

Parallel Syntactic Annotation of Multiple Languages

Owen Rambow*, Bonnie Dorr†, David Farwell‡, Rebecca Green†, Nizar Habash*
Stephen Helmreich‡, Eduard Hovy♣, Lori Levin♥, Keith J. Miller♣
Teruko Mitamura♥, Florence Reeder♣, Advaith Siddharthan◇

* Center for Computational Learning Systems, Columbia University, New York, NY, USA
{rambow, habash}@cs.columbia.edu

† University of Maryland, College Park, MD, USA
{bonnie, rgreen}@umd.edu

‡ New Mexico State University, Las Cruces, NM, USA
{david, shelmrei}@crl.nmsu.edu

♣ ISI, University of Southern California, Marina Del Rey, CA, USA
hovy@isi.edu

♣ MITRE, McLean, VA, USA
{keith, freedre}@mitre.org

♥ LTI, Carnegie Mellon University, Pittsburgh, PA, USA
{lsl, teruko}@cs.cmu.edu

◇ 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 projects with semantic components.

1. Introduction: 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).¹ 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 [4.]).

These requirements lead us to push the syntactic annotation as “deep” as possible without becoming semantic. It also means that choices in one language are coordinated with choices in the other languages.

This paper is structured as follows. We first discuss related work in Section [2.]. We then lay out the basics of our syntactic annotation in Section [3.], and illustrate the effect of multilingual annotation in Section [4.]. We discuss the features used in Section [5.], and some more constructions in Section [6.]. We finish with some comments on the practical aspects of annotation.

2. Related Work

The IL0 level of representation is very similar to (and inspired by) the tectogrammatical level of representation of the Prague theory (Sgall et al., 1986).² Annotated corpora

²The deep-syntactic level of representation of Meaning-Text Theory (Mel’čuk, 1988) is also similar, though we are not aware of annotated corpora. The English annotation manual is based on (Rambow et al., 2002), which in turn reflects the influences discussed in this paragraph.

¹This work has been supported by NSF ITR Grant IIS-0325887.

are available for Czech and English in the Prague Dependency Treebank (Hajič et al., 2001). Our IL0 takes from the tectogrammatical representation the notion that the linguistic contribution of (most) function words should be represented by features rather than by nodes in the tree (though IL0 keeps prepositions as separate nodes). The principal difference is that the tectogrammatical representation is a hybrid syntactic-semantic level of representation, with some arguments and all adjuncts annotated with semantic labels, while our scheme postpones any semantic label to further levels of annotation (IL1 and IL2). A secondary difference is that we keep prepositions in our IL0.

The PropBank (Kingsbury et al., 2002) shares many characteristics with IL0. IL0 is a purely syntactic level of annotation, while PropBank captures some aspects of lexical semantics. In particular, for a given set of alternations of one verb, the arguments are labeled consistently for that alternation, and the arguments are given labels specific to that set of alternations. For example, in both *John loaded the truck with hay* and *John loaded hay into the truck*, *hay* would have the same role label in PropBank, but different role labels in IL0 (it would be the object of a prepositional argument in the first sentence, the direct object in the second). Thus, both the Tectogrammatical Representation and PropBank are a level of representation intermediate between our IL0 and IL1. For a fuller discussion of these representational choices, see (Rambow et al., 2003).

Projects which might be seen as in some sense similar to the IAMTC annotation effort include Eurotra, EuroWordNet and the Universal Networking Language initiative (UNL). A crucial difference between our annotations and these projects is that our work is conceived of as an annotation project, while none of these projects included annotation.

Eurotra (Allegrezza et al., 1991) is similar to our effort in that it was a multi-site, multilingual effort but focused on developing a common framework for describing different natural languages on a range of levels: lexical, morphological, syntactic and semantic. However, Eurotra assumed a transfer-based approach to MT and so each language had its own syntactic and semantic processes and representations which were to be interconnected by pairwise transfer rules. There was no concern with developing an Interlingua and the methodology was essentially linguistic, motivating the framework on the basis of counter-examples rather than by way of corpus analysis and annotation.

