Data-Driven Approach towards Automated Deep Lexical Acquisition

Yi Zhang
yzhang@coli.uni-sb.de
Supervisors: Hans Uszkoreit, Valia Kordoni
Department of Computational Linguistics & Phonetics
Saarland University, Germany

Outline

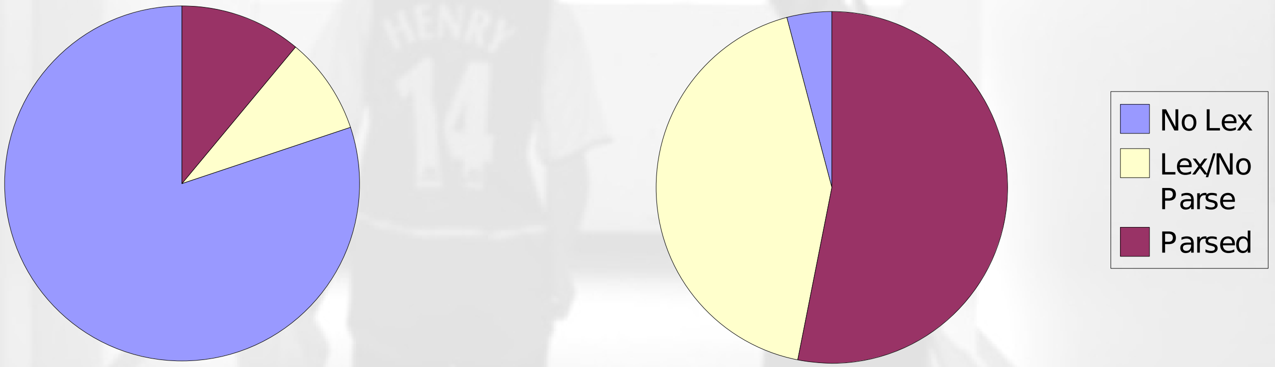
- Why automated DLA
- Previous work
- Data-driven approach
- Experiments
  - LinGO ERG
  - Alpino
- Work in progress/future
- Summary

Why automated DLA

- Broad coverage linguistically deep processing is desirable for advanced NL applications.
- State-of-the-art deep grammars can only achieve moderate coverage:
  - Coverage test of LinGO ERG on BNC shows
    - Full lexical coverage for 32% of strings
    - Of these, parse generated for 57% (83% correct)
    - For parsing failure
      - Missing lexical entries 26%
      - Missing constructions 17%
      - Garbage strings 17%
      - Others

Case Study: Manual Lexical Extension

- Corpus *“Shanghai”*
  - 1600 English sentences/strings about tourism in Shanghai (similar to the *“rondane”* corpus in LOGON).
- Discover new word/MWE; map it to one of the leaf lexical types in ERG



- *1500 entries are merged into official ERG lexicon since Apr-05

Case Study: Manual Lexical Extension

- Amount of work
  - 1575 entries, mostly nouns, adjectives
  - 5 person*days
  - Extension of verbs are much more difficult
    - Extension for 2000 verbs observed in BNC took several months of hard work.
- Conclusion:
  - Lexical extension is crucial for broad coverage text processing
  - Manual extension requires sufficient linguistic sufficiency, and is laborious.

Previous Work in Automated DLA

- Unification-based approach
  - [Erbach(1990)]
    - Parse the sentence with the unknown word
    - Collect the lexical information from the syntactic structure of the parse
    - Create new lexical entry according to the collected lexical information
  - [Barg and Walther(1998)]
    - *Generalizable* and *Revisable* information
  - [Fouvry(2003)]
    - Use external sources to reduce the computational complexity
- Problems
  - Grammar dependent
  - Underspecified lexical entries: overgeneration, comp. complexity

