Multilingual feature selection for a human sentence processing model

Marisa Ferrara Boston and Zhong Chen
Department of Linguistics and Languages
Michigan State University
East Lansing, MI 48824
{mferrara,chenzho3}@msu.edu

Abstract

This study develops a connection between human parsing preferences and feature selection rankings in a multilingual dependency parser. The results reveal that feature weights reflect the typological characteristics of three languages. Accounting for these differences leads to greater precision in modeling garden path data.

1 Introduction[1]

Human sentence processing models are often implemented by probabilistic parsers (Jurafsky, 1996; Roark, 2004; Hale, 2001; Demberg and Keller, 2007; Boston et al., 2008b), which use statistics derived from corpus data to determine sentential analyses. The probability space that informs these parsers can be partitioned into a wide variety of features that are based on characteristics of the internal parser state, the string, or any suitable combination that allows the parser to appropriately model the human sentence processor. This leads to a high-dimension feature space that requires exponential amounts of resources, and can be costly to compute.

Feature selection is a machine learning technique that helps to reduce the number of dimensions in a feature space, and thus avoid the “curse of dimensionality” (Guyon and Elisseeff, 2003). The technique determines the relevancy of features for a model according to a particular optimization function, and can be implemented using a variety of classification techniques. It also allows for better accuracy in parsing for natural language processing (Attardi et al., 2007), which indicates it may also be sensitive to the typological characteristics of languages.

In this paper we address the question of whether feature selection relates to typological differences between Chinese, German, and English. Further, we demonstrate that feature selection improves parsing accuracy for human sentence processing models. In the next section, we introduce our model.

2 A human sentence processing model

Our model is a statistical dependency parser that uses Dependency Grammar, a linguistic framework that specifies syntactic structure in terms of word-to-word connections. It is based on Nivre’s (2004) design, which defines parser states in terms of four data structures, detailed in Table 1. The σ data structure contains already-parsed words while the τ data structure lists unparsed words. The **h** and **d** functions aggregate the dependency information between words.

σ	A stack of already-parsed unreduced words.
τ	An ordered input list of words.
h	A function from dependent words to heads.
d	A function from dependent words to arc types.

Table 1: Nivre-defined parser configuration.

Parser states are manipulated with four operations, or transitions. The `Left-Arc` and `Right-Arc` transitions draw dependency relations between the elements at the top of σ and the top of τ. The `Shift` and `Reduce` transitions manipulate σ. This architecture renders the parser equivalent to a stack-based automaton.

The parser accurately models human garden path data in English, German, and Chinese (Boston and Hale, 2007; Boston et al., 2008a). Garden path sentences are temporary, local ambiguities the human sentence processor is susceptible to (Frazier, 1987).

[1] The authors thank John Hale and Rong Jin for their valuable comments and guidance.

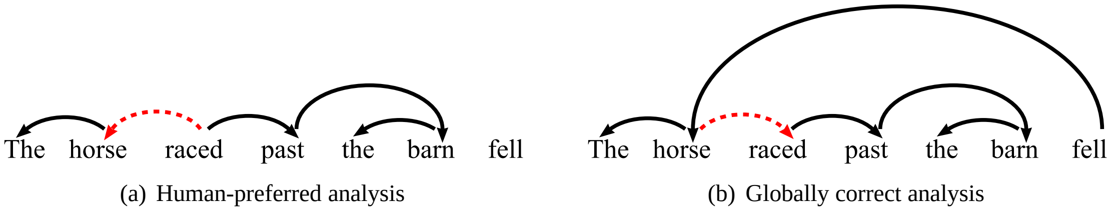

The horse raced past the barn fell
(a) Human-preferred analysis
The horse raced past the barn fell
(b) Globally correct analysis

Figure 1: Main verb-reduced relative ambiguity parsing preferences.

The sentence in Figure 1(a) exemplifies a strong garden path in English, where the verb “raced” is initially considered as the main verb of the sentence. This is marked by a red dashed arc to “horse”, signifying the noun is its dependent in the Dependency Grammar framework. Further input however reveals a second verb, “fell”, that is the actual main verb of the sentence. This reading requires an analysis of “raced” as the beginning of a reduced relative modifying the noun, which is marked by a red dotted line emanating from “horse” to “raced” (Figure 1(b)).

