFocus>				nf-sol		
<CFocus>						

(48)

<WORDS>	What	sort	of	books	does	Mary	read	?
<NFocus>	nf							
<CFocus>								

<WORDS>	She	reads	books	on	linguistics	.
<NFocus>				nf-sol		
<CFocus>						

- discontinuous focus domain: instances of discontinuous focus domains are given when a question is so explicit that it asks for two or more non-adjacent parts of an utterance. The index shows that the parts annotated for focus belong to one and the same focus domain that is interrupted by discourse-given material. This is useful to distinguish cases of discontinuous focus domains from those of multiple foci (cf. 4.4).

(49)

<WORDS>	What	did	Paul	do	with	the	book	?
<NFocus>	nf							
<CFocus>								

<WORDS>	He	gave	it	to	Mary	.
<NFocus>		nf_1		nf_1		
<CFocus>						

Unsolicited new-information focus (*nf-unsol*)

In running texts, for example in a narrative, report etc., the domain of *unsolicited new-information focus* extends over that part of the information that carries forward the discourse. It applies, for instance, to newly added discourse

referents, i.e. new individuals like persons, events, facts, states/qualities, time intervals and locations which can be refered to by pronouns in the following discourse. *Nf-unsol* further applies to new relations between given discourse referents, i.e. to all sorts of predicates: verbal and nominal predicates, quantificational determiners (*every, all, each, always, often* etc.).

In order to determine the domain of *nf-unsol*, we adopt a strategy already used for the identification of the focus domain in cases of question-answer sequences. We assume that for each sentence in a running text a preceding implicit question exists. That part of the sentence that supplies the new or missing information according to the implicit question is the information that carries the discourse further and has therefore to be annotated for *nf-unsol.*

Comment: Note that the domain of *nf-unsol* can also vary and be discontinuous as described for *nf-sol* above.

Text-initial sentences are usually all-focus sentences (also called presentational sentences which introduce new discourse referents). The entire initial sentence is annotated for focus.

With *non-initial sentences*, pay attention to the relation between given and newly established information, the latter being the domain of *nf-unsol*. In order to determine *nf-unsol*, try to formulate the *most general* question for each sentence on the basis of the given material, according to specific discourse types and the (probable) intention of the speaker to highlight that information which is able to develop the discourse.

The following is a sample annotation of *nf-unsol* in a narrative sequence:

(50) [1] Once upon a time, there was a wizard. [2] He lived in a beautiful castle. [3] All around the castle, there were green fields full of precious flowers. [4] One day, the wizard decided to leave his castle.

<WORDS>	Once	upon	a	time	there	was	a	wizard	.
<NFocus>	nf-unsol								
<CFocus>									
<FOCUS QUEST.>	no focus question possible / Who/What is the story going to be about?								

(51)

<WORDS>	He	lived	in	a	beautiful	castle	.
<NFocus>		nf-unsol					
<CFocus>							
<FOCUS QUEST.>	What about the wizard?						

In (51), questions like *"Where did he live?"* as well as *"What about his dwelling?"* are possible, too, but nevertheless they do not fit as a proper continuation of the discourse as established so far.

(52)

<WORDS>	All	around	the	castle	,	…
<NFocus>						…
<CFocus>						…
<FOCUS QUEST.>	What about the castle?					

<WORDS>	…	there	were	green	fields	full	of	precious	flowers	.
<NFocus>	…	nf-unsol								
<CFocus>	…									
<FOCUS QUEST.>	What about the castle?									

(53)

<WORDS>	One	day	,	the	wizard	decided	to	leave	his	castle	.
<NFocus>				nf-unsol							
<CFocus>											
<FOCUS QUEST.>	What happened then?										

Note that in (53), the role of the sentence in discourse structure plays a crucial role in formulating the focus question and assigning the domain of *nf-unsol*. As

the sentence in (53) opens a new paragraph, its function is similar to that of the text-initial sentence in (50). Consequently, "*the wizard*" – though mentioned before – belongs to the information necessary to complete the implicit question and is therefore part of *nf-unsol.*

5.3 Contrast

5.3.1 Core Annotation Scheme

We understand contrastive focus (*cf*) as that element of the sentence that evokes a notion of contrast to (an element of) another utterance.

