

Automatic Recognition of Logical Relations for English, Chinese and Japanese

Author Names Withheld for Blind Submission

Abstract

1 Introduction

For decades, computational linguists have paired a surface syntactic analysis with another analysis intended to represent something “deeper”. The work of Harris (Harris, 1951; Harris, 1968), Chomsky (Chomsky, 1957; Chomsky, 1981) and many others showed that one could use these deeper analyses to regularize differences between ways of expressing the same idea.¹ For statistical methods, these regularizations, in effect, reduce the number of significant differences between observable patterns in data and raise the frequency of each difference. Patterns are thus easier to learn from training data and easier to recognize in test data, thus somewhat compensating for the sparseness of data. In addition, deeper analyses are often considered semantic in nature because conceptually, two expressions that share the same regularized form also share some aspects of meaning. The specific details of this “deep” analysis have varied quite a bit, perhaps more than the surface syntax.

In the 1970s and 1980s, Lexical Function Grammar’s way of dividing C-structure (surface) and F-structure (deep) led to parsers such as (Hobbs and Grishman, 1976) which produced these two levels, typically in two stages. However, enthusiasm

¹It is not possible to give a full list of references for this idea as it has existed in one form or another for thousands of years, e.g., Panini’s *karaka* from 2000 years ago provide yet another way of regularizing different expressions of the same idea.

for these two-stage parsers was eclipsed by the advent of one stage parsers with much higher accuracy (about 90% vs about 60%), the now-popular treebank-based parsers including (Charniak, 2001; Collins, 1999) and many others. Currently, many different “deeper” levels are being manually annotated and automatically transduced, typically using surface parsing and other processors as input. One of the most popular, Semantic role labels (annotation and transducers based on the annotation) characterize relations anchored by select predicate types like verbs (Palmer et al., 2005), nouns (Meyers et al., 2004a), discourse connectives (Mitsakaki et al., 2004) or those predicates that are part of particular semantic frames (Baker et al., 1998). The CONLL tasks for 2008 and 2009 (Surdeanu et al., 2008) has focused on unifying many of these individual efforts to produce a logical structure for multiple parts of speech and multiple languages.

Our approach is much like the CONLL shared task. We link particular surface levels to particular logical levels and we do this for three languages. However, there are several differences:

1. The logical structures produced automatically by our system can reasonably be expected to be superior to the comparable CONLL systems. As we will explain, this expectation primarily relates to the level of granularity of our semantic roles. Our English result achieved a significantly higher F-score (89.9% on News, 76.3% to 90.3% on other text, including spoken language transcripts) than the best CONLL 2008 system (81.8% on News, 69.1% on the Brown Corpus). The CONLL 2009 task for English,

Japanese and Chinese has not been completed as of this date.

2. Each of the languages in our system uses the same linguistic framework, using the same types of relations, same analyses of comparable constructions, etc. (name and references omitted for blind submission). In one case, this required a conversion from a different framework to our own. In contrast, the 2009 CONLL task puts several different frameworks into one compatible input format.
3. The logical structures produced by our system typically connect all the words in the sentence. While this is true about some of the CONLL 2009 languages, e.g., Czech, but it is not true about all the languages. In particular, the CONLL 2009 English and Chinese logical structures only include noun and verb predicates.

In this paper, we will describe the XXXX framework² and a system for producing XXXX output. XXXX provides a logical structure for English, Chinese and Japanese with an F-score that is within a few percentage points of the best parsing results for that language. Like Lexical Function Grammar’s (LFG) F-structure, our logical structure is less fine-grained than many of the popular semantic role labeling schemes, but also has two main advantages over these schemes: it is more reliable and it is more comprehensive in the sense that it covers all parts of speech and the resulting logical structure is a connected graph.

2 The XXXX framework

Our system creates a multi-tiered representation in the XXXX framework, a framework which combines the theories underlying the Penn Treebank for English (Marcus et al., 1994) and Chinese (Xue et al., 2005) (which is heavily influenced by Chomskian linguistics of the 1970s and 1980s) with Relational Grammar’s graph-based way of representing “levels” as sequences of relations, feature structures in the style of Head-Driven Phrase Structure

²We use the name XXXX for the framework to facilitate blind review of this paper. In a similar vein, we will use YYYY and ZZZZ to represent other internal systems or frameworks.