Previous Work in Automated DLA

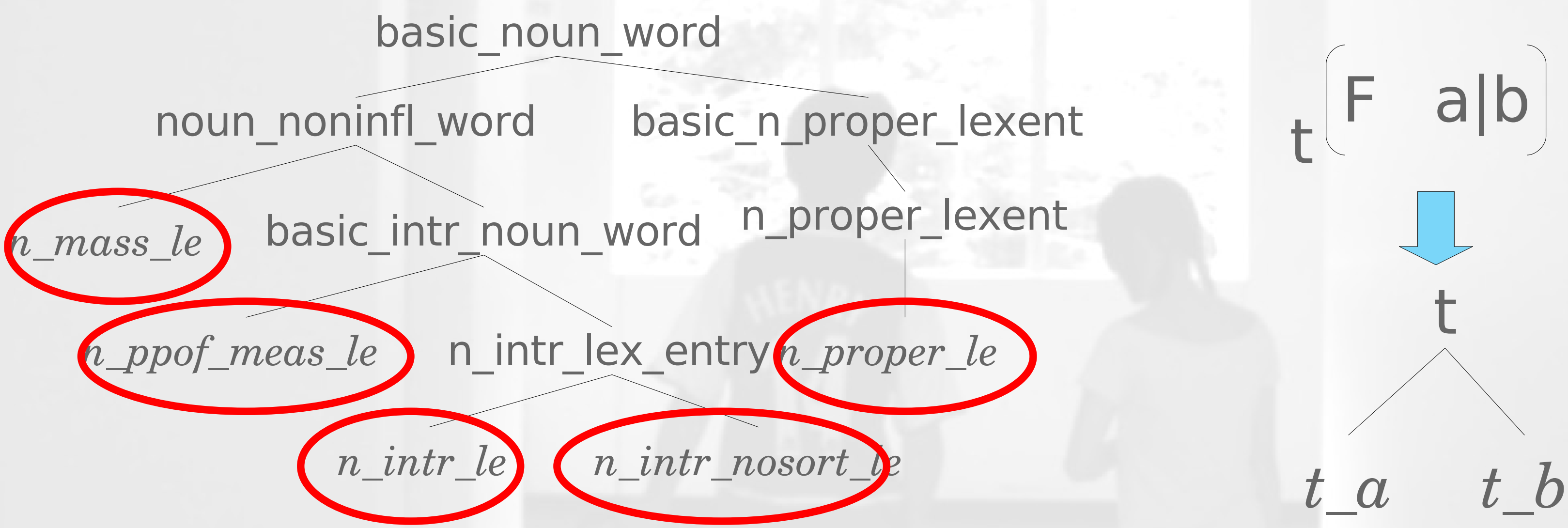
- Corpus-driven approach
  - [Brent(1991)]
    - To learn the SF of verbs from untagged text (Shallow).
  - ... ...
  - [Baldwin(2005)]
    - Bootstrap deep lexicon from secondary language resource, with the help of shallow processing tools
- Problems
  - Most of the approach focuses on some specific aspect of lexicon (SF for verbs, countability for nouns, etc)
  - All relies on the availability of secondary language resource.

Ideas!

Is the grammar itself (plus a set of raw text) capable of predicting unknown words?

DLA as Classification Task

- The lexical entries can be constructed with the lexeme and one of the atomic types.



- DLA assigns an atomic type to each unknown word/lexeme.

Tagger-based Model

- Use general purpose POS tagger
  - *TnT*: HMM-based trigram tagger [Brants(2000)]
  - *MXPOST*: ME-based tagger [Ratnaparkhi(1996)]
- Use atomic lexical types as tag-set
- Train tagger with corpus annotated with lexical types
- Tag the input sequence and use the tagger output for unknowns to create new lexical entries
- *Is general purpose POS tagger capable of handling large tag-set?*

Maximum Entropy based Model

- Maximum Entropy models
  - General feature representation
  - Capable of handling large feature set
  - No independence assumption between features