(54) from OHG Tatian 229, 28 – 230, 01 (John 11, 9-10):

oba uuer gengit In tage / ni bispurnit. […] /[…] *oba her get In naht / bispurnit.* […] (If anyone walks in the day, he does not stumble […]. But if he walks in the night, he stumbles.)

<WORDS>	oba	uuer	Gengit	In	tage
<GLOSS>	if	anyone	Walks	in	day
<NFocus>					
<CFocus>				cf	
<TRANS>	If anyone walks in the day, …				

<WORDS>	oba	her	get	In	naht
<GLOSS>	if	he	walks	in	night
<NFocus>					
<CFocus>				cf	
<TRANS>	But if he walks in the night, …				

Contrastive focus may also extend over different domains of an utterance. In alternative questions and the answers to them it covers the whole CP, cf. (55).

(55)

<WORDS>	Is	it	raining	or	not	?
<NFocus>						
<CFocus>	cf				cf	

<WORDS>	Yes	,	it	is	.
<NFocus>			nf		
<CFocus>			cf		

In other cases, it will cover only a part of a lexical constituent, for example prefixes, the auxiliary part of analytical tense forms etc., cf. (56).

(56)

<WORDS>	We	do	not	export	but	import	goods	.
<MOPRH>	We	do	not	ex- port	but	im- port	goods	.
<NFocus>								
<CFocus>				cf		cf		

In case there is more than only one contrast in a sentence, an index is used to identify the contrasted pairs, cf. (57).

(57)

<WORDS>	Mary	likes	apples	but	Bill	prefers	strawberries	.
<NFocus>								
<CFocus>	cf_1		cf_2		cf_1		cf_2	

5.3.2 Extended Annotation Scheme: Subcategories of Contrastive Focus (cf)

Contrastive subtype replacing (*cf-repl*)

This subtype of contrastive focus corrects the contextually given information by replacing parts of it for suppletive information.

(58)

<WORDS>	I	heard	that	Mary	is	growing	vegetables	now	?
<NFocus>	nf-unsol								
<CFocus>									

<WORDS>	No	,	she	is	growing	bananas	.
<NFocus>							
<CFocus>						cf-repl	

Contrastive subtype selection (*cf-sel*)

An element out of a given set of explicitly expressed alternatives is selected. The classic instance of a selective focus is found in answers to alternative questions with *or*, as in the following example.

(59)

<WORDS>	Do you want to go	to	the	red	or	to	the	blue	house	?
<NFocus>										
<CFocus>				cf				cf		

<WORDS>	I	want	to	go	to	the	red	one	.
<NFocus>									
<CFocus>							cf-sel		

Contrastive subtype partiality (*cf-part*)

The *cf* introduces a (new) part or subset of a previously mentioned entity.

(60)

<WORDS>	What	are	your	sisters	doing	?
<NFocus>	nf					
<CFocus>						

(61)

<WORDS>	My	older	sister	works	as	a	secretary	,
<NFocus>				nf-sol				
<CFocus>		cf-part_1		cf_2				

<WORDS>	but	my	younger	sister	is	still	going	to	school	.
<NFocus>					nf-sol					
<CFocus>			cf-part_1		cf_2					

Contrastive subtype implication (*cf-impl*)

An utterance with this subtype of contrastive focus implies that the requested information holds true not for the information provided explicitly in the answer but for other alternatives that are accessible in the context.

(62)

<WORDS>	Where	is	the	weather-cock	?
<NFocus>	nf				
<CFocus>					

<WORDS>	Well	,	on	the	red	roof	,	there is no weather-cock	.
<NFocus>									
<CFocus>					cf-impl				

Here, the speaker implies that the weather-cock is on a roof other than the red one. Difference to *cf-part* is difficult. Pay attention that in *cf-part* the set of alternatives is explicitly given. For example, a question like "*Where on the roofs is the weather-cock?"* allows for *cf-part* in the answer because the set of alternatives, "*the roofs"*, is explicitly given.