In this paper we explore the features that allow the parser to model a milder form of garden path, the prepositional phrase ambiguity, because the variation in attachment preferences for this garden path in the three languages demonstrates their diverse typological characteristics. Further, it is one of only a few garden paths that is available in all three languages.

3 Dependency Parsing Features

The parser uses probabilistic models, or features, to inform parser decisions. The features are trained on four newspaper corpora: the Wall Street Journal portion of the Penn Treebank (78,000 sentences) (Marcus et al., 1993), the Negra and Tiger Version 2.0 German treebanks (70,602 sentences) (Skut et al., 1997; Brants et al., 2004), and the Penn Chinese Treebank Version 4.0 (15,162 sentences) (Xue et al., 2004). The corpora were transformed into dependency format using Yamada’s Ptb-conv 3.0 tool (2004) for the English treebank, and Dubey’s (2004) and Ding’s (2006) head-finders for the German and Chinese treebanks, respectively.

Of the fourteen features that inform the dependency parser, six were found to be useful for distinguishing cross-linguistic differences in human sentence processing preferences (Table 2). Four of these features are state-based, or rely on the internal parser state information (Stack1, Stack2, TopLeft, and TopRight), and two are string-based (Distance and Position).

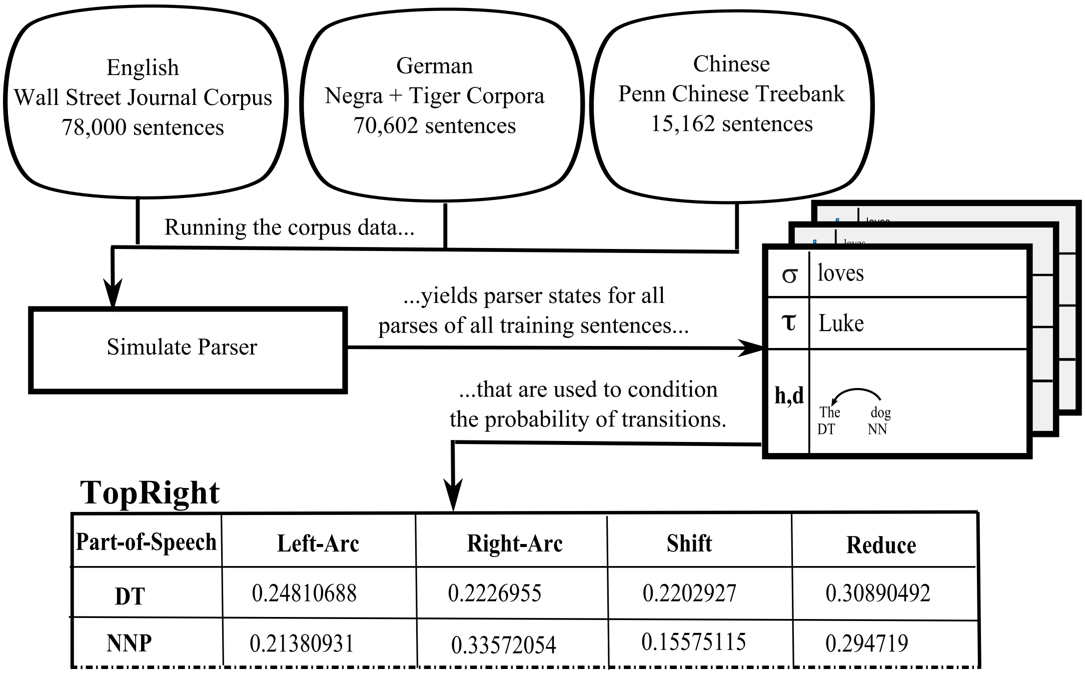
Feature	Description
Stack1	σ_1 and τ_1.
Stack2	$\sigma_{1,2}$ and τ_1.
TopLeft	The left-most dependent of σ_1.
TopRight	The right-most dependent of σ_1.
Distance	Surface distance between σ_1 and τ_1.
Position	The string position of τ_1.