EuroWordNet (Vossen, 1998), initially an effort to build WordNet resources for six European languages in parallel, is essentially lexical in nature. The central methodology was to translate the original Princeton WordNet for English into the other language, most importantly facing up to the problems of lexical mismatches or overlaps of the target language and filling in any lexical gaps in the original English resource. It was not concerned with sentence meaning or how it is represented. With the introduction of Inter-Lingual-Indexes, an effort was made to establish a cross-language mapping at the lexical level but, again, the developers did not follow a corpus based methodology and there was no related annotation effort.

Universal Networking Language (UNL) is a formal language designed for rendering automatic multilingual infor-

mation exchange (Martins et al., 2000). It is intended to be a cross-linguistic semantic representation of sentence meaning consisting of concepts (e.g., ‘cat’, ‘sit’, ‘on’, or ‘mat’), concept relations (e.g., ‘agent’, ‘place’, or ‘object’), and concept predicates (e.g., ‘past’ or ‘definite’). UNL syntax supports the representation of a hypergraph whose nodes represent “universal words” and whose arcs represent “relation labels”. Several semantic relationships may hold between universal words including synonymy, antonymy, hyponymy, hypernymy, meronymy, etc. Like the IAMTC effort, the UNL consortium is looking to create a practical IL by comparing translations across multiple languages at multiple sites and the results of both efforts may prove to be mutually informative both methodologically (multilingual, multi-site annotation) and at the level of formal representation.

Our goals are in some way similar to the goals of the ParGram project (Butt et al., 2002), in which grammars for several languages are developed in close consultation and in parallel; 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.

3. Our Syntactic Annotation

In Section [1.], we motivated our IL0 representation, and we concluded that we wanted a representation that concentrates on meaning-bearing (autosemantic) lexemes, and that reduces cross-linguistic differences. These requirements have led us to define IL0 as an unordered deep syntactic dependency representation. Only content words are represented. The dependency relations reflect syntactic predicate-argument structures, not (necessarily) surface-syntactic relations (such as case marking or agreement; see Section [6.2.] for an example). 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. 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 or simply MOD for modifiers (adjuncts). The argument roles are normalized for regular syntactic transformations, which include active/passive alternation. 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.

4. Cross-Linguistic Aspects

There are two ways in which IL0 succeeds in making different languages look alike already at the syntactic level:

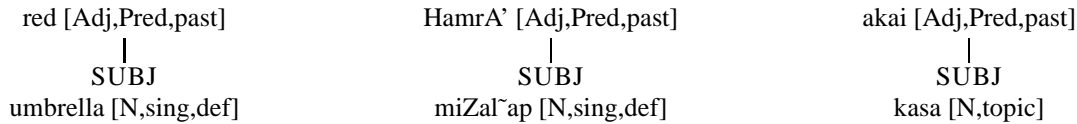


Figure 2: IL0 deep-syntactic representation for *the umbrella was red*, *kAnat AlmiZal~apu HamrA'F*, and *kasa-wa akakatta*



Figure 1: IL0 deep-syntactic representation for *llegará Juan* and *Juan will arrive*

- The basic definition of IL0 presented in Section [3.] equalizes certain differences, by not representing word order, and by representing function words as features.
- The basic definition leaves many option for defining the structure given a certain construction. When choosing the syntactic analysis for IL0, we look at all languages, and choose a uniform analysis for related constructions. Here, we may end up with an analysis which gives some languages a syntactic structure which at first sight may not be the most obvious one.

We discuss and exemplify these cases in turn. Many syntactic differences between languages are removed by removing word order and function words. For example, English forms the future tense with an auxiliary, while Spanish has an inflectional morpheme, and also a postposed subject:

- (1) *llegará Juan*
arrive_{FUT} Juan
Juan will arrive

However, both sentences are structurally identical at IL0, as seen in Figure 1.