Contrastive subtype: truth-value (verum) (*cf-ver*)

This subtype of contrastive focus emphasizes the truth-value of the proposition. The annotation domain for truth-value focus is the whole proposition. (Note: In the literature, it is common to mark only the focus exponent [here: did].)

(63) context:

A: The exam was difficult, nevertheless lots of students passed.

B: Yes, that's true. Lots of students did pass.

<WORDS>	Lots	of	students	did	pass	.
<NFocus>						
<CFocus>	cf-ver					

Comment: There are cases in which the truth-value of the proposition is set and emphasized at the same time.

(64)

<WORDS>	Nobody believed that	,	but	Mary	did	go	to	Berlin	.
<NFocus>				nf					
<CFocus>				cf-ver					

In this case the truth-value of the proposition that *Mary went to Berlin* which is open in the context is being specified and emphasized at the same time.

-

(65)

<WORDS>	Who	met	whom	?
<NFocus>	nf		nf	
<CFocus>				

<WORDS>	An	American	farmer	met	a	Canadian	farmer	.
<NFocus>	nf				nf			

-

(66)

<WORDS>	What	happened	?
<NFocus>	nf		
<CFocus>			

<WORDS>	An	American	farmer	met	a	Canadian	farmer	.
<NFocus>	nf							

- *cf* and *nf* can completely diverge from each other:

(68') (An adapted example from Jacobs 1991: 201f.)

The children left the remainings of their meals everywhere in the apartment. Mary is responsible for the dirt in the bedroom and John for that in the bathroom.

(68)

<WORDS>	And	who	has	eaten	in	the	living	room	?
<NFocus>		nf							
<CFocus>							cf		

5.4 Operator-bound focus (…+*op*)

All kinds of foci given above can occur as bound by focus operators like the particles *only, even, also etc.* as well as negation operators. Different focus association is also possible. In the cases given below, the focus operator *only* triggers two different foci.

(69a) (Rooth 1985)

<WORDS>	Mary	only	introduced	Bill	to	Sue	.
<CLASS>		foc-prt					
<NFocus>				nf+op			
<CFocus>							

(69b) (Rooth 1985)

<WORDS>	Mary	only	introduced	Bill	to	Sue	.
<CLASS>		foc-prt					
<NFocus>						nf+op	
<CFocus>							

5.5 Annotation Procedure

Please complete the following steps:

Q1: Is the sentence a declarative or a non-declarative one?

- if non-declarative (imperative, question): go to *Q3*
- if declarative: go to *Q2*

Q2: Does the utterance complete an explicit wh-question?

- Yes: the constituent which is congruent to the wh-word is to be annotated "*nf-sol*"
- No: go to *Q3*
-

Q5: Which part of the utterance reveals the new and most important information in discourse? Try to identify the domain by asking implicit questions as done in the example in 4.2.2!

- annotate the identified costituent or domain as "*nf-unsol*"
-

Q7: Does the sentence contain a focus operator?

- Yes: annotate the constituent that is bound by it for "*+op*"
- No: no additional specification is necessary

6 Recommended Annotation Procedure

(1) Preparation of the Data

Make sure that the data is prepared for the annotation with information structure. In particular, check for the annotation of sentences and NPs and PPs according to the Syntax Annotation Guidelines.

If the data is not annotated accordingly, do this annotation first!

(2) Annotation step 1: Information Status and Topic

Start from the beginning of the discourse.

For every sentence:

(a) Check for the referentiality of each NP and PP in the sentence (cf. Section 2.2.1).

(b) Specify the Information Status of every referring NP- and PP-marked constituent. Follow the instructions in 2.3.!

(c) Test for the Topic status of each NP and PP in the sentence. Follow the guidelines in Section 3!

(3) Annotation step 2: Focus

Start from the beginning of the discourse. For every sentence:

- Apply the annotation procedure for the Focus Annotation Scheme in Section 4.6.