Table 2: Dependency Parsing Features

The flowchart in Figure 2 depicts the feature-making process for the dependency parser. The converted training data is input to a parser simulation, which outputs state and transition information for each sentence. This collection of parser states and transitions is used to derive the probabilities that inform parser decisions in the form of features. Probabilities for the TopRight feature are shown in the figure. The probabilities indicate that when the right-most element of the top of σ is a determiner (DT), the most probable transition is `Reduce`. When it is a proper noun (NNP), the most probable transition is `Right-Arc`. The process is repeated for all features for all languages to derive a collection of probabilistic models that inform parser decisions.

4 SVMs and Feature Selection

Because of the number of features for each language, we use feature selection to decrease the dimensions and to interpret the features in terms of their typological characteristics. This is implemented with the LIBLINEAR Support Vector Machine (SVM) classifier (Lin et al., 2008). SVMs are often used to induce classifiers for determin-


English
Wall Street Journal Corpus
78,000 sentences
German
Negra + Tiger Corpora
70,602 sentences
Chinese
Penn Chinese Treebank
15,162 sentences
Running the corpus data...
Simulate Parser
...yields parser states for all
parses of all training sentences...
σ
loves
τ
Luke
h,d
The
DT
dog
NN
...that are used to condition
the probability of transitions.
TopRight

Part-of-Speech	Left-Arc	Right-Arc	Shift	Reduce
DT	0.24810688	0.2226955	0.2202927	0.30890492
NNP	0.21380931	0.33572054	0.15575115	0.294719

Figure 2: The feature-making process for the dependency parser.

istic parsing, and provide high accuracy in dependency parsing (Hall et al., 2006; Attardi et al., 2007). They combine a maximum margin strategy with kernel functions to map the original feature space into a higher-dimensional space (Vapnik, 1995; Vapnik, 1998).

LIBLINEAR has a least squares SVM option based on a logistic regression function for multi-class classification (Lin et al., 2008), depicted in Equation 1.

$$\min_{\mathrm{w}} f(\mathrm{w}) \equiv \frac{1}{2}\mathrm{w}^T\mathrm{w} + C\sum_{i=1}^{l}\log(1+\mathrm{e}^{-\mathrm{y}_i\mathrm{w}^T\mathrm{x}_i}) \quad (1)$$

The second term in Equation 1 can be considered a loss function, which the least squares SVM optimizes according to the function in Equation 2.

$$\min_{\mathrm{w}} f_{\mathit{2}}(\mathrm{w}) \equiv \frac{1}{2}\mathrm{w}^T\mathrm{w} + C\sum_{i=1}^{l}(\max(0, 1-\mathrm{y}_i\mathrm{w}^T\mathrm{x}_i))^{\mathit{2}} \quad (2)$$

For this study, the input x_i is the series of configurations derived from the training data with transition prediction probabilities for each feature. The class label y_i is the set of four transitions (`Left-Arc`, `Right-Arc`, `Shift`, and `Reduce`). The output w^* is an optimal weight for each feature.

The wrapper feature selection method (Kohavi and John, 1997) uses weights output from a classifier as a measure of the relevance of features (Guyon and Elisseeff, 2003). The weights output from LIBLINEAR, w^*, therefore represent a feature selection technique, and the weights straight-forwardly correlate to the utility of the features for each language.

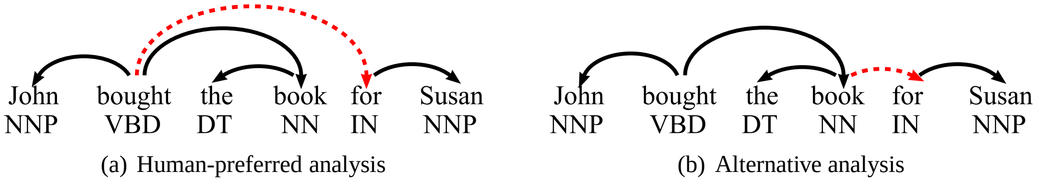
5 Results and Discussion

Table 3 provides the LIBLINEAR-derived weights for each feature in each language. The features are organized by rank for the languages, with feature weights adjacent. The weights determine the strength of the feature for informing parser decisions during parsing, with Position in English having the strongest weight overall (1.77), and Distance in English having the lowest (-1.89).

