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): 1–27
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]

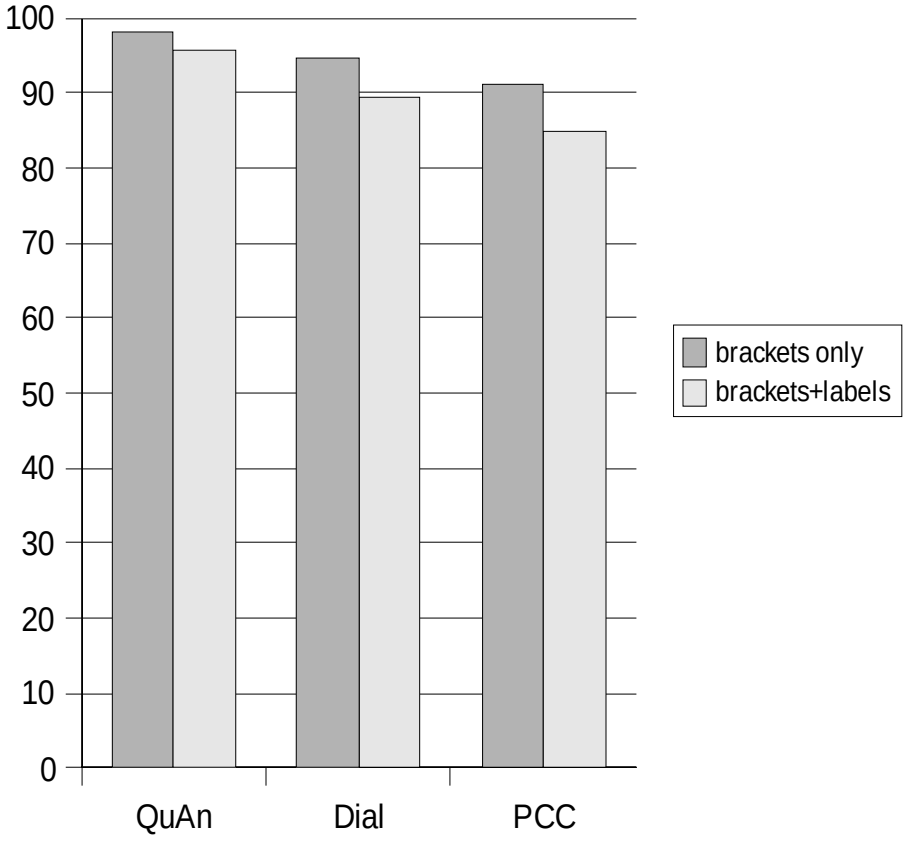
(4) $$matchRate(match_{A1,A2}) = \frac{2 \times \#sharedTokens(A1,A2)}{\#tokens(A1) + \#tokens(A2)}$$

[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)