(4) Check for Completeness

Check for the completeness of the Annotation:

(a) Check for the complete annotation of Information Status for all referring NPs and PPs.

(b) Check for the complete annotation of new-information focus: for each sentence a new-information focus should be assigned.

(5) Finishing the Annotation

Don't forget to save the annotation!

7 References

Nissim, M., Dingare, S., Carletta, J., and Steedman, M. 2004. An Annotation Scheme for Information Structure in Dialogue. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal, May.

Jacobs, Joachim. 1991. *Informationsstruktur und Grammatik*. Opladen: Westdeutscher Verlag.

Rooth, Mats. 1985. *Association with Focus*. Ph.D. dissertation, University of Massachusetts.

Prince, Ellen F. 1992. “The ZPG Letter: Subjects, Definiteness, and Information-status”. In Discourse Description, Mann, William C. and Sandra A. Thompson (eds.), 295 ff.

Appendix I: Annotation sample

This appendix presents a fully-annotated example for illustration of the presented guidelines. The transcribed text has been spontaneously elicited through the elicitation task “Fairy Tale” which is part of the *Questionnaire of Information Structure* which is a collaborative product of the project D2, SFB 632 (see http://www.sfb632.uni-potsdam.de/homes/d2/index.php).

Table 1: Annotation layers in the annotation sample

words	orthography
phones	SAMPA
stress	primary stress
accent	realized stress
php	PP
ip	IP
int-tones	ToBi transcription
morph	morphemic transcription
pos	part of speech
gloss	glossing
trans	free translation
cs1	constituent structure, first layer
cs2	constituent structure, second layer
infostat	information status
defp	definiteness
c	countability
a	animacy
topic	topic
focus	focus

Interdisciplinary Studies on Information Structure 07 (2007): 189–208
Dipper, S., M. Götze, and S. Skopeteas (eds.):
Information Structure in Cross-Linguistic Corpora

[1]

words	Heute		ist	mir	was	ganz	tolles		passiert		.
phones	heU	t@	Ist	mI6	vas	gans	tO	l@s	pa	sI6t	
stress	1						1			1	
accent						1	1				
php	PP										
ip	IP										
int-tones						L*+H	H*+	L			Li
morph	heute		ist	mir	was	ganz	toll-es		passiert		
pos	ADV		VAUX	PRONPRS	PRON	ADV	A		VINTR		
gloss	today		be:3.SG	1.SG.DAT	something: N.SG[NOM]	totally	fantastic -N.SG[NOM]		happen: PTCP.PRF		
trans	Something totally fantastic has happened today to me.										
cs1											
cs2											
cs3				NP-IO-EXP	NP-SUBJ-THEME				V		
cs4	S-MAIN										
infostat				ACC-SIT	NEW						
defp				U	SP						
c				C	C						
a				H	I						
topic	FS										
focus	NF-UNSOL										
focus											

[2]

words	Da		sollten			nämlich		,	nämlich		,	ähm	,	der	
phones	da:		zOl	t@n		nE:m	lIC		nE:m	lIC		E:m		dE6	
stress			1			1			1						
accent			1												
php	PP													PP	
ip	IP														
int-tones			H*												
morph	da		soll-t-en			nämlich			nämlich			ähm		der	
pos	PRONEXPL		VMOD			ADV			ADV					DET	
gloss	there		shall-PST-3.PL			namely			namely			hmm		DEF:M.SG.NOM	
trans	There should namely, namely... hmm...														
cs1														NP	
cs2														NP-SUBJ	
cs3	S-MAIN														
cs4	S-MAIN														
infostat														ACC-GEN	
defp														SP	
c														C	
a														H	
focus	NF-UNSOL														
focus															