In the following sections, we interpret the weights with respect to typological differences between the three languages. In particular, the results are compared to the prepositional phrase (PP) attachment ambiguity. Figure 3 shows an English example of the ambiguity, where the PP can attach high, as in 3(a) or low, as in 3(b). The two attachments give rise to two different but grammatical readings of the sen-

Relative Rank	Chinese		German		English	
1	Position	0.43	Distance	0.48	Position	1.77
2	TopLeft	0.32	Stack1	0.36	Stack1	0.40
3	Stack2	0.16	Position	-0.02	TopRight	0.22
4	TopRight	-0.16	Stack2	-0.34	Stack2	-0.40
5	Stack1	-0.20	TopLeft	-1.41	TopLeft	-0.58
6	Distance	-0.57	TopRight	-1.41	Distance	-1.89

Table 3: Features Weights and Rankings



(a) Human-preferred analysis

(b) Alternative analysis

Figure 3: PP-attachment ambiguity parsing preferences in English.

tence, where the PP either modifies the verb or the noun. For example, Figure 3(a) represents the benefactive reading of the sentence, where the PP modifies the verb "bought". Figure 3(b) represents the alternative reading where the PP modifies the noun "book". In English, the human sentence processor prefers the high attachment (Bever, 1970), but as will be described below, this is not the case for all three languages.

5.1 TopLeft in English vs. TopRight in Chinese

An interesting difference between the English and Chinese rankings is the relative weights of the TopLeft and TopRight features, described in Table 2. In Table 3, TopLeft has a positive weight and is ranked highly for Chinese, but low for English. Alternatively, TopRight has a positive weight in English, but a negative weight in Chinese. This disparity may arise from one of the main typological differences between Chinese and English, head position. Chinese is a head-final language (Huang and Li, 1995), which means that many of the heads occur to the right of the clauses. Therefore, the left-most dependents of the top element of σ would provide more information to a parser than the right-most dependents. English, on the other hand, is head-initial. In this case, we expect the right-most dependents to be more informative.

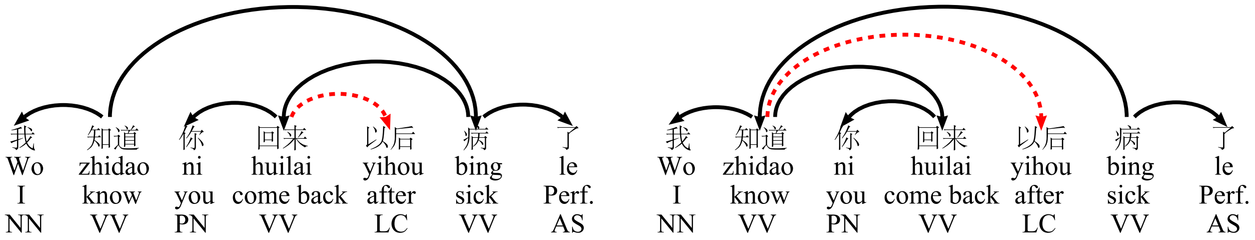
The weights not only reflect this typological disparity, but they also result in a better processing model for human garden paths. Figures 4(a) and 4(b) demonstrate alternative readings of the Chinese PP-Attachment ambiguity. The first is glossed as "I know that you became sick after coming back", whereas the alternative is "I became sick after I knew you came back". Unlike English, Chinese prefers the low-attachment for the PP, as in Figure 4(a).[2]

This difference in attachment preferences is predicted by the relative rankings of TopLeft and TopRight as well. As the transition probabilities in Table 5 demonstrate, TopLeft chooses the human-preferred attachment in Chinese but not in English, while TopRight chooses the human-preferred attachment in English but not in Chinese. Figure 6 depicts this preference at the parser-internal level, where the transitions that lead to the human-preferred analysis have higher probabilities for each language. This indicates that the weights from this feature selection method allow the parsing model to accurately predict the PP-Attachment preferences in both Chinese and English.