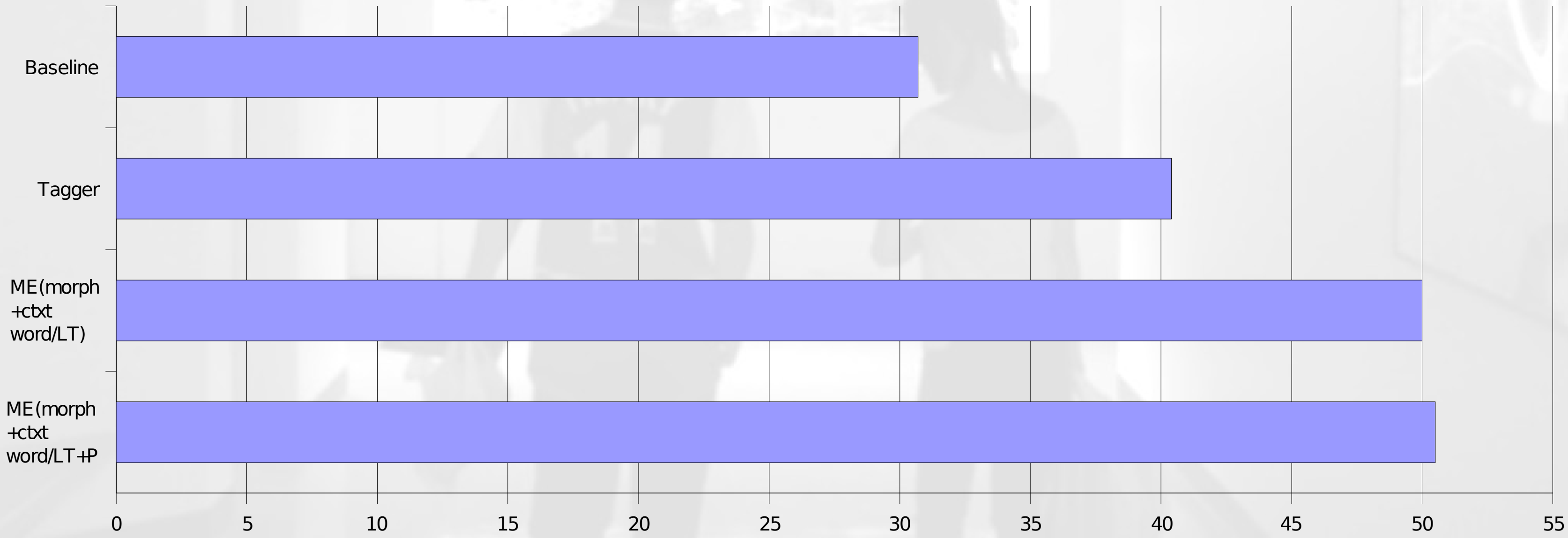
$$p_\Lambda(t|x) = \frac{\exp(\sum_i \lambda_i f_i(x,t))}{\sum_{t' \in T} \exp(\sum_i \lambda_i f_i(x,t'))},\ \Lambda = \{\lambda_i\}.$$

Classification Features

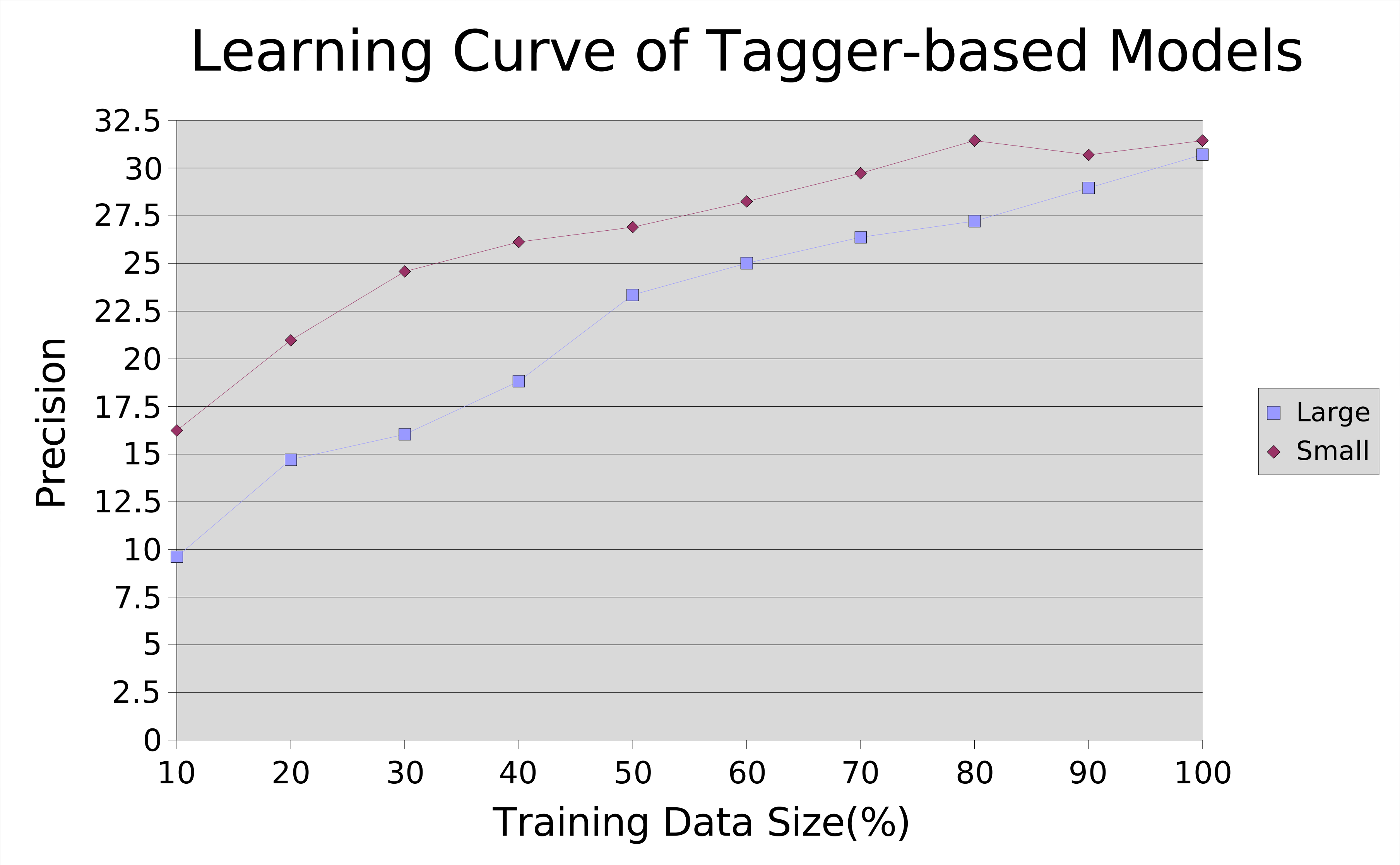
- Morphological features
  - Prefix/Suffix
- Syntactic features
  - Adjacent words/lexical types
  - Partial parse chart/chunks
  - Dependency head/daughters/labels
- Semantic features
  - (R)MRS fragments