[3]

words	Thomas		und	der	Ludwig		,
phones	to:	mas	Unt	dE6	lu:t	wIC	
stress	1				1		
accent							
php	[... PP]		PP				
ip	[... IP]						
int-tones	L+H*	Hp			L+H*	Hp	
morph	Thomas		und	der	Ludwig		
pos	NPRP		COOR	DET	NPRP		
gloss	Thomas:M.SG[NOM]		and	DEF:M.SG.NOM	Ludwig:M.SG[NOM]		
trans	Thomas and Ludwig...						
cs1	[... NP]			NP			
cs2	[... NP-SUBJ]						
cs3	[... S-MAIN]						
cs4	[... S-MAIN]						
infostat	[... ACC-GEN]			ACC-GEN			
defp	[... SP]			SP			
c	[... C]			C			
a	[... H]			H			
focus	[... NF-UNSOL]						
focus							

[4]

words	die		sollten			Tomaten			holen			gehen		und	die
phones	di:		sOl	t@n		to	ma:	t@n	ho:	l@n		ge:	h@n	Unt	di:
stress			1				1		1			1			
accent							1								
php	PP													PP	
ip	IP													IP	
int-tones						L+	H*					Hi			
morph	die		soll-t-en			Tomate-n			holl-en			geh-en		und	die
pos	PRONDEM		VMOD			NCOM			VTR			VINTR		and	DET
gloss	these [M.PL.NOM]		shall-PST-3.PL			tomato. F-PL[ACC]			bring-INF			go-INF		and	DEF [F.SG.NOM]
trans	they should go to bring tomatos and														
cs1						NP-OBJ-THEME			V						
cs2	NP-SUBJ-THEME					S-ARG						V			NP-SUBJ-AG
cs3	S-MAIN														S-MAIN
cs4	[... S-MAIN]														
infostat	GIV-ACTIVE					ACC-GEN									ACC-GEN
defp	SP					GEN									SP
c	C					C									C
a	H					I									H
topic	AB														
focus						NF-UNSOL									NF-UNSOL
focus															

[5]
[5]

words	Mama		hat	zuerst		den	Thomas	
phones	ma	ma	hat	tsu	E6st	de:n	to:	mas
stress	1				1		1	
accent	1						1	
php	[... PP]		PP					
ip	[... IP]							
int-tones	H*	Hp					H*	
morph	Mama		hat	zuerst		den	Thomas	
pos	NCOM		VAUX	ADV		DET	NPRP	
gloss	mam:F.SG[NOM]		have:3.SG	first		DEF:M.SG.ACC	Thomas:M.SG[ACC]	
trans	mam has sent first Thomas,							
cs1						NP-OBJ-THEME		
cs2	[... NP-SUBJ-AG]							
cs3	[... S-MAIN]							
cs4	[... S-MAIN]							
infostat	[... ACC-GEN]					GIV-ACTIVE		
defp	[... SP]					SP		
c	[... C]					C		
a	[... H]					H		
focus	[... NF-UNSOL]							
focus						cf-sel		

[6]

words	los	geschickt		und	der	ist	dann	los
phones	lo:s	g@	SIkt	Unt	dE6	Ist	dan	lo:s
stress			1					
accent								
php	[... PP]			PP				
ip	[... IP]			IP				
int-tones			Hi					
morph	los	geschickt		und	der	ist	dann	los
pos	ADV	VTR		COOR	PRONDEM	VAUX	ADV	ADV
gloss	off	send:PTCP.PRF		and	this:M.SG.NOM	be:3.SG	then	off
trans	and then he went off,							
cs1								
cs2		V			NP-SUBJ-THEME			
cs3	[... S-MAIN]				S-MAIN			
cs4	[... S-MAIN]							
infostat					GIV-ACTIVE			
defp					SP			
c					C			
a					H			
topic					AB			
focus						NF-UNSOL		
focus	[... cf-sel]							