Grammar and the Zelig Harris style goal of attempting to regularize multiple ways of saying the same thing into a single representation. Our approach differs from LFG F-structure in several ways: we have more than two levels; we have a different set of relational labels; and finally, our approach is designed to be compatible with the Penn Treebank framework and therefore, Penn-Treebank-based parsers.³

For each sentence, we generate a feature structure representing a phrase structure-based instantiation of our full analysis. We then distill this into an Dependency Representation representing a subset of this information and making certain theoretical assumptions, e.g., about identifying *functors* of phrases. In our framework, each dependency is between a functor and an argument, where functor is either the head of a phrase, a conjunction (in the case of conjoined structures), or function words like complementizers in the case of special constructions in which the term *head* does not clearly identify a child that can represent the whole phrase for purposes of selecting the phrasal category or selectional restrictions. For many applications, this dependency analysis is sufficient (references omitted) other applications have been designed to use the feature structure representation (references omitted).⁴ The dependency representation describes each sentence as a set of 23 fields (23-tuples), such that each 23-tuple represents up to three relations between two words: a SURFACE relation which essentially labels the relation between a functor and an argument in the parse of a sentence; a LOGIC1 relation which regularizes for various lexical and syntactic phenomena, e.g., passive, relative clauses, infinitival subjects, parentheticals, etc.; and a LOGIC2 relation which corresponds to relations in PropBank, Nombank and the Penn Discourse Treebank (PDTB). While the full output has all this information, we will limit this paper to a discussion of the LOGIC1 relations. Figure 1 gives a 5 tuple subset of the 23 tuple XXXX

³For example, it is possible to derive a Penn Treebank style parse from our XXXX representation, which would for example, include all part of speech correction and phrase structure changes made as part of our processing.

⁴The choice of representation used by applications relates to specifics of our system and has nothing to do with intrinsic properties of dependency versus phrase structure representations.

L1	Surf	L2	Func	Arg
NIL	SENT	NIL	Who	was
PRD	PRD	NIL	was	eaten
COMP	COMP	ARG0	eaten	by
OBJ	NIL	ARG1	eaten	Who
NIL	OBJ	NIL	by	Grendel
SBJ	NIL	NIL	eaten	Grendel

Figure 1: 5-tuples for *Who was eaten by Grendel*

analysis of the sentence *Who was eaten by Grendel?*. The tuples listed are: logic1 label, surface label, logic2 label, functor and argument. NIL indicates that there is no relation of that type. Figure 2 represents this as a graph. For edges with two labels, the ARG0 or ARG1 label indicates a LOGIC2 relation. Edges with an L- prefix are LOGIC1 labels (the edges are curved); edges with S- prefixes are SURFACE relations (the edges are dashed); and other (thick) edges bear unprefix labels representing combined SURFACE/LOGIC1 relations. Deleting the dashed edges yields a LOGIC1 representation; deleting the curved edges yields a SURFACE representation; and a LOGIC2 representation is created by only including the nodes connected by the ARG0 and ARG1 relations, with the proviso that when an argument is a preposition, complementizer or other function word that takes exactly one argument: an NP or an S, one should include the NP or S as part of the argument.⁵ together, the SURFACE relations for a sentence form a tree; the LOGIC1 relations form directed acyclic graph; and the LOGIC2 relations form directed graphs which may include some cycles and, due to PDTB relations, may actually connect this sentence to previous ones, e.g., adverbs like *however*, take the previous sentence as one of their arguments.

LOGIC1 relations (based on Relational Grammar) are designed to regularize across grammatical and lexical alternations, but not to make fine grained distinctions. For example, subcategorized verbal arguments include: SBJect, OBJect and IND-OBJ (indirect Object), COMPLEMENT, PRT (Particle), PRD (predicative complement); other verbal modifiers include AUXilliary, PARENthetical, AD-

⁵Additional provisos are needed to handle other special structures, e.g., coordination, named entities, etc.

