

# Parallel Syntactic Annotation of Multiple Languages

Owen Rambow<sup>\*</sup>, Bonnie Dorr<sup>†</sup>, David Farwell<sup>‡</sup>, Rebecca Green<sup>†</sup>, Nizar Habash<sup>\*</sup>  
Stephen Helmreich<sup>‡</sup>, Eduard Hovy<sup>♣</sup>, Lori Levin<sup>♡</sup>, Keith J. Miller<sup>♣</sup>  
Teruko Mitamura<sup>♡</sup>, Florence Reeder<sup>♣</sup>, Advait Siddharthan<sup>◇</sup>

<sup>\*</sup> Center for Computational Learning Systems, Columbia University, New York, NY, USA

{rambow, habash}@cs.columbia.edu

<sup>†</sup> University of Maryland, College Park, MD, USA

{bonnie, rgreen}@umd.edu

<sup>‡</sup> New Mexico State University, Las Cruces, NM, USA

{david, shelmrei}@crl.nmsu.edu

<sup>♣</sup> ISI, University of Southern California, Marina Del Rey, CA, USA

hovy@isi.edu

<sup>♣</sup> MITRE, Reston, VA, USA

{keith, freeder}@mitre.org

<sup>♡</sup> LTI, Carnegie-Mellon University, Pittsburgh, PA, USA

{lsl, teruko}@cs.cmu.edu

<sup>◇</sup> Cambridge University, Cambridge, UK

as372@cl.cam.ac.uk

## Abstract

This paper describes an effort to investigate the incrementally deepening development of an interlingua notation, validated by human annotation of texts in English plus six languages. We begin with deep syntactic annotation, and in this paper present a series of annotation manuals for six different languages at the deep-syntactic level of representation. Many syntactic differences between languages are removed in the proposed syntactic annotation, making them useful resources for multilingual NLP project with semantic components.

## 1. Background: Goals of Annotation

The IAMTC project (Farwell et al., 2004) aims at defining a level of interlingual annotation (the information needed to translate a text from one language to the next) based on annotating parallel multilingual texts (i.e., multiple translations into English of source texts in six foreign languages). As a first step in the sequence of annotations, we annotate texts for syntax. This level of annotation is called IL0. Subsequently, we augment IL0 with semantic disambiguation annotations, namely concepts from an ontology and semantic roles (IL1). This annotation does not change the structure of IL0. We then reconcile different IL1s from parallel texts into the common interlingual representation (IL2). In this paper, we discuss annotation standards for IL0 for Arabic, English, French, Hindi, Japanese, Korean, and Spanish. For details on the other levels of annotation, see (Farwell et al., 2004).

There has been much activity in syntactic annotation of corpora, starting with the Penn Treebank for English (Marcus et al., 1993), and more recently, there has also been semantic annotation on top of the Treebank, such as PropBank (Kingsbury et al., 2002). However, our project imposes specific requirements on syntactic annotation, which are not faced by other annotation projects:

- Because our goal is in fact *interlingual* annotation and syntax is just an intermediate representation, we are only concerned with the syntactic predicate-argument structure amongst the meaning-bearing words of a sentence, but not with certain details of syntax, such as function words.

- Because in IL2 we reconcile representations based on the augmented syntactic representations from different languages (as well as paraphrases from the same language), we want to choose representations that eliminate non-semantic syntactic differences as much as possible (see the example in Section 3.).

The second requirement is similar to the goal of the ParGram project (Butt et al., 2002); however, the ParGram project is motivated by the theoretical assumption that grammars of different languages are in fact similar (Universal Grammar), an issue we are agnostic on. Furthermore, ParGram is a grammar development project, while our project is a text annotation project.

## 2. Our Syntactic Annotation

These two requirements led us to define IL0 as an unordered deep syntactic dependency representation, inspired by the Deep-Syntactic Structure of Meaning-Text Theory (Mel'čuk, 1988) and the Analytical and Tectogrammatical Representation of the Prague School (Sgall et al., 1986). Only content words are represented. Function words (auxiliaries, determiners) are omitted and their meaning represented as features on the content nodes. Missing arguments (such as embedded subjects in control constructions) are added as lexically empty nodes with coindexation information, since some languages (or same-language paraphrases) may represent these arguments with overt pronouns. Nodes are annotated with the citation form of the inflected word, its base part-of-speech (noun, verb, etc), and several POS-specific morphological and morpho-syntactic

features (such as voice, aspect, number, gender, etc). Arcs are annotated with the underlying syntactic relation, which is either a type of argument (“0” for subject, “1” for direct object, and so on), or simply “adjunct”. The argument roles are normalized for regular syntactic transformations, which include active/passive alternation, and, in English, dative shift. We do not normalize alternations which always involve at least one PP such as *load trucks with hay/load hay into trucks*. For such constructions, the IL1 annotation expresses their similar meaning. Note that representations very similar to our IL0 are sometimes called “semantic”, but the relevant criteria for IL0 are in fact purely syntactic.