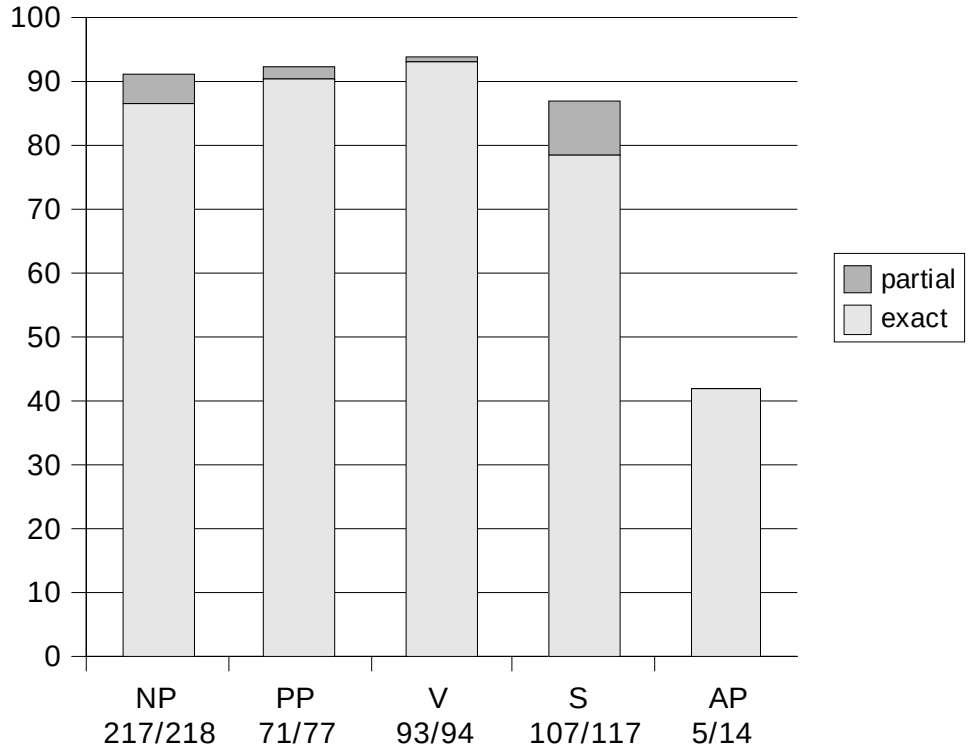


Figure 2 displays the results for use of individual labels within the PCC dataset.[6] For each category, we report the number of times it was used by each annotator (e.g., the label "NP" was used 217 times by one of the annotators, and 218 times by the other). The F-scores of NP, PP, and V are comparably 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.

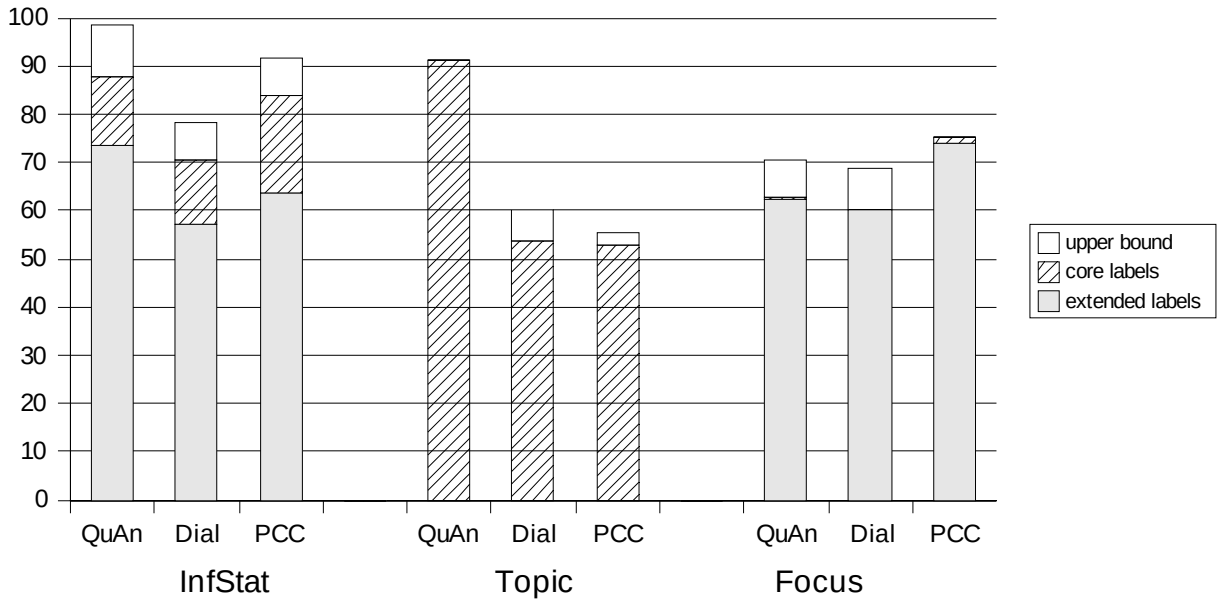
with an F-score of 42.11%, which can be attributed to the fact that one of the annotators found 14 APs and the other only 5. The top parts of the columns, which correspond to the (weighted) portions of partial matches, indicate that partial agreement occurs more prominently with S and NP segments than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