[7]

words	gegangen			und	kam	aber		ohne		Tomaten			wieder		und	dann
phones	g@	gaN	@n	Unt	ka:m	a:	b6	o:	n@	to	ma:	t@n	wi:	d6	Unt	dan
stress		1				1		1			1		1			
accent		1						1								
php	[... PP]			PP											PP	
ip	[... IP]														IP	
int-tones		H*	Hp					H*						Hi		
morph	gegangen			und	kam	aber		ohne		Tomate-n			wieder		und	dann
pos	VINTR			COOR	VINTR	ADV		P		NCOM			ADV		COOR	ADV
gloss	go: PTCP.PRF			and	come. PST[3]SG	but		without		tomato: F-PL[ACC]			again		and	then
trans	but he came again without tomatos, and then															
cs1										NP-ARG						
cs2	V				V_1		PP-ADJ-MAN						_1			
cs3	[... S-MAIN]				S-MAIN											S-MAIN
cs4	[... S-MAIN]															
infostat										ACC-GEN						
defp										GEN						
c										C						
a										I						
focus	[...NF-UNSOL]				NF-UNSOL											NF-UNSOL
focus																

[8]

words	sollte		der	der	Ludwig		los	gehen		
phones	sOl	t@	dE6	dE6	lu:t	wIC	los	ge:	@n	
stress					1			1		
accent					1					
php	[... PP]									
ip	[... IP]									
int-tones					H*				Lp	
morph	soll-t-e		der	der	Ludwig		los	geh-en		
pos	VMOD		DET	DET	NCOM		ADV	VINTR		
gloss	shall-PST-3.SG		DEF:M.SG.NOM	DEF:M.SG.NOM	Ludwig:M.SG[NOM]		off	go-INF		
trans	Ludwig should go off,									
cs1				NP-SUBJ-THEME				V		
cs2										
cs3	[... S-MAIN]									
cs4	[... S-MAIN]									
infostat				GIV-INACTIVE						
defp				SP						
c				C						
a				H						
topic				AB						
focus	[... NF-UNSOL]									
focus				cf-sel						

[9]

words	und	dem	ist	dann	genau		dasselbe		
phones	Unt	de:m	Ist	dan	g@	naU	das	sEl	b@
stress						1		1	
accent								1	
php	PP								
ip	[... IP]								
int-tones								H*	
morph	und	dem	ist	dann	genau		das-selbe		
pos	COOR	PRONDEM	VAUX	ADV	ADV		PRON		
gloss	and	DEM:M.SG.DAT	be:3.SG	then	exactly		DEF:N.SG[NOM]-same[N.SG.NOM]		
trans	and exactly the same happened to him,								
cs1									
cs2		NP-IO-EXP					NP-SUBJ-THEME		
cs3		S-MAIN							
cs4	[... S-MAIN]								
infostat		GIV-ACTIVE					GIV-INACTIVE		
defp		SP					SP		
c		C					C		
a		H					I		
focus		NF-UNSOL							
focus									

[10]

words	passiert		und	dann	sollte		ich	los	gehen		und	
phones	pa	sI6t	Unt	dan	sOl	t@	IC	lo:s	ge:	@n	Unt	
stress		1			1				1			
accent							1					
php	[... PP]		PP								PP	
ip	[... IP]		IP									
int-tones		Hi					H*			Lp		
morph	passiert		und	dann	soll-t-e		ich	los	geh-en		und	
pos	VINTR		COOR	ADV	VMOD		PRONPRS	ADV	VINTR		COOR	
gloss	happen:PTCP.PRF		and	then	shall-PST-1.SG		1.SG.NOM	off	go-INF		and	
trans	and then I should go off, and											
cs1												
cs2	V						NP-SUBJ-THEME		V			
cs3				S-MAIN								
cs4	[...S-MAIN]											
infostat							GIV-INACTIVE					
defp							U					
a							H					
focus					NF-UNSOL							
focus												

[11]

words	ich	bin	in	die	Stadt	gegangen			und
phones	IC	bIn	In	di:	Stat	g@	gaN	@n	Unt
stress							1		
accent					1				
php	[... PP]								PP
ip	[... IP]								IP
int-tones					H*			Hi	
morph	ich	bin	in	die	Stadt	gegangen			und
pos	PRONPRS	VAUX	P	DET	NCOM	VINTR			COOR
gloss	1.SG.NOM	be:1.SG	in	DEF[F.SG.ACC]	city.F[SG.ACC]	go:PTCP.PRF			and
trans	I went to the city, and								
cs1				NP-ARG					
cs2	NP-SUBJ-THEME		PP-ARG-LOC			V			
cs3	S-MAIN								
cs4	[...S-MAIN]								
infostat	GIV-ACTIVE		ACC-GEN						
defp	U		SP						
c	C		C						
a	H		I						
topic	AB								
focus		NF-UNSOL							
focus									

