Machine Learning and the Cognitive Basis of Natural Language

Shalom Lappin

Department of Philosophy
King's College London

Abstract

Machine learning and statistical methods have yielded impressive results in a wide variety of natural language processing tasks. These advances have generally been regarded as engineering achievements. In fact it is possible to argue that the success of machine learning methods is significant for our understanding of the cognitive basis of language acquisition and processing. Recent work in unsupervised grammar induction is particularly relevant to this issue. It suggests that knowledge of language can be achieved through general learning procedures, and that a richly articulated language faculty is not required to explain its acquisition .

1 Introduction

The past fifteen years have seen a massive expansion in the application of information theoretic and machine learning (ML) methods to natural language processing. This work has yielded impressive results in accuracy and coverage for engineering systems addressing a wide variety of tasks in areas like speech recognition, morphological analysis, parsing, semantic interpretation, and dialogue management. It is worth considering whether the inductive learning mechanisms that these methods employ have consequences not simply for natural language engineering, but also for our understanding of the cognitive basis of human language acquisition and processing.

A view common among computational linguists is that the information theoretic methods used to construct robust NLP systems are engineering tools. On this approach language engineering does not bear directly on the cognitive properties of grammar and human language processing. Shieber (2004 pc) suggests that the success of information theoretic methods in NLP may have implications for the scientific understanding of natural language. Pereira (2000) argues that current work on grammar induction has revived Harris' (1951), (1991) program for combining formal grammar and information theory in the study of language.

Most machine learning has used supervised learning techniques. These have limited implications for theories of human language learning, given that they require annotation of the training data with the structures and rules that are to be learned. However, recently there has been an increasing amount of promising research on unsupervised machine learning of linguistic knowledge. The results of this research suggest the computational viability of the view that general cognitive learning and projection mechanisms rather than a richly articulated language faculty may be sufficient to support language acquistion and intepretation.

In Section 2 I briefly consider the poverty of stimulus argument that has been traditionally invoked to motivate a distinct language faculty. Section 3 reviews

major developments in the application of supervised machine learning to different areas of NLP. Section 4 considers some recent work in unsupervised learning of NLP systems. In Section 5 I attempt to clarify the central issues in the debate between the distinct language faculty and generalized learning model approaches. Finally Section 6 states the main conclusions of this discussion and suggests directions for future work.

2 The Poverty of Stimulus Argument for a Language Faculty

Chomsky (1957), (1965), (1981), (1986), (1995), (2000) argues for a richly articulated language acquisition device to account for the speed and efficiency with which humans acquire natural language on the basis of "sparse" evidence. This device defines the initial state of the language learner. Its structures and constraints represent a Universal Grammar (UG) that the learner brings to the language acquisition problem. In earlier versions of the language faculty hypothesis (as in Chomsky (1965), for example) this device is presented as a schema that defines the set of possible grammars and an evaluation metric that ranks grammars conforming to this schema for a particular natural language. Since Chomsky (1981) UG has been described as an intricate set of principles containing parameters at significant points in their specification. On the Principles and Parameters (P&P) approach exposure to a very limited amount of linguistic data allows the learner to determine parameter values in order to produce a grammar for his/her language. Chomsky (2000) (p. 8) describes this Principles and Parameters model in the following terms.

> We can think of the initial state of the faculty of language as a fixed network connected to a switch box; the network is constituted of the principles of language, while the switches are options to be determined by experience. When switches are set one way, we have Swahili; when they are set another way, we have Japanese. Each possible human language is identified as a particular setting of the switches—a setting of parameters, in technical terminology. If the research program succeeds, we should be able literally to deduce Swahili from one choice of settings, Japanese from another, and so on through the languages that humans acquire. The empirical conditions of language acquisition require that the switches can be set on the basis of the very limited properties of information that is available to the child.