### 3. Cross-Linguistic Aspects

As a result of these decisions, many syntactic differences between languages are removed, at the cost of giving some languages a syntactic analysis which at first sight may not be the most obvious one. For example, we uniformly analyze predicative nouns, adjectives, and prepositions as the syntactic head, and any copula as an auxiliary which is omitted. Thus, Japanese (where adjectives are morphologically like verbs in that they inflect for tense), Arabic (where the copula is omitted for present tense) and English (which always uses a copula in main clauses) all have the same syntactic analysis for such predicative constructions, as shown in Figure 1. The adjective gets the feature *Pred*, which means it is being used predicatively, and it then can also have verbal features, including tense. In Figure 1 we show the past tense examples, and the present tense examples simply have the feature *present*. The IL1 we derive (in all cases) is shown in Figure 2.

- (1) a. al-muzlap  $\emptyset$ /kun al-aHmaru (Arabic)  
the-umbrella  $\emptyset$ /was the-red  
the umbrella is/was red
- b. kasa-wa akai/akakatta (Japanese)  
umbrella<sub>TOP</sub> red<sub>PRES</sub> /red<sub>PAST</sub>  
the umbrella is/was red

Similarly, other constructions such as control are treated similarly across languages. Some constructions do not exist in all languages. For example, Arabic does not have raising or exceptional case-marking (raising-to-object), while not all languages have serial verb constructions (for example, Hindi does while French does not). We will give more complete details of differences in the final version of the paper, as we consider this an important contribution of our work.

### 4. Practical Aspects

In our project, we constructed IL0 by hand-correcting the output of a dependency parser or from scratch, depending on the language. The IL0-annotated structures were subsequently augmented with IL1 by annotators; (Passonneau et al, in preparation) reports on the inter-annotator agreement of that effort and shows that IL0 indeed was a successful starting point for IL1 annotation.

We will make the annotation manuals available to the community.

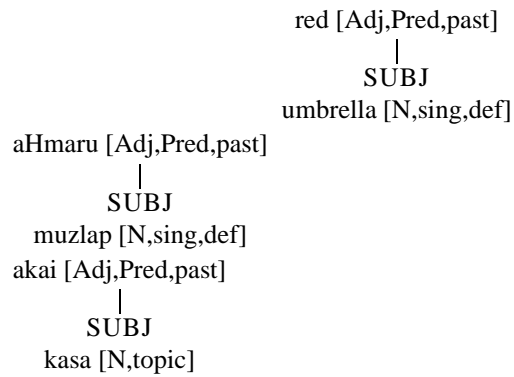


Figure 1: IL0 deep-syntactic representation for *the umbrella was red*, *al-muzlap kun al-aHmaru*, and *kasa-wa akakatta*

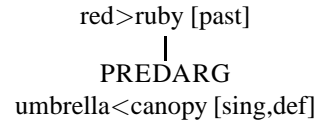


Figure 2: IL1 (semantically annotated) representation for *al-muzlap kun al-aHmaru*, *kasa-wa akakatta*, and *the umbrella was red*; *umbrella<canopy* and *red>ruby* are pointers to nodes in the ontology

## 5. References

- Butt, Miriam; Dyvik, Helge; King, Tracy Holloway; Masuichi, Hiroshi; and Rohrer, Christian (2002). The parallel grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7, Taipei, Taiwan.
- Farwell, David; Helmreich, Stephen; Reeder, Florence; Dorr, Bonnie; Habash, Nizar; Hovy, Eduard; Levin, Lori; Miller, Keith; Mitamura, Teruko; Rambow, Owen; and Siddharthan, Advait (2004). Interlingual annotation of multilingual text corpus. In *Proceedings of the NAACL/HLT Workshop: New Frontiers in Corpus Annotation*.
- Kingsbury, Paul; Palmer, Martha; and Marcus, Mitch (2002). Adding semantic annotation to the Penn Treebank. In *Proceedings of the Human Language Technology Conference*, San Diego, CA.
- Marcus, Mitchell M.; Santorini, Beatrice; and Marcinkiewicz, Mary Ann (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19.2:313–330.
- Mel’čuk, Igor A. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.
- Sgall, P.; Hajičová, E.; and Panevová, J. (1986). *The meaning of the sentence and its semantic and pragmatic aspects*. Reidel, Dordrecht.