[12]

words	habe		den	richtigen			Weg		gefunden			und
phones	ha:	b@	de:n	rIC	tI	g@n	ve:g		g@	fUn	d@n	Unt
stress	1			1						1		
accent				1			1					
php	[... PP]											PP
ip	[... IP]											IP
int-tones				H*			L*+H				Hi	
morph	hab-e		den	richtig-en			Weg		gefunden			und
pos	VAUX		DET	A			NCOM		VTR			COOR
gloss	have-1.SG		DEF:M.SG.ACC	right-[M.SG.ACC]			way.M[SG.ACC]		find:PTCP.PRF			and
trans	I found the right way, and											
cs1												
cs2			NP-DO-THEME						V			
cs3	S-MAIN											
cs4	[... S-MAIN]											
infostat			ACC-INF									
defp			SP									
c			C									
a			I									
focus	NF-UNSOL											
focus												

[13]

words	habe		Tomaten			mitgebracht			und	da	hat	sich
phones	ha:	b@	to	ma:	t@n	mIt	g@	bRaxt	Unt	da	hat	zIC
stress	1			1		1						
accent				1								
php	[... PP]								PP			
ip	[... IP]								IP			
int-tones			L+	H*				Li				
morph	hab-e		Tomat-en			mit-gebracht			und	da	hat	sich
pos	VAUX		NCOM			VTR			COOR	ADV	VAUX	PRONRFL
gloss	have-1.SG		tomato.F-PL[ACC]			together-bring:PTCP.PRF			and	there	have:3.SG	REFL.3.SG.ACC
trans	I have brought tomatos, and											
cs1												
cs2			NP-DO-THEME			V						NP-DO
cs3	S-MAIN									S-MAIN		
cs4	[...S-MAIN]											
infostat			GIV-INACTIVE									GIV-INACTIVE
defp			GEN									SP
c			C									C
a			I									A
focus	NF-UNSOL									NF-UNSOL		
focus												

[14]

words	die	Mama		sehr	drüber		gefreut	.
phones	di:	ma	ma	ze:6	dRy	b6	g@ fROIt	
stress		1		1		1		
accent		1						
php	[... PP]							
ip	[... IP]							
int-tones		H*						Li
morph	die	Mama		sehr	drüber		gefreut	
pos	DET	NCOM		ADV	ADV		VTR	
gloss	DEF: [F.SG.NOM]	mam. F[SG.NOM]		very	there:over		make.happy: PTCP.PRF	
trans	mam was very happy about that.							
cs1								
cs2		NP-SUBJ-EXP			PP-OBJ-THEME		V	
cs3	[...S-MAIN]							
cs4	[...S-MAIN]							
infostat		GIV-INACTIVE			GIV-ACTIVE			
defp		SP						
c		C						
a		A						
topic		AB						
focus	[... NF-UNSOL]							
focus								

Appendix II: Annotation guidelines tagset declarations

Section	Layer	Core tagset	Definition	Extended tagset	Definition
Phonology	info	(free)	relevant useful information		
	words	(free)	orthography or transliterated		
		<P>	pause		
	phones	IPA/SAMPA			
	stress	1	primary stress		
		2	secondary stress		
	accent	1	'realized stress'		
	php	PP		aPP	abstract PP
				rPP	realized PP
	ip	IP			
	lex-tones	(language specific)			
	int-tones	('ToBI' inventory)			
	surface	(language specific)			
	phontones	l	low unstressed syllable		
		L	low stressed syllable		
		h	high unstressed syllable		