The language faculty view relies primarily on the poverty of stimulus argument to motivate the claim that a powerful task-specific mechanism is required for language acquistion. According this argument the complex grammar that a child achieves within a short period, with very limited data cannot be explained through general learning procedures of the kind involved in other cognitive tasks.

This is, at root, a "What else could it be?" argument. It asserts that, given the complexity of grammatical knowledge and the lack of evidence for discerning its properties, we are forced to the conclusion that much of this knowledge is

not learned at all but must already be present in the initial design of the language learner. The basis for this claim is the assumption that first language learners have access to very limited amounts of data, where this data is, in general, free of negative information (corrections), and inadequate to support inductive projection of grammars under the attested conditions of acquisition. In fact, this assumption has been increasingly subject to effective challenges. So, for example, Pullum and Scholz (2002) argue that there is no poverty of stimulus in language learning. The linguistic data to which children are exposed is far richer than poverty of stimulus theorists suggest. Similarly Chouinard and Clark (2003) present the results of detailed case studies showing that parents provide children with a wealth of negative evidence that plays a significant role in the language acquisition process.

One of Chomsky's (1957) original arguments against statistical induction of grammar turns on the absence of many (most) grammatical structures in corpora. This is an instance of the sparse data problem. Pereira (2000) points out that smoothing techniques, introduced in Good (1953), permit the assignment of probability values to unobserved linguistic events. When enriched with smoothing, statistical modelling of NL learning can deal effectively with sparse data. This development undermines the poverty of stimulus argument from the perspective of the power of task-general computational devices for inductive learning. A plausible alternative to the UG hypothesis is that, given reasonable initial settings for a set of linguistic categories and a search space for grammar rules, general learning strategies of the sort used in AI and NLP can account for language acquisition and processing without the assumption of a task-specific language learning device.

3 Supervised Learning

In supervised learning a corpus is annotated with the structures or features that the system must learn to recognize. This corpus provides the gold standard for training and evaluation. Symbolic machine learning algorithms extract rules or classifying procedures from the corpus to identify the marked properties in unlabelled corpora. Statistically driven learning methods construct models from the training corpus that determine the probability distributions for the marked properties over a set of possible linguistic contexts.

Supervised acquistion of part of speech (POS) tags has produced successful broad coverage taggers. Church (1988), (1992) proposes a stochastic POS tagger. Brill (1992), (1994) constructs a transformation-based tagger that applies rule-based machine learning. These systems and other current taggers generally have a lower bound of 96% accuracy evaluated against POS annotated corpora like the British National Corpus (BNC) and the Penn Tree Bank.

Probabilistic parsing is another domain in which supervised learning has yielded impressive results Charniak (1997) and Collins (1998) develop Probabilistic Context Free Grammars (PCFGs) which extract CFG rules with specified probability values from the Penn Tree Bank. The PCFG rules are lexicalized to identify the lexical head of a constituent and dependency relations. Charniak's PCFG includes

lexical subcategorization features. Collins' grammar represents argument-adjunct distinctions. Both grammars achieve recall and precision scores of close to 90% on the Wall Street Journal (WSJ) of the Penn Tree Bank.

Clark and Curran (2004) describe a Maximum Entropy statistical Combinatory Categorical Grammar (CCG) parser. The CCG parser represents unbounded dependencies as well as local function-argument structures. It is trained on a corpus annotated with CCG lexical tags and lexical dependency structures (Hockenmaier and Steedman (2002)). Dependency structures are represented as features, and the system computes a model of the most likely set of dependencies, given a sentence, from the annotated gold standard. The authors report recall and precision scores of over 86% for labelled dependencies and over 92% for unlabelled dependencies on a test set from the WSJ.