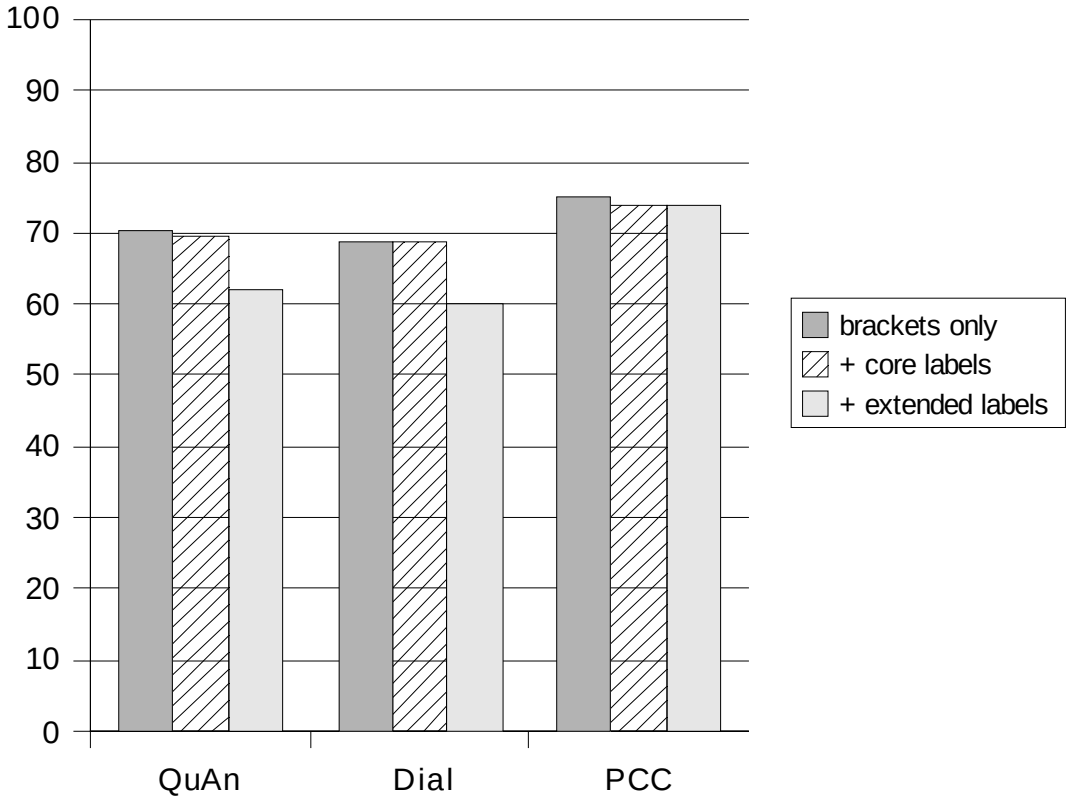
this text type contains highly constrained language content, in the form of short question-answer pairs, which appear to be suitable input for the Topic annotations.