Section	Layer	Core tagset	Definition	Extended tagset	Definition
		H	high stressed syllable		
		m	mid unstressed syllable		
		M	mid stressed syllable		
		i	end of Implementation Domain		
		-	interpolation		
Morphology	morph	<new cell>	word boundary		
		-	morpheme boundary		
		=	clitic boundary		
		_	union of sublexical components		
		0	zero affix		
	gloss	x:y	non-segmentable morphemes		
		x.y	semantic components		
		x_*n*	discontinuous morphemes		
		x/y	alternating meanings		
		{x}	non-realized in this context		
		[x]	non-overtly encoded		
		XXX	grammatical meaning		
			(see the list of abbreviations for glosses, e.g., SG, NOM, etc., see Table 4, page 80)		

Section	Layer	Core tagset	Definition	Extended tagset	Definition
	pos	A	adjective		
		ADV	adverb		
		CLF	classifier		
		COOR	coordinative conjunction		
		DET	determiner		
		N	noun	NCOM	common noun
				NPRP	proper noun
				VN	verbal noun
		P	preposition/postposition		
		PRON	pronoun	PRONDEM	demonstrative pronoun
				PRONEXPL	expletive pronoun
				PRONINT	interrogative pronoun
				PRONPOS	possessive pronoun
				PRONPRS	personal pronoun
				PRONQUANT	quantifier
				PRONREL	relative pronoun
				PRONRFL	reflexive pronoun
				AT	attributive
				SU	substantive
		PTC	particle		

Section	Layer	Core tagset	Definition	Extended tagset	Definition
		SUB	subordinative conjunction	SUBADV	adverbial subordinator
				SUBCOM	complementizer
		V	verb	VAUX	auxiliary verb
				VCOP	copula verb
				VDITR	ditransitive verb
				VINTR	intransitive verb
				VLEX	lexical verb
				VMOD	modal verb
				VN	verbal noun
				VTR	transitive verb
				CLIT	clitic form
				FULL	full form
Syntax	cs	AP	adjectival phrase		
		NP	noun phrase		
		PP	prepositional phrase		
		V	verbal head		
		S	sentence/clause		
	function	ADJ	adjunct		
		ADV	adverbial clause		

Section	Layer	Core tagset	Definition	Extended tagset	Definition
		ARG	argument	DO	direct object
				IO	indirect object
				OBJ	unspecified object
				SUBJ	subject
		ATTR	relative clause		
		MAIN	main clause		
		PRDNOM	predicate nominal		
	role	AG	agent		
		CAUSE	cause		
		COM	comitative		
		EXPER	experiencer		
		GOAL	goal		
		INSTR	instrument		
		LOC	location		
		MAN	manner		
		POSS	possessor		
		THEME	theme		
		TIME	time		
Semantics	QuP	All	universal quantifier		

Section	Layer	Core tagset	Definition	Extended tagset	Definition
		Exist	existential quantifier		
		Q	other quantifier		
		Gen	generic quantifier		
		Num	numerals		
	IN_adv	R	restrictor		
		N	nucleus		
		Qadv	quantificational adverb		
	IN_scope	All	universal quantifier		
		Exist	existential quantifier		
	DefP	Gen	generic		
		Sp	specific		
		U	unique		
		Usp	unspecific		
	C	C	count		
		M	mass		
	A	A	animate non-human		
		H	animate human		
		I	inanimate		
		IA	non classifiable		
IS	infostat	giv	given	giv-active	active

Section	Layer	Core tagset	Definition	Extended tagset	Definition
				giv-inactive	inactive
		acc	accessible	acc-sit	situationally accessible
				acc-aggr	aggregation
				acc-inf	inferable
		new	new	acc-gen	general
	topic	ab	aboutness topic		
		fs	frame-setting topic		
	focus	nf	new-information focus	nf-sol	solicited new-inf. focus
				nf-unsol	unsolicited new-inf. focus
		cf	contrastive focus	cf-repl	replacement
				cf-sel	selection
				cf-part	partiality
				cf-impl	implication
				cf-ver	truth value (verum)
		..+op	bound by a focus operator		