5. Features on Nodes

We record all syntactic information in IL0 so that the surface form (both morphological and syntactic) can deterministically be generated from it. Since the morphology and morphosyntax of different languages express different features, we accept that we cannot have a uniform feature set cross-linguistically. By way of example, we will discuss the part-of-speech feature, and then the features found on verbs.

5.1. Parts of Speech

The lists of parts of speech is the same in all languages we deal with.

- V: verbs, but not auxiliary verbs (=Aux)
- N: common nouns and personal pronouns
- PN: proper nouns
- Adj: adjectives
- Adv: adverbs

- P: prepositions and subordinating conjunctions
- Conj: coordinating conjunctions, but not subordinating conjunctions; also includes the comma used in enumerations instead of repeated *and*
- Det: determiners; only used for demonstratives and so on, since the *and* and *a* do not appear in IL0
- Aux: auxiliary verbs; at IL0, only modal auxiliaries are included, not the auxiliaries for passive, progressive, etc.
- Pun: punctuation marks, but not the comma used in conjunctions
- Sym: various symbols (dollar signs and the like)
- Uh: speech-specific sounds, even if meaningful (such as /UH HUH/)
- Misc: everything else, including greetings (*Hi*, *Hello*) and interjections (*Okay*)

For some of the languages, not all parts of speech are always recognized in the traditional analyses. For example, in Arabic, adjectives are not traditionally distinguished from nouns, since their morphology is identical. However, the distinction can be made in Arabic as well by referring to English cases.

We now discuss features present for verbs and predicative nouns, adjectives, and prepositions. Here, the morphology and morphosyntax of the languages imposes certain differences. These features do not capture semantics (this is handled at later stages of annotation), but rather morphological and morphosyntactic forms (morphemes, auxiliaries) that have been removed in IL0.

- Progressive (prog): a binary feature that marks whether a verbal complex is progressive. Present in English (*is sneezing*, *will have been eating*) and Spanish (*está realizando* ‘is carrying out’).
- Perfective (perf): a binary feature that marks whether a verbal complex is perfective. Present in English (*has eaten*, *will have been eating*), Spanish (*ha comido*, and French *a mangé*), where the perfective is marked with an auxiliary. This feature is also used in Arabic to make the rather different distinction between the perfective and imperfective verbal forms, neither of which carries an auxiliary. The Arabic perfective is often considered semantically equivalent to the past tense in other languages, but this meaning is only normalized at later levels of annotation.

- Tense (tense): a feature that takes as value different possible tenses. In English, French, and Spanish, it marks whether a verbal complex is past (*ate, mangea*), present *eats, mange*), or future (*will eat, mangera*). Note that the feature is insensitive to whether there is a bound morpheme or an auxiliary expressing it. In Korean and Japanese, there is only a past/non-past distinction. In Arabic, there is no tense at all (see “perfective”).
- Mood (mood): a feature that marks for English whether a verbal complex is indicative (*eats*), imperative (*Eat!*), or subjunctive (*eat in lest he eat*). Different languages have different moods. While the indicative and imperative are common, the subjunctive is less so, and Arabic alone also has a jussive. In many cases, the subjunctive carries no meaning per se and is lexico-syntactically conditioned and carries no meaning (French *je ne crois pas qu’il vienne*, I NEG think NEG that he come/subj, ‘I don’t think he will come’), while in other cases the choice among moods is meaningful and will be transformed into semantic features at later levels of annotation (e.g., choice between indicative and imperative).

6. Some Constructions

Many constructions such as clausal embedding are treated similarly across languages. We discuss in this section three constructions in more detail as they differ cross-linguistically in interesting ways: copula constructions, the causative, and serial verbs.