5.2 Stack2 in Chinese vs. Stack1 in German

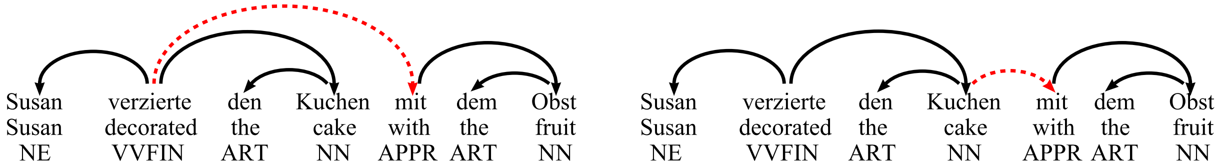
One of the main differences between the German and Chinese weights is for the Stack1 and Stack2 features. Although we would expect Stack2 to be ranked more highly across languages because

[2] Based on preferences from a Chinese native speaker judgment task (n=10).



(a) Human-preferred analysis (b) Alternative analysis

Figure 4: PP-attachment ambiguity parsing preferences in Chinese.

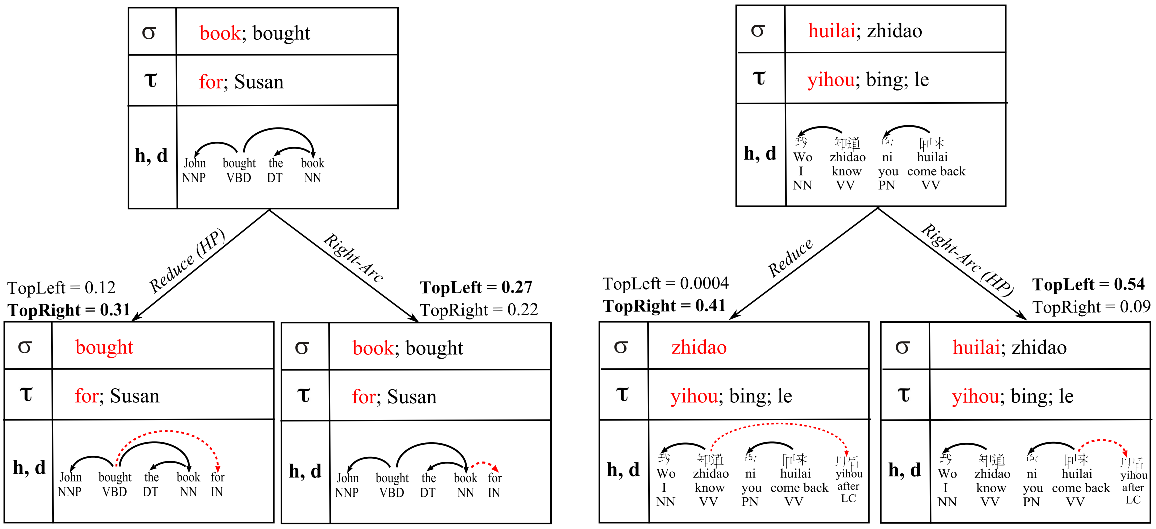


(a) Human-preferred analysis (b) Alternative analysis

Figure 5: PP-attachment ambiguity parsing preferences in German.

	Chinese		English	
Feature	TopLeft	TopRight	TopLeft	TopRight
Human-preferred	**0.54**	0.09	0.12	**0.34**
Alternative	0.0004	**0.41**	**0.27**	0.20

Table 4: TopLeft vs. TopRight Transition Probabilities.