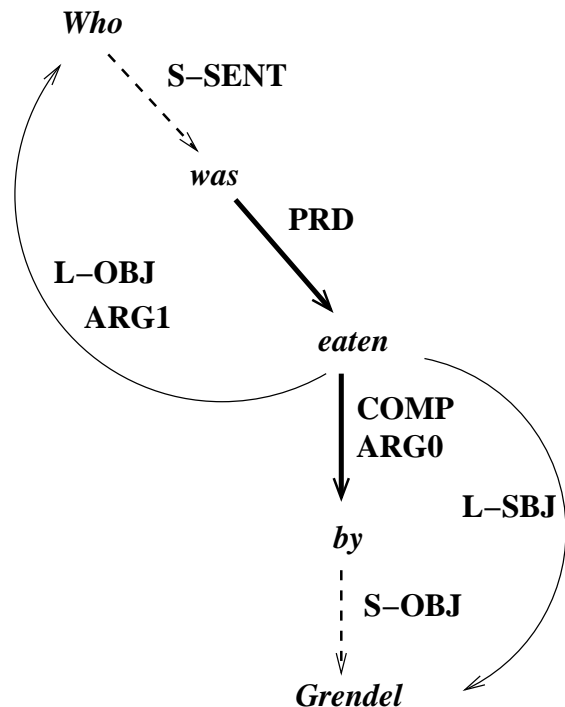


Figure 2: Graph of *Who was eaten by Grendel*

Verbial. While a subset of these are available for every verb, FrameNet and PropBank clearly make finer distinctions, including those which are lexically specific. For example, both PP arguments of *consulted* in the following sentence bear COMP relations with the verb in XXXX, but would have different ARG numbers in PropBank or different role names in FrameNet: *John consulted with Mary about the project*. Thus one would expect Semantic Role Labeling to be more difficult than recognizing LOGIC1 relations.⁶

Beginning with Penn Treebank II, Penn Treebank annotation includes Function tags, hyphenated additions to phrasal categories which indicate their function. There are several types of function tags:

- **Argument Tags** such as SBJ, OBJ, IO (IND-OBJ), CLR (COMP) and PRD—These are limited to verbal relations and not all are used in all treebanks. For example, OBJ and IO are used in the Chinese, but not the English treebank. These labels can often be directly trans-

⁶XXXX includes relations in which all parts of speech are predicates, not just verbs. However, space limitations prevent us from going into detail for most of these relations.

L1	Surf	L2	Func	Arg
NIL	SBJ	ARG1	is	ketchup
PRD	PRD	ARG2	is	nutritious
SBJ	NIL	NIL	nutritious	Ketchup
ADV	ADV	NIL	nutritious	very
N-POS	N-POS	NIL	ketchup	Banana
NIL	PAREN	NIL	is	claims
SBJ	SBJ	ARG0	claims	agency
Q-POS	Q-POS	NIL	agency	the
COMP	NIL	ARG1	claims	is

Figure 3: 5-tuples for *Banana Ketchup*, *the agency claims*, *is very nutritious*

lated into XXXX LOGIC1 relations.

- **Adjunct Tags** such as ADV, TMP, DIR, LOC, MNR, PRP—These tags often transate into a single LOGIC1 tag (ADV). However, some of these also correspond to LOGIC1 arguments. In particular, some DIR and MNR tags are realized as LOGIC1 COMP relations (based on dictionary entries). The fine grained semantic distinctions are maintained in other features that are part of the XXXX description.

In addition, XXXX treats Penn’s PRN phrasal category as a relation rather than a phrasal category. For example, given a sentence like, *Banana ketchup, the agency claims, is very nutritious*, the phrase *the agency claims* is a sentence in XXXX bearing a (surface) PAREN relation to the main clause. Furthermore, the whole sentence is a COMP of the verb *claims*. Since PAREN is a SURFACE relation, not a LOGIC1 relation, there is no LOGIC1 cycle as shown by the set of 5-tuples in Figure 3.