6.1. Copular Constructions

The second case (in which the basic definition of ILO is not sufficient to make two languages look similar) is illustrated by the copular construction (predicative nouns, adjectives, and prepositions). Consider the following predicative adjective sentences. In Arabic, the copula is omitted for present tense but present for past tense. In Japanese, adjectives are morphologically like verbs in that they inflect for present or past tense. English always uses a copula in main clauses, no matter what the tense.³

(2) a. AlmiZal~apu HamrA’N (Standard Arabic)

the-umbrella the-red

the umbrella is red

b. kAnat AlmiZal~apu HamrA’F
was the-umbrella_{NOM} the-red
(Standard Arabic)

the umbrella was red

c. kasa-wa akai (Japanese)
umbrella_{TOP} red_{PRES}

the umbrella is red

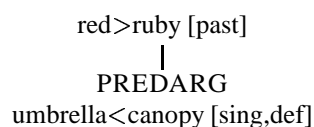


Figure 3: IL1 (semantically annotated) representation for *kAnat AlmiZal~apu HamrA’F*, *kasa-wa akakatta*, and *the umbrella was red*; *umbrella<canopy* and *red>ruby* are pointers to nodes in the ontology

d. kasa-wa akakatta (Japanese)
umbrella_{TOP} red_{PAST}
the umbrella was red

We uniformly analyze predicative nouns, adjectives, and prepositions as the syntactic head, and any copula as an auxiliary. The auxiliary is omitted and its contribution is represented by features, following the basic ILO definition. Thus, Arabic, Japanese, and English all have the the same syntactic structure for such predicative constructions, as shown in Figure 2. 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 2 we show the past tense examples, and the present tense examples are identical, but have the feature *present*. The IL1 we derive (in all cases) is shown in Figure 3.

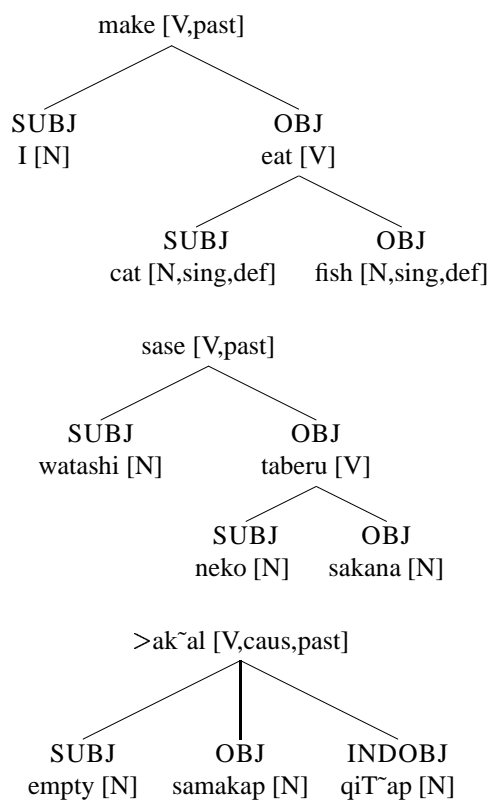


Figure 4: ILO deep-syntactic representation for *I made the cat eat the fish* (top), *watashi ha neko ni sakana wo tabe-sase-ta* (Japanese, middle), and *>ak altu AlqiT~apa Al-samakapa* (Arabic, bottom)

³For Arabic, we use the Buckwalter transcription of diacritized orthography (<http://www.qamus.org/transliteration.htm>).

6.2. The Causative and Exceptional Case Marking Verbs

Japanese and Korean have morphemes which can be added to verbs productively to make the verb a causative. Here is a Japanese example:

- (3) watashi-ha neko-ni sakana-wo tabe-sase-ta
 I_{TOP} fish_{DAT} cat_{OBJ} eat-CAUSE-PAST
 I made the cat eat the fish

When analyzing this construction on its own, it would be conceivable to consider the verb (*tabesasetu* in our example) as a single item with an additional syntactic argument. However, our cross-linguistic approach leads us to propose that the morpheme *-sareru* (also *-seru*) in fact gets its own node, since it corresponds to what are clearly full verbs in most other languages, such as English (as shown in the gloss). The resulting ILO structures for Japanese and English are shown in Figure 4. The English analysis is an example of an ECM (exceptional case marking) verb, where the embedded subject gets accusative case through an “exceptional” mechanism from the matrix verb (the Mechanism does not interest us here). (We know that *cat* is the lower subject since we can have semantically vacuous words in that position which are only licensed as subjects: *he made there be a fish* but **he made there* and **he invited there to be a fish*).