σ
τ
h, d
book; bought
for; Susan
John NNP bought VBD the DT book NN
Reduce (HP)
Right-Arc
TopLeft = 0.12
TopRight = 0.31
TopLeft = 0.27
TopRight = 0.22
bought
for; Susan
John NNP bought VBD the DT book NN for IN
book; bought
for; Susan
John NNP bought VBD the DT book NN for IN
huilai; zhidao
yihou; bing; le
我 Wo I NN 知道 zhidao know VV 你 ni you PN 回来 huilai come back VV
Reduce
Right-Arc (HP)
TopLeft = 0.0004
TopRight = 0.41
TopLeft = 0.54
TopRight = 0.09
zhidao
yihou; bing; le
我 Wo I NN 知道 zhidao know VV 你 ni you PN 回来 huilai come back VV 以后 yihou after LC
huilai; zhidao
yihou; bing; le
我 Wo I NN 知道 zhidao know VV 你 ni you PN 回来 huilai come back VV 以后 yihou after LC

(a) English (b) Chinese

Figure 6: State space diagrams of the ambiguous regions for the relevant transitions. While TopRight chooses the human-preferred (HP) transition for English, TopLeft chooses the HP transition for Chinese.

it takes into account more information, we actually find that the Stack1 feature often performs better at modeling garden path data cross-linguistically (Boston et al., 2008a). The fact that Chinese prefers

	Chinese		German	
Feature	Stack1	Stack2	Stack1	Stack2
Human-preferred	0.20	**0.53**	**0.43**	0.37
Alternative	**0.40**	0.21	0.48	**0.49**

Table 5: Stack1 vs. Stack2 Transition Probabilities.

Stack2 may be an artifact of a typological difference between the two languages. Chinese is an analytic language which relies more on the use of particles (i.e., separate function words) than inflection to provide meaning (Comrie, 1989). Therefore, the parser would require more stack information to derive meaning than it would for a synthetic language like German.

The difference in weights for these features predicts the PP-Attachment preferences for both languages as well. German, like English, prefers high-attachment for prepositions, where they modify the verb rather than the noun (Figure 5(a)) (Agricola, 1968). The alternative reading in Figure 5(b), while available, would only arise for particular lexical items. Stack1, which is given a positive weight for German, is able to predict the human-preferred analysis because the human-preferred transition's probability is higher in Table 5. Stack2, however, predicts the alternative analysis. In Chinese, the opposite is the case: Stack2 predicts the human-preferred reading, whereas Stack1 does not. These results demonstrate that the feature selection method is sensitive to the typological differences between isolating and synthetic languages.

5.3 Distance in German vs. Position in English

English is a configurational language, where the word order is highly constrained within a sentence (Ross, 1967). German, on the other hand, does not constrain many sentential structures to set positions (Uszkoreit, 1987). This typological difference is reflected by the feature selection ranking in Table 3. In a configurational language, the position a word has in a sentence would be an important characteristic, which is reflected by the high weight and rank of the Position feature for English. In a language like German, however, Position would be less informative than the relative distance between two words, which is reflected by the high weight of Distance but low weight of Position.

The feature ranking once again leads to the correct parsing of the PP-attachment ambiguities for the two sentences. Distance, which has a high, positive weight for English, predicts the high-attachment preference, whereas Position predicts the low-attachment preference (Table 6). On the other hand, Position chooses the human-preferred high-attachment for German, but Distance does not. These results indicate that the feature selection method is able to distinguish the differences between configurational and non-configurational languages, and counsels the parser to choose the human-preferred attachment in each language.

6 Conclusion

Our results demonstrate that a wrapper feature selection method implemented with an SVM learner is able to distinguish three main typological characteristics of the languages tested: head position in English and Chinese, morphological constituents in Chinese and German, and configurational preferences in German and English. The language-specific weights for the features additionally counsel the parser to choose the human-preferred transition for the PP-Attachment ambiguity despite attachment preference differences. This indicates that feature selection can reveal how typological conditions interact with human parsing preferences in a sentence processing model, and further supports the use of probabilistic parsers for sentence processing research (Jurafsky, 1996).

7 Acknowledgements

This work was partially supported by NSF IGERT grant DGE-0114378.