Experiment I: LinGO ERG

- More than 700 atomic lexical types
- Redwoods Treebank (5th)
  - 16.5K sentences with 122K tokens
- 10-fold cross validation

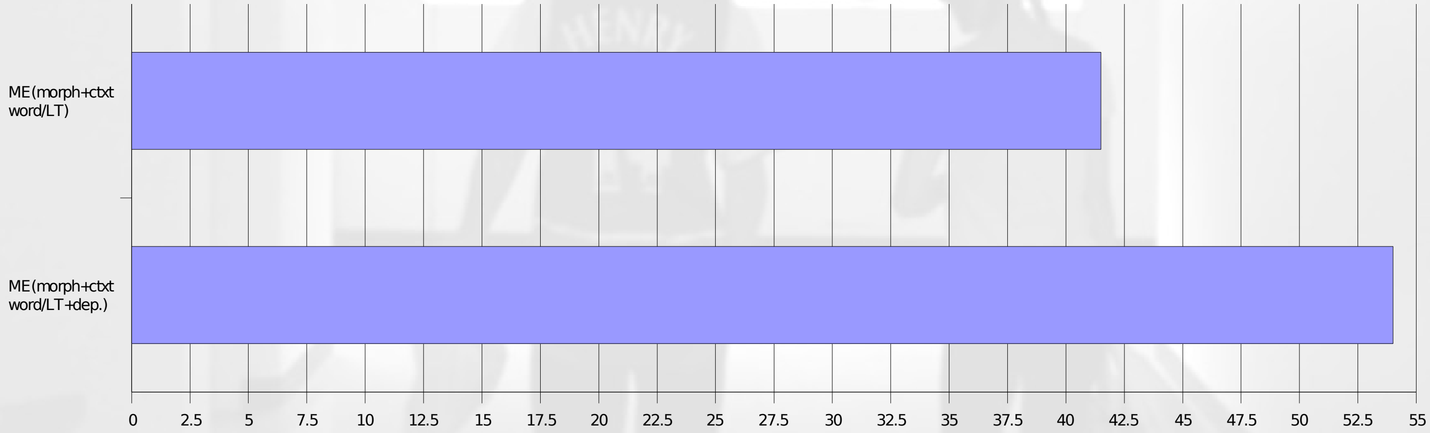


Effect of Large Tag-set



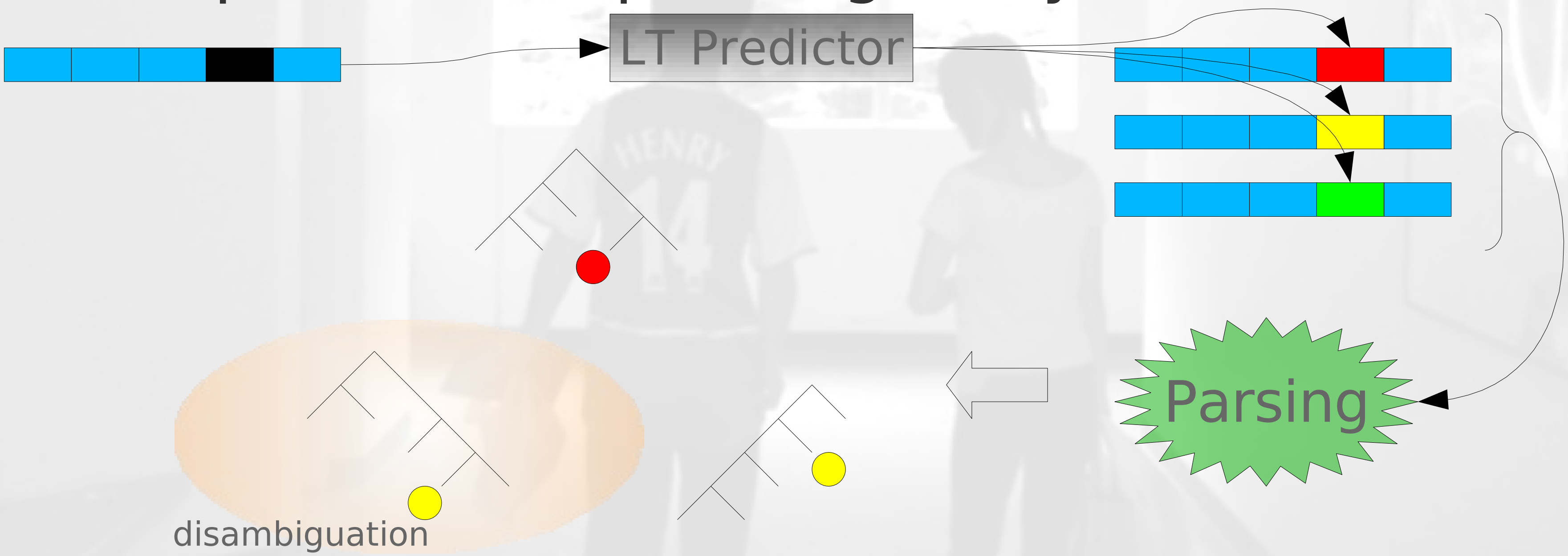
Experiment II: Alpino

- Broad coverage Dutch HPSG grammar
- Large dependency treebank
- Predict ~500 possible SF combinations
- +/- dependency features



Feedback from Full Parsing

- Predictor outputs *n* types
- Full parsing with these new entries
- Select best parse (disambiguation task)
- Keep the corresponding entry



Enhancing Performance with Voting

- Current approach: unknowns are predicted per occurrence
- Most words have no more than five entries
- For the same unknown word in multiple sentences, vote for the best lexical type.

Importing Lexicon

- WordNet 2.0
  - 152,059 words, 203,145 word-sense pairs
- LinGO ERG Apr-05
  - 21,000 entries
- Assumption: Semantically similar (open class) words generally also show syntactic similarity. (vice versa)
- Classifying WordNet word, using sharing lexicon with ERG as training data.

Automated Grammar Extension

- Lexical coverage only counts for part of the *robustness* problem
- Missing construction is another obstacle
- Automated grammar adaption for specific domain

A Larger Theme

- Restricted domain question answering with deep processing
  - More complicated questions
  - Less information redundancy for data intensive approach
  - Domain knowledge available

Summary

- Necessity for automated DLA explained (with manual extension case study)