Another important feature of the XXXX transparency is transparency. A relation between two words is transparent if: the functor fails to characterize the selectional properties of the phrase, but its argument does. For example, relations between conjunctions (e.g., *and*, *or*, *but*) and their conjuncts are transparent CONJ relations. Thus although *and* links together *John* and *Mary*, it is these dependents that determine that the resulting phrase is noun-like (an NP in phrase structure terminology) and sentient (and thus can occur as the subject of verbs like *say*). Another common example of transparent relations are the relations connecting certain nouns and

the prepositional objects under them, e.g., *bag of sandwiches* is edible, because sandwiches are edible even though bags are not. These features are marked in the NOMLEX-PLUS dictionary (Meyers et al., 2004b).

The above description most accurately describes English XXXX. However, Chinese XXXX has most of the same properties, the main exception being that PDTB arguments are not currently marked. For Japanese, we have only a preliminary representation of LOGIC2 relations and they are not derived from PropBank/NomBank/PDTB. More complete XXXX specifications are available online (website omitted for blind review).

2.1 Scoring the LOGIC1 Structure

For purposes of scoring, we chose to focus on LOGIC1 relations, our proposed high-performance level of semantics. We scored with respect to: the LOGIC1 relational label, the identity of the functor and the argument, and whether the relation is transparent or not. If the system output differs in any of these respects, the relation is marked wrong. The following sections will briefly describe each system and present an evaluation of its results.

3 English XXXX

We generate English XXXX output by applying a procedure that combines:

1. The output of the 2005 version of the Charniak parser described in (Charniak, 2001). While the cited paper cites label precision and recall scores in the 85% range, we believe the updated version of the parser seems to perform closer to 90% on input similar to the Penn Treebank Wall Street Journal corpus, and perform lower on other genres. That performance would reflect reports on other versions of the Charniak parser for which performance statistics are available (Foster and van Genabith, 2008).
2. Named entity tags from the YYYY NE system (citation omitted), which achieves F-scores ranging from X% for Z genre to Z% for Newswire on ACE data.
3. Several Machine Readable dictionaries including COMLEX (Macleod et

al., 1998) and the dictionaries that come with NomBank <http://nlp.cs.nyu.edu/meyers/nombank/those-other-nombank-dictionaries.pdf>.

4. A sequence of hand-written rules (citations omitted) such that: (1) the first set of rules convert the Penn Treebank into a Feature Structure representation; and (2) each rule N after the first rule is applied to an entire Feature Structure that is the output of rule $N - 1$.

For this paper, we evaluated the English output for several different genres, all of which approximately track parsing results for that genre. For written genres, we chose between 40 and 50 sentences. For speech transcripts, we chose 100 sentences—we chose this larger number because a lot of so-called sentences contained text with empty logical descriptions, e.g., single word utterances contain no relations between pairs of words. Each text comes from a different genre. For NEWS text, we used 50 sentences from the aligned Japanese-English data created as part of the JENAAD corpus (Utiyama and Isahara, 2003); the web text (BLOGs) was taken from some corpora provided by the Linguistic Data Consortium through the GALE (<http://projects.ldc.upenn.edu/gale/>) program; the LETTer genre (a letter from Good Will) was taken from the ICIC Corpus of Fundraising Texts (Indiana Center for Intercultural Communication); Finally, we chose two spoken language transcripts: a TELEphone conversation from the Switchboard Corpus (http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html) and one NARRative from the Charlotte Narrative and Conversation Collection (<http://newsouthvoices.uncc.edu/cncc.php>). In both cases, we assumed perfect sentence splitting (based on Penn Treebank annotation). The ICIC, Switchboard and Charlotte texts that we used are part of the Open American National Corpus (OANC), in particular the shared subcorpus designated described at <http://nlp.cs.nyu.edu/wiki/corpuswg/ULA-OANC-1> (Meyers et al., 2007).