	English		German	
Feature	Distance	Position	Distance	Position
Human-preferred	0.01	**0.32**	**0.29**	0.24
Alternative	**0.24**	0.07	0.24	**0.32**

Table 6: Distance vs. Position Transition Probabilities.

References

E. Agricola. 1968. *Syntaktische Mehrdeutigkeit (Polysyntaktizitöt) bei der Analyse des Deutschen und des Englischen: Schriften zur Phonetik, Sprachwissenschaft und Kommunikationsforschung*. Akademie Verlag, Berlin.

G. Attardi, F. Dell'Orletta, M. Simi, A. Chanev, and M. Ciaramita. 2007. Multilingual dependency parsing and domain adaptation using desr. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

T. G. Bever, 1970. *The cognitive basis for linguistic structures*, pages 277–360. Wiley and Sons, New York.

M.F. Boston and J.T. Hale. 2007. Garden-pathing in a statistical dependency parser. In *Proceedings of the Midwest Computational Linguistics Conference (MCLC) 2007*.

M.F. Boston, Z. Chen, and J.T. Hale. 2008a. Modeling garden paths in a statistical dependency parser: Chinese, german, and english. Poster presented at the CUNY Human Sentence Processing Conference in Chapel Hill, N.C., March 13-15.

M.F. Boston, J.T. Hale, R. Kliegl, U. Patil, and S. Vasishth. 2008b. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*.

S. Brants, S. Dipper, P. Eisenberg, S. Hansen-Schirra, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit. 2004. TIGER: Linguistic interpretation of a german corpus. *Research on Language and Computation*, 2:597–619.

B. Comrie. 1989. *Language Universals and Linguistic Typology*. Basil Blackwell, Oxford.

V. Demberg and F. Keller. 2007. Eye-tracking evidence for integration cost effects in corpus data. In *Proceedings of the 29th meeting of the Cognitive Science Society (CogSci-07)*, Nashville, Tennessee. Cognitive Science Society.

Y. Ding. 2006. *Machine translation using probabilistic synchronous dependency insertion grammars*. Ph.D. thesis, University of Pennsylvania.

A. Dubey. 2004. *Statistical Parsing for German: Modeling syntactic properties and annotation differences*. Ph.D. thesis, Saarland University, Germany.

L. Frazier. 1987. Sentence processing: A tutorial review. In M. Coltheart, editor, *Attention and performance XII*, pages 559–586. Lawrence Erlbaum Associates.

I. Guyon and A. Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

J.T. Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of NAACL 2001*, pages 1–8.

J. Hall, J. Nivre, and J. Nilssson. 2006. Discriminative classifiers for deterministic dependency parsing. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*.

C.T. James Huang and Y. H. Audrey Li. 1995. Recent generative studies in Chinese syntax. In C.T. James Huang and Y. H. Audrey Li, editors, *New Horizons in Chinese Linguistics*, pages 49–95. Kluwer, Dordrecht.

D. Jurafsky. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20:137–194.

R. Kohavi and G. John. 1997. Wrappers for feature selection. *Artificial Intelligence*, 97(1-2):273–324.

C.-J. Lin, R. C. Weng, and S. S. Keerthi. 2008. Trust region newton method for large-scale regularized logistic regression. *Journal of Machine Learning Research*, 9.

M. P. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

J. Nivre. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the Workshop on Incremental Parsing (ACL)*, pages 50–57.

B. Roark. 2004. Robust garden path parsing. *Natural Language Engineering*, 10(1):1–24.

J. Ross. 1967. *Constraints on variables in syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, Washington, DC.

H. Uszkoreit. 1987. *Word order and constituent structure in German*. CSLI, Stanford, CA.

V. Vapnik. 1995. *The nature of statistical learning theory*. Springer.

V. Vapnik. 1998. *Statistical learning theory*. John Wiley and Sons, New York.

N. Xue, F. Xia, F. Chiou, and M. Palmer. 2004. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 10(4):1–30.

H. Yamada. 2004. Ptb-conv 3.0: Dependency generator tool. Available online at http://www.jaist.ac.jp/ hyamada/.