Similar methods have been used for wide coverage semantic representation. Bos, Clark, Curran, Hockenmaier and Steedman (2004) construct a system for assigning logical forms to the parse structures of a statistical CCG. Typed λ-terms are assigned to tagged lexical items, and λ-terms are composed for phrases in accordance with the dependency relations of the parse structure. The reduced λ-terms for sentences are first-order formulas with Davidsonian event variables. The authors report that the system assigns well-formed logical forms to 92% of parse input when tested on WSJ text. These logical forms are not sufficiently expressive for many sentences. They do not handle higher-order quantificational determiners or intensional modifiers. However, the system could be extended to generate more realistic representations to deal with these phenomena.

Supervised machine learning is also being fruitfully applied to ellipsis resolution. Nielsen (2003) uses a set of alternative machine learning algorithms to identify elided VP's in a sample of the BNC on the basis of contexts specified with POS features. Transformation Based Learning (TBL) obtained an F score (a weighted average of recall and precision) of 76.61%, and a Maximum Entropy system yielded 76.01%. Nielsen (2004) tests a subset of these algorithms on a much larger corpus containing text from both the BNC and the Penn Tree Bank, and using full parse structures as input. The Maximum Entropy system achieved an F score of approximately 71% for this text.

Liakata and Pulman (2004) propose a method for learning a domain theory from a text by Inductive Logic Programming (ILP). The theory consists of a set of Horn clause rules that express relations among the main properties, relations, and entities cited in the text. It is encoded as a probabilistic Finite State Automata whose transition arcs are labelled with simple Horn clauses that are assigned probability values. The method is applied to text from the Penn Tree Bank and from the ATIS corpus.

Fernández, Ginzburg and Lappin (2005) apply supervised machine learning techniques to the task of identifying the intepretational type of non-sentential utterances (NSUs) in dialogue from a set of possible readings. They construct a procedure for automatically annotating a corpus of 1109 NSUs, extracted from the BNC, with 9 features. The automatic feature annotation achieves 89% accuracy when evaluated against a randomly selected 10% sample of the NSU corpus.

They apply four machine learning algorithms, C4.5-based decision trees (Quinlan (1993)), SLIPPER, a greedy rule learning system with confidence rated rule boosting (Cohen and Singer (1999)), TiMBL, a memory-based learner (Daelemans, Zavrel, van der Sloot and van den Bosch (2003)), and a maximum entropy system (Le (2003)), to this annotated data set. The four ML algorithms achieve F scores of between 87% and 90%.

The striking achievements of supervised learning in NLP show that powerful symbolic and statistical induction procedures applied to corpora can acquire knowledge of the structure and interpretation of NL quickly and efficiently. However, these procedures require that the information to be learned is explicitly represented in the training data. This information defines the initial conditions from which learning proceeds. Therefore the success of supervised learning in NLP does not, in itself, provide decisive evidence against the assumption of a language specific learning device.

4 Unsupervised Learning

In unsupervised learning the training corpus is not annotated with the structures or features to be acquired. The search space is constrained to a set of possible entities to be learned (such as lexical classes, constituent structures, CFG rules, etc.). Learning is achieved largely through the identification of distributional and clustering patterns by which classes of similar objects are recognized. Successful acquisition of linguistic structure and content through unsupervised methods would provide significant motivation for the view that weak assumptions concerning linguistic structure or rule hypothesis space, combined with general cognitive mechanisms of induction and projection are sufficient for language learning.

Goldsmith (2001) uses Minimal Description Length (MDL) criteria to select between alternative morphological anayses of words in unmarked text. He employs several probabilistic methods as heuristic procedures to generate morphological signatures that split the words of a corpus into stems and suffixes. Goldsmith uses MDL to identify the optimal morphological system as the one which provides the most compressed description of the full range of data with the most compact set of signatures. He reports that in a test set of 1000 alphabetically consecutive words taken from a 500,000 word English corpus 82.9% of the analyses that his system produces are good.

