Parallel Text Annotation for Information Structure

1. Introduction: As with other linguistic investigations, a deep analysis of Information Structure (hereafter, IS) requires the creation of language resources, in which linguistic features related to the phenomena in question are annotated in a fine-grained way. Languages use different phonological, morphological, and syntactic means of marking IS in sentences, and for many languages, the full range of IS marking possibilities remains unknown. Thus, the most comprehensive way of delving into cross-linguistic structuring of information is to analyze multilingual texts. Exploiting multilingual texts allows us to determine how IS strategies in different languages are related to each other, as well as to find systematic methods to identify topics and foci in monolingual texts.

2. Goal: This study ultimately aims to provide a fully annotated multilingual treebank that covers IS itself and linguistic domains relevant to IS. This data makes a significant contribution in at least two areas. First, this study can be used to support previous theoretical work completed on IS (e.g. Engdahl and Vallduví, 1996; Lambrecht, 1996; Gundel, 1999; Büring 1999). Second, this data can be used to aid in the development of computational models involving IS.

3. Previous Work: There have been several corpus studies for IS studied by Calhoun et al. (2005) and Dipper et al. (2007). Both provide guidelines for annotating IS in multilingual texts as well as monolingual texts. However, their studies differ from ours. While we use a fully parallel text, in which a sentence in one language is aligned with the same sentence translated in the other language(s), their multilingual corpora are sets of monolingual texts written in several languages, rather than parallel corpora. In the case of bitexts, there have also been previous studies which make a parallel analysis (e.g. Japanese to English (Komagata, 1999), Swedish to English (Johansson, 2001), and Norwegian to English (Bouma et al., 2010). But, in order to give a comparative explanation about distributional differences of IS in different languages, it is necessary to harness multilingual texts rather than just bitexts. Another advantage of our annotation schema is its coverage of dropped elements, making NP omissibility one of the important criteria for distinguishing topics from foci.

4. Basic Data: The running text used in this study is *The Little Prince* (originally written in French by Antoine de Saint-Exupéry). It was chosen because it has been translated into many languages, the sentence structure is relatively simple, and the size of the text is manageable. Additionally, since it is a naturally occurring text, each sentence can be analyzed in relation to its given context. We are presently investigating four languages: English, Spanish, Russian, and Korean.

5. Annotation: The preliminary steps include obtaining raw texts and sentence-aligning. The original texts in the respective languages were taken from websites or books, and a Python script was made to align sentences.

The main annotation tool for this study is EXMARaLDA (Extensible Markup Language for Discourse Annotation), which has been already used in the SFB632 project in Germany (http://www.sfb632.uni-potsdam.de). Dealing with datasets in a XML format, this software allows annotation of linguistic features at various layers (varying from phonology to discourse), using multiple tiers consisting of cell(s) for each word or phrase (http://www.exmaralda.org).

Our annotation schema was taken from Dipper et al. (2007), and adapted to fit our research on IS. First, the phonological layer has been eliminated, as there is currently no equivalent spoken data for *The Little Prince*. Second, most of the syntactic and semantic layers except for NP types (e.g. definiteness) have been removed. We were able to omit these layers in our annotation because we plan to supplement the extracted parts with the DELPH-IN grammars, given than several resource grammars are already available (e.g. ERG for English (Copestake and Flickinger, 2000), SRG for Spanish (Marimon et al., 2007), and KRG for Korean (Song et al., 2010)), or are under construction for all our target languages (e.g. RRG for Russian (Avgustinova and Zhang, 2010)). By parsing our data with these HPSG/MRS-based grammars, we will resolve syntactic and semantic constructions in a (semi-)automatic way when we build up treebank. Third, morphological and IS layers layers have not been significantly modified from Dipper et al. (2007). The former is composed of MORPH and GLOSS, and the latter consists of INFOSTAT (information status, such as *given*, *new*, *accessible*), TOPIC (*aboutness* or *frame-setting*), FOCUS (*new (un)solicited focus*), and CONTRAST (*contrastive topic* or *contrastive focus*). Finally, we have introduced three additional layers to our annotation schema; namely, INDEX, DROPPED, and IF. DROPPED layers point to the missing expression (DROPPED_WORD), the properties of the dropped element (DROPPED_FEAT), and index of its antecedent (DROPPED_IDX). INDEX layer, which represents word and phrase alignment among our target languages, will be determined using GIZA++ (Och and Ney, 2003) in an automatic way. IF (Inner Frame) and OF (Outer Frame) layers have been introduced to differentiate two types of discourses: dialogues between characters within the story (e.g. between

the little prince and the tippler) and author's narration, IF and OF respectively. The IF and OF layers have the same set of fields (i.e. parallel IS layers).