In Arabic some verbs have a causative version through a change in the templatic morphology. Most frequently, this is from Form I to Form II (which results in a gemination of the middle consonant) or Form IV.

- (4) >ak~altu AlqiT~apa Alsamakpa
 eat.CAUS cat_{DEF, ACC} fish_{DEF, ACC}
 I made the cat eat the fish (or: I fed the cat the fish)

However, this is not a productive morphological process as in Japanese: it does not apply to all verbs, and not all Form II verbs have a causative meaning.⁴ Furthermore, there is no single morpheme which is *added* to get the causative reading and which could serve as root node in the tree. Therefore, in Arabic, we analyze the Form II verb which has a causative meaning as a single lexical item with an additional argument. We mention this case to illustrate that, while we strive to make constructions in different languages that are similar in meaning look similar syntactically, we only do so to the extent that the lexicon, morphology, and syntax of the language actually allow it. ILO is *not* a semantic level of representation.

6.3. Compound Verbs

There is a small class of Hindi verbs that function as light verbs in verb compounds. The main light verbs are *ja/gaya* ‘go/went’, *le* ‘take’, *de* ‘give’, *daal* ‘put’, but there are several more. For example, Examples:

- (5) a. hum santra kha gaye
 we oranges eat went

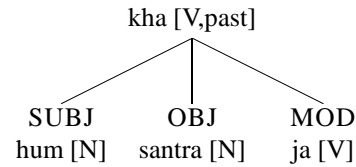


Figure 5: ILO deep-syntactic representation for Hindi *hum santra kha gaye* (5a)

We ate the oranges

- b. maine santra kha liya
 I-did orange eat take

I ate the orange

The function of these verbs is similar to modal auxiliary verbs in languages such as English in that light verbs carry the agreement features with the arguments of the verb compound; however the arguments are determined by the main verb solely. Semantically, the light verb adds aspectual information to the meaning of the main verb. We therefore treat these light verbs as modal auxiliaries and make the auxiliary dependent on the main verb, as shown in Figure 5. Note that the specific semantic contribution of the light verbs is not specific at ILO but rather at later levels of annotation.

There are some tricky cases where what appears to be a light verb is actually not semantically void. In these cases, they should not be removed.

- (6) a. Ram santra kha-kar jayega
 Ram orange eat-then go_{FUT}
 Ram will eat the orange and then leave
- b. Ram santra kha-ye jayega
 Ram oranges eating go_{FUT}
 Ram will go on eating oranges

In the the above examples, *ja* ‘go’ is not functioning as a light verb, since it actually carries its usual meaning of locomotion. *ja* contributes meaning to the sentence and should be preserved as a node. In these cases, *ja* is the head of the sentence, and the other verb (in this case, *kha* ‘eat’) should be a dependent of it. In both sentences, the embedded clause has an empty subject, which is indicated in the ILO structure (Figure 6) with a coindexed empty node. In (6a), the *kar* clitic indicates sequencing; in (6b), the *ye* suffix indicates an ongoing action. This is illustrated in Figure 6.

Note that the choice between a main verb analysis for *ja* or an auxiliary-type analysis depends on the annotator’s assessment of the meaning of *ja*. While ILO is a syntactic representation, the correct syntactic representation (i.e., the choice among many possible syntactic representations for a string of words) of course depends on the interpretation given to the string of words (ideally, in context) by the annotator. This comment applies to all syntactic annotation work.

⁴In fact, we have no consensus on the acceptability of our example among a group of educated Arabic speakers.