Schone and Jurafsky (2001) improve on Goldsmith's results with an algorithm for unsupervised morphological analysis of stems and affixes that combines three main factors to identify pairs of morphologically related words. It makes use of induced semantic relations among lexical items, othographic connections among words measured in terms of transformability operations, and local syntactic contexts of distribution common to attested morphological variants. It also uses weighted transitive closures of identified morphological relations among words to extend sets of variants. Schone and Jurafsky test their algorithm on English (6.7 million words of newswire text), German (2.3 million words), and Dutch (6.7 million words). They use the hand annotated CELEX lexicon for each language

as the gold standard for evaluation. Their algorithm gives an F score of 88.1 % for English suffixes, compared to 81.8% that Goldsmith's system achieves for the same text, 92.3% for German (Goldsmith 84%), and 85.8% for Dutch (Goldsmith 75.8%).

Clark (2000) describes an unsupervised distributional method for identifying lexical syntactic categories through clustering. The tag set which this method generates compares favourably to the POS tag set of the BNC (CLAWS). Clark reports that a probabilistic finite state model for tagging gave lower perplexity values for the tag set obtained by unsupervised learning than for CLAWS, showing that the unsupervised tags express significant distributional generalizations. Reliable unsupervised POS tagging provides the basis for unsupervised grammar acquisition.

Clark (2001) describes an unsupervised system that learns a stochastic CFG from a tagged text using distributional information concerning local contexts for tag sequences. This system gives a somewhat disappointing F score of 41% on the ATIS corpus.

The grammar induction system proposed by Klein and Manning (2002) is an unsupervised method that learns constituent structure from part of speech (POS) tagged input by assigning probability values to sequences of tagged elements as constituents in a tree. They bias their model to parse all sentences with binary branching trees, and they use an Expectation Maximization (EM) algorithm to identify the most likely tree structure for a sentence. Their method relies on recognizing (unlabeled) constituents through distributional clustering of corresponding sequences in the same contexts, where a tree structure is constrained by the requirement that sister constituents do not overlap (have non-null intersections of elements).

The Klein and Manning procedure achieves an F score of 71% on WSJ text, using Penn Tree Bank parses as the standard of evaluation. This score is impressive when one considers a limitation that the evaluation procedure imposes on their system. The upper bound on a possible F-score for their algorithm is 87% because the Penn treebank assigns non-binary branching to many constituents. In fact, many of the system's "errors" are linguistically viable parses that do not conform to analyses of the Penn Treebank. So, for example, the Treebank assigns flat structure to NPs, while the Klein and Manning procedure analyses NPs as having iterated binary branching. Parses of the latter kind can be motivated on linguistic grounds.

One might object to the claim that Klein and Manning's parser is genuinely unsupervised on the grounds that it uses the POS tagging of the Penn Treebank as input. They run an experiment in which they apply their procedure to WSJ text annotated by an unsupervised tagger, and obtain an F score of 63.2%. However, as they point out, this tagger is not particularly reliable. Other unsupervised taggers, like the one that Clark (2000) describes, yield very encouraging results, and outputs of these taggers might well permit the parser to perform at a level comparable to that which it achieves with the Penn Treebank tags.

Klein and Manning (2004) describe a probabilistic model for unsupervised learning of lexicalized head dependency grammars. The system assigns probabil-

ities to dependency structures for sentences by estimating the likelihood that each word in the sentence is a head that takes a specified sequence of words to its left and to its right as argument or adjunct dependents. The probabilities are computed on the basis of the context in which the head appears, where this context consists of the words (word classes) occurring immediately on either side of it. Like the constituent structure model, their dependency structure model imposes binary branching as a condition on trees. The procedure achieves an F score of 52.1% on Penn Treebank test data. This result underrates the success of the dependency model to the extent that it relies on strict evaluation of the parser's output against the dependency structures of the Penn Treebank, in which NPs are headed by N's. Klein and Manning report that in many cases their dependency parser identifies the determiner as the head of the NP, and this analysis is, in fact, linguistically viable.