Figures 1 through 4 below illustrate our annotation schema. To help reader place this data in a context, the question preceding the sentence given in the examples below is "What are you doing there?" asked by the Little Prince and addressed to the tippler. In Figure 1, the subject 'I', corresponding to 'you', has *given-active* status (marked as *act*) and is topicalized, and the predicate 'am drinking' which corresponds to 'what' element in the preceding question is focused (i.e. wide-focus). In the cases like Spanish in which the subjects tend to be dropped, '#' is used in the position where a dropped element Ø is most likely to occur if it were not dropped. In this case, '1SG' on the fourth line indicates features of the dropped element. The cell in the next line is empty because there is no antecedent in the previous context. Figure 3 shows how our multilingual indexing works. Superscripts in each cell stand for word-alignment: for example, a sentence Bebo 'drink.1SG' in Spanish corresponds to 'am drinking' in English, as marked as (2-3). The dropped subjects in Spanish, Korean, and Russian are specified by '#', but they carry the same INFOSTAT (*given-active*) and TOPIC (*aboutness*) equivalent to 'I' in English. The VPs in all languages likewise get focused (*new-focus*) as new and solicited information. Though IS is the same across all four languages, the sentential form of English is *topic-focus*, unlike the other languages, in which this sentence has *all-focus* form. That is, even though dropped subjects Ø presumably can participate in syntactic configuration covertly, they cannot take overt part in it. Finally, Figure 4 demonstrates how IF annotation differs from OF annotation. The narration within direct quotation marks in this case, is the new information given by the sentence as a whole (the OF) and is thus annotated as the structure's focused element. The IF layer, however, concerns itself with the IS within the quoted dialogue, and within this frame, "am drinking" is focused as the predicate solicited by the question, "What are you doing?"

	I	am	drinking
MORPH	I	am	drink-ing
GLOSS	1SG	be.1SG.PRS	drink-PROG
DROPPED_FEAT			
DROPPED_IDX			
OF-INFOSTAT	act		
OF-TOPIC	ab		
OF-FOCUS		nf-sol	

Figure 1: An English Sample

	#	Bebo
MORPH		Beb-o
GLOSS		drink.1SG
DROPPED_FEAT	1SG	
DROPPED_IDX		
OF-INFOSTAT		
OF-TOPIC	ab	
OF-FOCUS		nf-sol

Figure 2: A Spanish Sample

	12.6.1	12.6.2	12.6.3	sentential form
English	I(1)	am(2)	drinking(3)	topic-focus
Spanish	#	Bebo(2-3)		all-focus
Russian	#	P'ju(2-3)		all-focus
Korean	#	swul(4)	masinta(3)	all-focus
OF-INFOSTAT	act			
OF-TOPIC/FOCUS				

Figure 3: A Multilingual Sample

	"I	am	drinking."	said	the	tippler.
OF-INFOSTAT					given-inactive	
OF-TOPIC						
OF-FOCUS	nf					
IF-INFOSTAT	act					
IF-TOPIC	ab					
IF-FOCUS		nf-sol				

Figure 4: A Sample for OF *vs.* IF

6. Progress: Currently, we are focusing on establishing more precise criteria for multilingual annotation. As is well-known, an IS category in one language does not always correspond to the same category in other languages. For example, topic/focus does not apply to the same lexical item in different languages, and the proportion of topic-drop differs from language to language as well. In addition, we are already seeing how English, Russian, and Spanish differently mark contrastive focus, due to the differences in the sentential structure of these languages. Our initial data shows that some of the focused lexical items marked contrastively in English and Spanish, are not marked contrastive in Russian. A possible way to resolve these and other mismatches in a long-term would be to develop an HPSG/MRS-based treebank to cover IS from a cross-linguistic perspective, which will be in line with the framework of the Redwoods Treebank (Oepen et al., 2004) and the Hinoki Treebank (Bond et al., 2004).