- 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

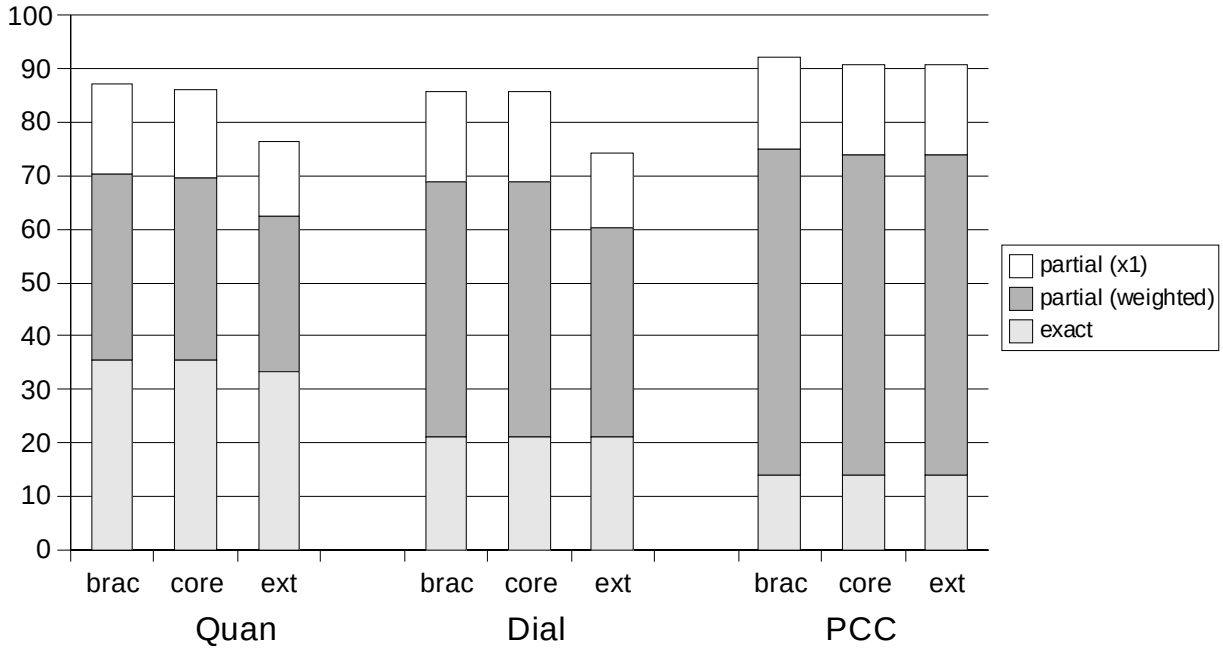


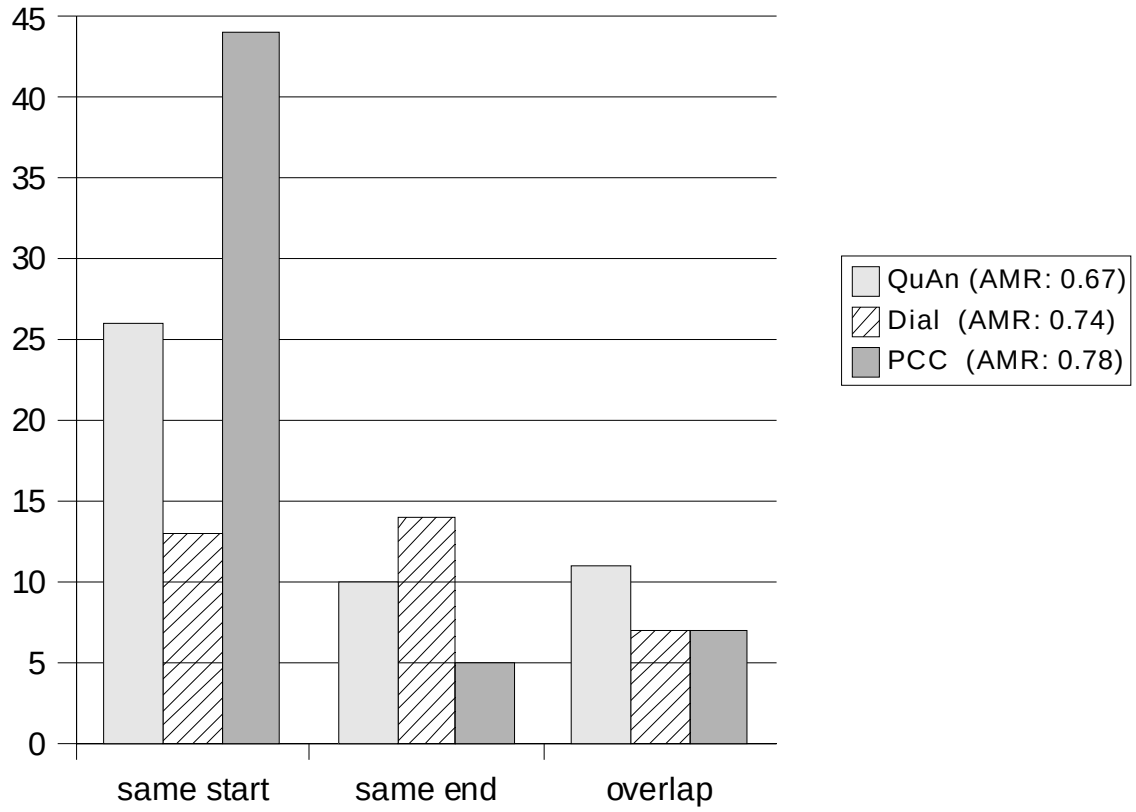
Figure 5 shows the F-scores of exact matches only (light-grey part), the F-scores when weighted partial matches are added (dark-grey part), and the F-scores that result if partial agreement is not weighted, i.e., not penalized at all (white part on top).[9]

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".