Comparable work for English includes: (1) (Gabbard et al., 2006), a system which reproduces the

Genre	Prec	Rec	F
NEWS	$\frac{731}{815} = 89.7\%$	$\frac{715}{812} = 90.0\%$	89.9%
BLOG	$\frac{844}{704} = 83.4\%$	$\frac{899}{704} = 78.3\%$	80.8%
LETT	$\frac{392}{434} = 90.3\%$	$\frac{392}{449} = 87.3\%$	88.8%
TELE	$\frac{472}{604} = 78.1\%$	$\frac{472}{610} = 77.4\%$	77.8%
NARR	$\frac{732}{959} = 76.3\%$	$\frac{732}{964} = 75.9\%$	76.1%

Table 1: English Aggregate Scores

Corpus	Prec	Rec	F	Sents
NEWS	90.5%	90.8%	90.6%	50
BLOG	84.1%	79.6%	81.7%	46
LETT	93.9%	89.2%	91.4%	46
TELE	81.4%	83.2%	84.9%	103
NARR	77.1%	78.1%	79.5%	100

Table 2: English Score per Sentence

function tags of the Penn Treebank with 89% accuracy and empty categories (and their antecedents) with varying accuracies, the meaningful categories ranging from 82.2% to 96.3% accuracies.⁷; (2) Current systems that generate LFG F-structure such as (Wagner et al., 2007) which achieve an F score of 91.1 on the F-structure PRED relations, which are similar to our LOGIC1 relations.

4 Chinese XXXX

The Chinese XXXX program takes a Chinese Treebank-style syntactic parse as input, and attempts to determine the relations between the head and its dependents within each constituent. It does this by first exploiting the structural information and detecting six broad categories of syntactic relations that hold between the head and its dependents. These are *predication*, *modification*, *complementation*, *coordination*, *auxiliary*, and *flat*. Predication holds at the clause level between the subject and the predicate, where the predicate is considered to be the head and the subject is considered to the dependent. Modification can also hold mainly within NPs and VPs, where the dependents are modifiers of the NP head or adjuncts to the head verb. Coordination holds almost for all phrasal categories where each non-punctuation child within this constituent is ei-

⁷They get nearly perfect accuracy for null complementizer detection. It is unclear, in our view, what role null complementizers play (if any) in the interpretation of sentences.

ther conjunction or a conjunct. The head in a coordination structure is underspecified and can be either a conjunct or a conjunction depending on the grammatical framework. Complementation holds between a head and its complement, with the complement usually being a core argument of the head. For example, inside a PP, the preposition is the head and the phrase or clause it takes is the dependent. An auxiliary structure is one where the auxiliary takes a VP as its complement. This structure is identified so that the auxiliary and the verb it modifies can form a verb group in the XXXX framework. Flat structures are structures where a constituent has no meaningful internal structure, which is possible in a small number of cases. After these six broad categories of relations are identified, more fine-grained relation can be detected with additional information. The same set of relations are used for Chinese as with the English system. For example, a temporal adjunct of a VP is in an ADV relation with the verb. The fact that it is temporal is encoded elsewhere in the feature structure and will be omitted from the simplified representation provide here. Figure ?? is a sample 4-tuple for the Chinese sentence ??? which means ?????.

***** take example from answer key and encode it in the same tuples format as the English 3-tuples above*****

Description of some additional routines borrowed from the English system, especially the incorporation of Chinese NE

cite the Harper parser, especially the function tagged version (**** references ****)

– possibly include comparison with and without function tags???

Chinese parsing: – GALE paper: Recall 81.8, precision 82.2, F score: 82.0 – Other paper: 84.1, 86.3, 85.2 84.2 85.3, 84.8

*** site 5W paper *** regarding another system that uses these function tags, but I can't really site scores for them ***

5 Japanese XXXX

Michiko: 1) description of KNP output; 2) example with English Gloss; 3) Tuple output; 4) describe data

A description of the Japanese system (and how it is derived from KNP output – and Kyoto Corpus

Type	Prec	Rec	F
No Function Tags Version			
Aggreg	$\frac{751}{1258} = 59.7\%$	$\frac{751}{1226} = 61.3\%$	60.5%
Average	60.0%	61.5%	60.5%
Function Tags Version			
Aggreg	$\frac{917}{1292} = 71.0\%$	$\frac{917}{1226} = 74.8\%$	72.8%
Average	70.1%	72.6%	72.2%

Table 3: Chinese Results on 48 Sentences from the ??? Newswire Corpus

Type	Prec	Rec	F
Aggreg	$\frac{764}{843} = 91.0\%$	$\frac{764}{840} = 90.6\%$	90.8%
Average	90.7%	90.6%	90.6%

Table 4: Japanese Results on 40 Sentences from the JENAA Corpus

output)

We scored Japanese XXXX on forty sentences of the Japanese side of the JENAA data (25 of which are parallel with the English sentences scored).