7. Implications: The data-oriented findings of this study, as well as the dataset itself will be of great help to other linguistic endeavors, such as grammar engineering and Machine Translation (MT). We can build up a grammar library (Bender et al. 2010) for IS, which aims to work robustly from both theoretical and empirical viewpoints. If the theoretical basis can be grounded upon empirical findings from this and like studies, we can draw more substantial and well-balanced generalization about IS. Furthermore, this study can contribute to transfer-based MT (Oepen, 2007). Since an essential part of translation is reshaping the means of conveying information (i.e. involving IS) instead of simply changing the words or reordering phrases, this data can aid in producing more felicitous translations.

References

Avgustinova, Tania and Yi Zhang. 2010. Conversion of a Russian dependency treebank into HPSG derivations. In *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories (TLT'9)*, Tartu, Estonia.

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar Customization. *Research on Language & Computation*, 8(1):23–72.

Bond, Francis, Sanae Fujita , Chikara Hashimoto , Kaname Kasahara , Shigeko Nariyama , Eric Nichols , Akira Ohtani , Takaaki Tanaka , and Shigeaki Amano. 2004. The Hinoki Treebank: A Treebank for Text Understanding. In *Proceedings the 1st IJCNLP*, pages 158–167.

Bouma, Gerlof, Lilja Øvrelid, and Jonas Kuhn. 2010. Towards a Large Parallel Corpus of Cleft Constructions. In *Proceedings of LREC*, pages 3585–3592.

Büring, Jeanette K. 1999. Topic. In Bosch, Peter and Rob van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 142–165. Cambridge University Press, Cambridge.

Büring, Jeanette K. 1999. Topic. In Bosch, Peter and Rob van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 142–165. Cambridge University Press, Cambridge.

Calhoun, Sasha, Malvina Nissim, Mark Steedman, and Jason Brenier. 2005. A Framework for Annotating Information Structure in Discourse. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 45–52. Association for Computational Linguistics.

Copestake, Ann and Dan Flickinger. 2000. An Open-Source Grammar Development Environment and Broad-Coverage English Grammar using HPSG. In *Proceedings of the 2nd conference on Language Resources and Evaluation*, Athens, Greece.

Dipper, Stefanie, Michael Goetze, and Stavros Skopeteas. 2007. *Information Structure in Cross-linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*. Universitätsverlag Potsdam.

Engdahl, Elisabet and Enric Vallduví. 1996. Information Packaging in HPSG. *Edinburgh Working Papers in Cognitive Science*, 12:1–32.

Gundel, Jeanette K. 1999. On Different Kinds of Focus. In Bosch, Peter and Rob van der Sandt, editors, *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 293–305. Cambridge University Press, Cambridge.

Johansson, Mats. 2001. Clefts in Contrast: a Contrastive Study of it Clefts and wh Clefts in English and Swedish Texts and Translations. *Linguistics*, 39(3):547–582.

Komagata, Nobo N. 1999. *A Computational Analysis of Information Structure Using Parallel Expository Texts in English and Japanese*. Ph.D. thesis, University of Pennsylvania.

Lambrecht, Knud. 1996. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge, UK.

Marimon, Montserrat, Núria Bel, Sergio Espeja, and Natalia Seghezzi. 2007. The Spanish Resource Grammar: pre-processing strategy and lexical acquisition. In *Proceedings of the Workshop on Deep Linguistic Processing*, pages 105–111. Association for Computational Linguistics.

Och, Franz J. and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1): 19-51.

Oepen, Stephan, Daniel Flickinger, Kristina Toutanova, and Christopher D. Manning. 2004. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. Journal of Research on Language and Computation, 2(4): 575–596.

Oepen, Stephan, Erik Velldal, Jan T. Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. 2007. Towards Hybrid Quality-Oriented Machine Translation. – On linguistics and probabilities in MT –. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, Skvde, Sweden.

Song, Sanghoun, Jong-Bok Kim, Francis Bond, and Jaehyung Yang. 2010. Development of the Korean Resource Grammar: Towards Grammar Customization. In *Proceedings of the 8th Workshop on Asian Language Resources*, pages 144–152. Beijing, China.