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

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:

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

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)

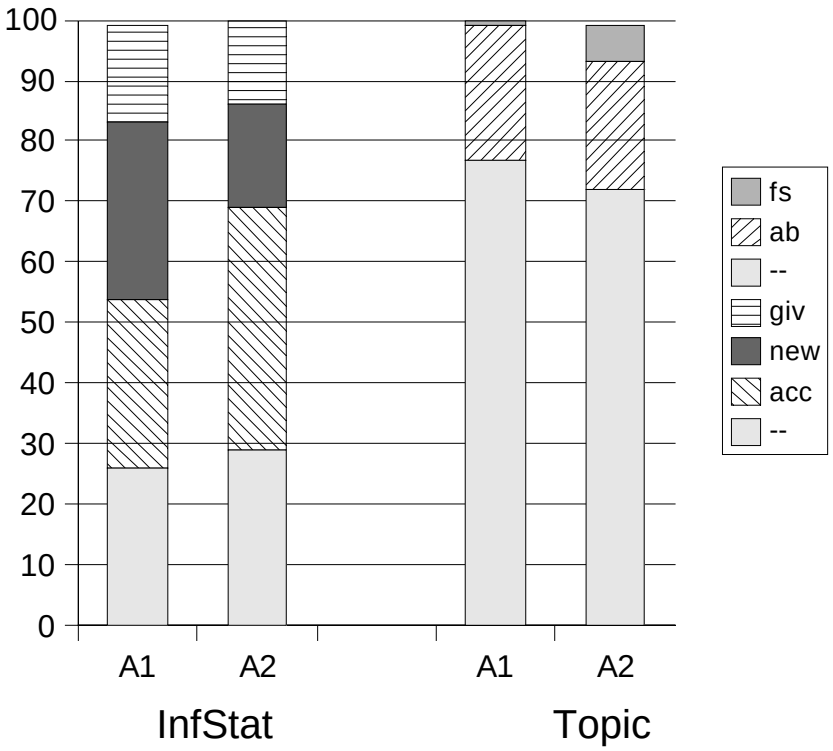
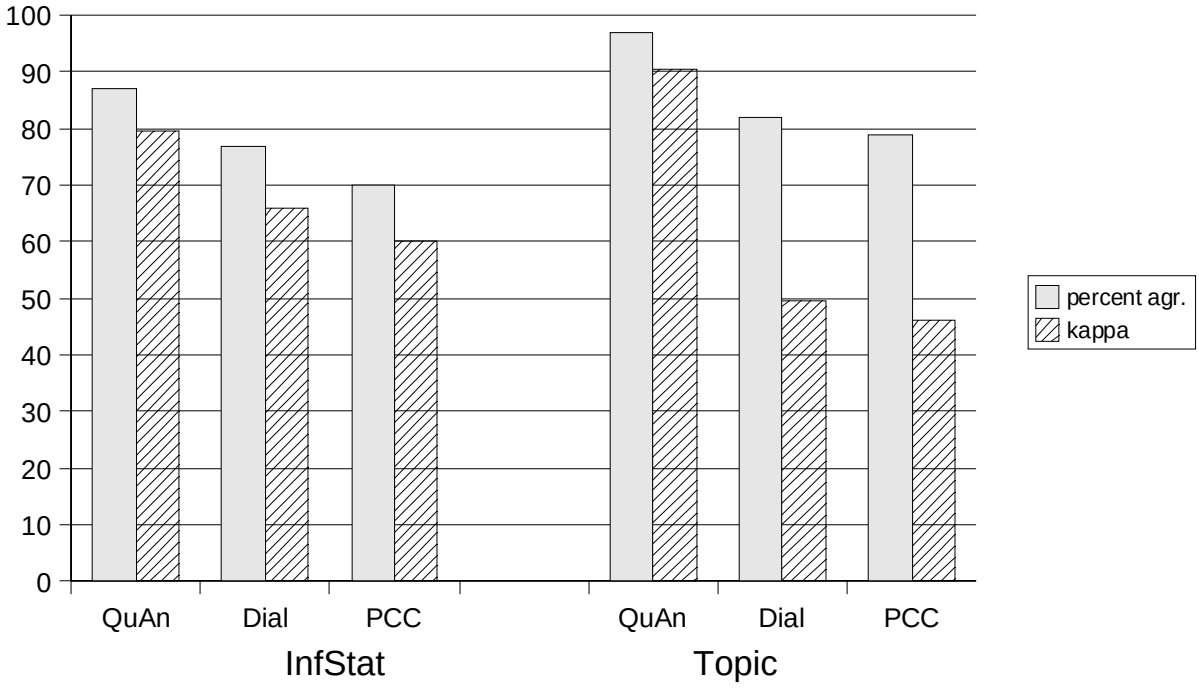


Figure 8: IS evaluation, percent agreement vs. kappa



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