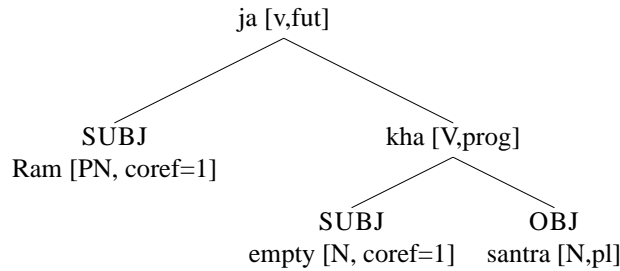


Figure 6: ILO deep-syntactic representation for Hindi *Ram santra kha-ye jayega (6b)*

7. Practical Aspects

In our project, we constructed ILO by hand-correcting the output of a dependency parser or from scratch, depending on the language. We used the TrEd annotation tool (Hajič et al., 2001) developed at Prague, which is easily configurable to any annotation format. Furthermore, it has the advantage that it is easy to convert the input and output to other formats, thus facilitating interfacing with a parser. The ILO-annotated structures were subsequently augmented with IL1 by annotators using a new tool which we developed; Passonneau et al. (2006) report on the inter-annotator agreement of that effort and shows that ILO indeed was a successful starting point for IL1 annotation.

8. Conclusion

Creating a syntactic annotation manual for a language amounts to writing a descriptive grammar with nearly complete coverage. It is a daunting task. Many choices must be made. These choices should be informed by an analysis of data, by syntactic theory (which one hopes is itself informed by an analysis of data), and/or by the goal of the annotation. Our syntactic annotation has two characteristics: it is only the first step in a semantic annotation effort; and it is intended to be used in the presence of parallel texts in different languages, i.e., different representations of the same content. We have taken these goals of the annotation task as our primary motivating forces in making decisions about annotation. We believe that just as parallel syntactic annotation leads to better semantic annotation, the parallel creation of syntactic annotation manuals leads to better-founded syntactic representations, and eliminates non-essential differences between languages which only complicate work in linguistics and natural language processing.

9. References

- Allegranza, V.; Bennett, P.; Durand, J.; Eynde, F. Van; Humphreys, L.; Schmidt, P.; ; and Steiner, E. (1991). Linguistics for machine translation: The eurotra linguistic specifications. In Copeland, C.; Durand, J.; Krauwer, S.; ; and Maegaard, B., editors, *The Eurotra Linguistic Specifications*, pages 15–124. CEC, Luxembourg.
- 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, Advaith (2004). Interlingual annotation of multilingual text corpus. In *Proceedings of the NAACL/HLT Workshop: New Frontiers in Corpus Annotation*.
- Hajič, Jan; Hajičová, Eva; Holub, Martin; Pajas, Petr; Sgall, Petr; Vidová-Hladká, Barbora; and Řezníčková, Veronika (2001). The current status of the prague dependency treebank. In *LNAI 2166*, LNAI 2166, pages 11–20. Springer Verlag, Berlin, Heidelberg, New York.
- 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.
- Martins, T.; Rino, L.H. Machado; Nunes, M.G. Volpe; Montilha, G.; ; and Novais, O. Osvaldo (2000). An interlingua aiming at communication on the web: How language-independent can it be? In *Proceedings of Workshop on Applied Interlinguas, ANLP-NAACL*.
- Mel’čuk, Igor A. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.
- Passonneau, Rebecca; Habash, Nizar; and Rambow, Owen (2006). Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of LREC*.
- Rambow, Owen; Creswell, Cassandra; Szekely, Rachel; Taber, Harriet; and Walker, Marilyn (2002). A dependency treebank for english. In *Proceedings of LREC*, Las Palmas, Spain. ELRA.
- Rambow, Owen; Dorr, Bonnie; Kipper, Karin; Kučerová, Ivona; and Palmer, Martha (2003). Automatically deriving tectogrammatical labels from other resources: A comparison of semantic labels across frameworks. *The Prague Bulletin of Mathematical Linguistics*, (79–80):23–36.
- Sgall, P.; Hajičová, E.; and Panevová, J. (1986). *The meaning of the sentence and its semantic and pragmatic aspects*. Reidel, Dordrecht.
- Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.