When the dependency system is combined with their unsupervised constituency grammar, the integrated model outperforms each of these systems. In the composite model the score for each tree is computed as the product of the individual models that the dependency grammar and the constituency structure grammar generate. This model uses both constituent clustering and the probability of head dependency relations to predict binary constituent parse structure. It yields an F score of 77.6% with Penn Treebank POS tagging. It also achieves an F score of 72.9% with an unsupervised tagger (Schuetze (1995)).

This work on unsupervised grammar induction indicates that it is possible to learn a grammar that identifies complex syntactic structure with a relatively high degree of accuracy using a model containing a weak assumptions concerning syntactic structure, specifically the restriction of binary branching, a non-overlap constraint for constituents, and limited conditions on head argument/adjunct dependency relations.

Recent research on unsupervised grammar induction offers support for the view that knowledge of language can be achieved through general machine learning methods on the basis of a minimal set of initial settings for possible linguistic categories and rule hypotheses. This work suggests a sequenced boot strap model of language learning in which each level of structure acquired provides the input to a higher successor component of grammar. In at least some cases both the basic categories and the hypothesis space might be derived from more general cognitive processing patterns (like the binary branching trees that Klein and Manning (2002), (2004) generate for constituent and dependency structures).

By contrast to machine learning NLP only a few small scale prototypes for grammar acquistion with the P&P model have been implemented, most notably by Fong (1991), (2005). They have not been extensively applied to real data from linguistic corpora. No robust, wide coverage systems using this model have been designed or tested. A common response of P&P advocates to this criticism is to argue that they are concerned with characterizing the set of possible languages rather than to develop broad coverage parsers for particular languages. This is, in effect, a circular argument, as it reduces to the assertion that the proper objective of linguistic theory is to specify the properties of the language faculty that constitute UG. But it is the existence of such a faculty, and hence the motivation for this

enterprise that is at issue in the current discussion.

Some critics of the view that machine learning can provide a viable model of human grammar acquisition have argued that an ML system learns only from a corpus of grammatical sentences, and so it is limited to recognizing well formed phrases.[1] This claim is misconceived. The parser that an ML system produces can be engineered as a classifier to distinguish grammatical and ungrammatical strings, and it can identify the structures assigned to a string under which this distinction holds. Such a classifier can also be refined to provide error messages that identify those properties of an unsuccessful parse that cause a sentence to be identified as ungrammatical. It is important to note that such a classifier is generated by machine learning applied to a data set, where the learning task is defined by the (relatively) weak grammatical assumptions of the language model that the learner invokes.

5 Clarifying the Issues of the Discussion

It has occasionally been suggested that the debate between ML-based NLP and P&P involves a choice between symbolic and non-sympbolic approaches to computing. In fact the role of symbolic computing methods in NLP is not an issue in this discussion. Many machine learning algorithms produce rule systems for classification, as is the case with TBL, ILP, and SLIPPER. Statistically driven learning procedures apply to data annotated with high-level linguistic information and, in many cases, produce abstract representations, including syntactic structures, semantic role information, and logical forms.

It is important to recognize that the need to assume a rich set of innate learning principles is not in dispute. The machine learning approach posits a powerful induction and projection device with initial categories and hypothesis settings. However, unlike the language faculty view, it takes most of this innate mechanism to be a set of general cognitive capacities that apply to a wide variety of learning and processing problems rather than a task-specific device that applies only to natural language learning. The constraints that the model imposes on the set of grammatical structures are minimal and define a large space of possible languages and grammars.

The debate does not concern ”mentalism”. The role of internal states and operations of the language learner is not in question. The machine learning approach requires them as part of its model of the computations involved in learning and processing. Whether these states and operations are conceived of as mental properties and events or as neurological phenomena is not relevant to either side of the debate.