(Kurohashi and Nagao, 1998) – describes KNP (Noro et al., 2005)

KNP is listed as 96.9% accuracy for segmentation, 91.32% for dependency, 60.07% for sentence

KNP parser (per user manual) Copyright University of Tokyo 2005. Version 2.0 dated September 7, 2005.

6 Conclusion

Acknowledgments

References

- C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of Coling-ACL98: The 17th International Conference on Computational Linguistics and the 36th Meeting of the Association for Computational Linguistics*, pages 86–90.
- Eugene Charniak. 2001. Immediate-head parsing for language models. In *Meeting of the Association for Computational Linguistics*, pages 116–123.
- Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.
- Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris Publications, Dordrecht.
- Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

- J. Foster and J. van Genabith. 2008. Parser Evaluation and the BNC: 4 Parsers and 3 Evaluation Metrics. In *LREC 2008*, Marrakech, Morocco.
- Ryan Gabbard, Mitchell Marcus, and Seth Kulick. 2006. Fully parsing the penn treebank. In *Proceedings of NAACL/HLT*, pages 184–191, New York, New York, USA. Association for Computational Linguistics.
- Zellig Harris. 1951. *Structural Linguistics*. University of Chicago Press, Chicago.
- Zellig Harris. 1968. *Mathematical Structures of Language*. Wiley-Interscience, New York.
- Jerry R. Hobbs and Ralph Grishman. 1976. The Automatic Transformational Analysis of English Sentences: An Implementation. *International Journal of Computer Mathematics*, 5:267–283.
- S. Kurohashi and M. Nagao. 1998. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of The 1st International Conference on Language Resources & Evaluation*, pages 719–724.
- C. Macleod, R. Grishman, and A. Meyers. 1998. COMLEX Syntax. *Computers and the Humanities*, 31:459–481.
- M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.
- A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004a. The NomBank Project: An Interim Report. In *NAACL/HLT 2004 Workshop Frontiers in Corpus Annotation*, Boston.
- A. Meyers, R. Reeves, Catherine Macleod, Rachel Szekeley, Veronkia Zielinska, and Brian Young. 2004b. The Cross-Breeding of Dictionaries. In *Proceedings of LREC-2004*, Lisbon, Portugal.
- A. Meyers, N. Ide, L. Denoyer, and Y. Shinyama. 2007. The shared corpora working group report. In *Proceedings of The Linguistic Annotation Workshop, ACL 2007*, pages 184–190, Prague, Czech Republic.
- E. Miltsakaki, A. Joshi, R. Prasad, and B. Webber. 2004. Annotating discourse connectives and their arguments. In A. Meyers, editor, *NAACL/HLT 2004 Workshop: Frontiers in Corpus Annotation*, pages 9–16, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.
- T. Noro, C. Koike, T. Hashimoto, T. Tokunaga, and Hozumi Tanaka. 2005. Evaluation of a Japanese CFG Derived from a Syntactically Annotated corpus with Respect to Dependency Measures. In *2005 Workshop on Treebanks and Linguistic theories*, pages 115–126.
- M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
- M. Surdeanu, R. Johansson, A. Meyers, Ll. Márquez, and J. Nivre. 2008. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *Proceedings of the CoNLL-2008 Shared Task*, Manchester, GB.
- M. Utiyama and H. Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *ACL-2003*, pages 72–79.
- J. Wagner, D. Seddah, J. Foster, and J. van Genabith. 2007. C-Structures and F-Structures for the British National Corpus. In *Proceedings of the Twelfth International Lexical Functional Grammar Conference*, Stanford. CSLI Publications.
- Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11:207–238.