The primary question at issue between the ML and P&P approaches is the status of the poverty of stimulus argument as the basis for postulating a highly articulated language faculty (UG) for language learning. The success of machine

[1] See, for example, Carson Schutze’s contributions of April 20 and May 5, 2005, on the *Linguist List*, to the discussion of Richard Sproat and Shalom Lappin, ”A Challenge to the Minimalist Community”, *Linguist List*, April 5, 2005.

learning methods shows that this argument does not go through. The fact that no robust P&P system for acquiring grammar from real data has yet been implemented casts serious doubt on the computational credibility of this approach to language acquisition. Refined information theoretic methods provide viable computational procedures for acquiring linguistic knowledge.

However, it is important to recognize that the success of unsupervised ML grammar induction does not, in itself, entail that these procedures actually apply in human language acquistion. To understand human language learning and interpretation it is necessary to achieve deeper insight into the psychological and neurological facts of the acquisition process.

6 Conclusions

We have seen that advances in the development of information theoretic learning methods have given rise to substantial progress in the computational modelling of a wide range of NLP tasks. This work is commonly regarded as an achievement in natural language engineering. In fact, it also has implications for theories of human language acquisition and processing. Recent work in unsupervised grammar acquisition is particularly significant in indicating the possibility of accounting for language learning through general procedures of induction and probabilistic learning.

Much remains to be done in order to further clarify the issues that have been raised here. Advocates of the general inductive learning approach must focus on the refinement and further development of unsupervised learning procedures for NLP tasks. Proponents of the language-specific learning device view should be concerned to implement precise and convincing versions of a P&P model for particular grammar acquisition tasks. To be credible these systems must achieve the same level of robustness over large corpora that ML acquisition and processing systems have succeeded in demonstrating. In order to determine the cognitive reality of their respective models adherents of both approaches must work more closely with psychologists and neuro-scientists to explore the relation of computationally viable models of language learning and processing to human performance.

7 Acknowledgements

Earlier versions of this paper were presented at the University College London Workshop on Computational Linguistics, Novermber 2004; Computational Linguistics in the Netherlands, Leiden, December, 2004; the Linguistics Colloquium of the University of Illinois at Urbana-Champaign, March, 2005; and the Construction of Meaning Workshop, Stanford, April, 2005. I am grateful to the participants of these forums for useful comments and criticism. I would also like to thank Eve Clark, Jennifer Cole, Crit Cremers, Jonathan Ginzburg, Dan Jurafsky, Chris Manning, John Nerbonne, Leif Nielsen, Dan Roth, Ivan Sag, and James Yoon for helpful discussion of some of the ideas presented here. I am particularly grateful to Stuart Shieber and Richard Sproat for conversations that were invaluable to me

in clarifying my thinking on many of the issues addressed in this paper. Of course I bear sole responsibility for its content and any errors it may contain.

References

Bos, J., Clark, S., Curran, J., Hockenmaier, J. and Steedman, M.(2004), Wide-coverage semantic representations from a CCG parser, *Proceedings of COLING 18*, Geneva, Switzerland.

Brill, E.(1992), A simple rule-based part of speech tagger, *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, pp. 152 –155.

Brill, E.(1994), Some advances in transformation-based part of speech tagging, *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 722–727.

Charniak, E.(1997), Statistical parsing with context-free grammar and word statistics, *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pp. 598–603.

Chomsky, N.(1957), *Syntactic Structures*, Mouton, The Hague.

Chomsky, N.(1965), *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

Chomsky, N.(1981), *Lectures on Government and Binding*, Foris, Dordrecht.

Chomsky, N.(1986), *Knowledge of Language: Its Nature, Origin, and Use*, Praeger, New York.

Chomsky, N.(1995), *The Minimalist Program*, MIT Press, Cambridge, MA.

Chomsky, N.(2000), *New Horizons in the Study of Language and Mind*, Cambridge University Press, Cambridge.

Chouinard, M. and Clark, E.(2003), Adult reformulations of child errors as negative evidence, *Journal of Child Language* **30**, 637–669.

Church, K.(1988), A stochastic parts program and noun phrase parser for unrestricted text, *Second Conference on Applied Natural Language Processing*, Austin, TX, pp. 136–143.

Church, K.(1992), Current practice in part of speech tagging and suggestions for the future, *in* Simmons (ed.), *Sbornik Praci: In Honor of Henry Kucera*, Michigan Slavic Studies, Michigan, pp. 13–48.

Clark, A.(2000), Inducing syntactic categories by context distribution clustering, *Proceedings of CoNLL 2000*, Lisbon, Portugal.

Clark, A.(2001), Unsupervised induction of stochastic context-free grammars using distributional clustering, *Proceedings of CoNLL 2001*, Toulouse, France.

Clark, S. and Curran, J.(2004), Parsing the WSJ using CCG and log-linear models, *Proceedings of the Forty-Second Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp. 104–111.

Cohen, W. and Singer, Y.(1999), A simple, fast, and effective rule learner, *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*.

Collins, M.(1998), *Head-Driven Statistical Models for Natural Language Parsing*, PhD thesis, University of Pennsylvania.

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A.(2003), TiMBL: Tilburg Memory Based Learner, v. 5.0, Reference Guide, *Technical Report ILK-0310*, University of Tilburg.

Fernández, R., Ginzburg, J. and Lappin, S.(2005), Using machine learning for non-sentential utterance classification, *Proceedings of the Sixth SIGdial Workshop on Discourse and Dialogue*, Lisbon, pp. 77–86.

Fong, S.(1991), *Computational Properties of Principled-Based Grammatical Theories*, PhD thesis, Massachuetts Institute of Technology.

Fong, S.(2005), Computation with probes and goals: A parsing perspective, *in* A. D. Sciullo and R. Delmonte (eds), *UG and External Systems*, John Benjamins, Amsterdam.

Goldsmith, J.(2001), Unsupervised learning of the morphology of a natural language, *Computational Linguistics* **27**, 153–198.

Good, I.(1953), The population frequencies of species and the estimation of population parameters, *Biometrika* **40**, 237–264.

Harris, Z.(1951), *Structural Linguistics*, University of Chicago Press, Chicago, IL.

Harris, Z.(1991), *A Theory of Language and Information: A Mathematical Approach*, Clarendon Press, Oxford, New York.

Hockenmaier, J. and Steedman, M.(2002), Generative models for statistical parsing with combinatory categorial grammar, *Proceedings of the Fortieth Meeting of the Association for Computational Linguistics*, Philadelphia, PA, pp. 335–342.

Klein, D. and Manning, C.(2002), A generative constituent-context model for improved grammar induction, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 128–135.

Klein, D. and Manning, C.(2004), Corpus-based induction of syntactic structure: Models of dependency and constituency, *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, Barcelnoa, Spain.

Le, Z.(2003), Maximum Entropy Modeling Toolkit for Python and C++. http://homepages.inf.ed.ac.uk/ s0450736/maxent_toolkit.php.

Liakata, M. and Pulman, S.(2004), Learning theories from text, *Proceedings of COLING 18*, Geneva, Switzerland.

Nielsen, L.(2003), Using machine learning techniques for VPE detection, *Proceedings of Recent Advances in Natural Lanuage Processing*, Borovets, Bulgaria, pp. 339–346.

Nielsen, L.(2004), Verb phrase ellipsis detection using automatically parsed text, *Proceedings of COLING 18*, Geneva, Switzerland.

Pereira, F.(2000), Formal grammar and information theory: Together again?, *Philosophical Transactions of the Royal Society*, Royal Society, London, pp. 1239–1253.

Pullum, G. and Scholz, B.(2002), Empirical assessment of stimulus poverty arguments, *The Linguistic Review* **19**, 9–50.

Quinlan, R.(1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, CA.

Schone, P. and Jurafsky, D.(2001), Knowledge-free induction of inflectional morphologies, *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, Pittsburgh, PA.

Schuetze, H.(1995), Distributional part-of-speech tagging, *in* 141-148 (ed.), *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 7)*.