- Previous works (unification based approach)
- Data-driven models for unknown word prediction
- Experiments with ERG and Alpino
- Work in progress
  - Using Feedback from Full Parsing
  - Improve accuracy with voting
  - Importing lexicon from WordNet
  - Grammar extension
  - Restricted domain question answering with deep processing

Reference

- [Baldwin(2005)] Timothy Baldwin. Bootstrapping deep lexical resources: Resources for courses. In ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition, 2005.
- [Baldwin and Bond(2003)] Timothy Baldwin and Francis Bond. Learning the countability of english nouns from corpus data. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 463{470, Sapporo, Japan, 2003.
- [Baldwin et al.(2004)] Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. Road-testing the English Resource Grammar over the British National Corpus. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, 2004.
- [Barg and Walther(1998)] Petra Barg and Markus Walther. Processing unknown words in HPSG. In Proceedings of the 36th Conference of the ACL and the 17th International Conference on Computational Linguistics, Montreal, Quebec, Canada, 1998.
- [Brants(2000)] Thorsten Brants. TnT - a statistical part-of-speech tagger. In Proceedings of the Sixth Confrence on Applied Natural Language Processing ANLP-2000, Seattle, WA, USA, 2000.
- [Brent(1991)] Michael R. Brent. Automatic acquisition of subcategorization frames from untagged text. In Meeting of the Association for Computational Linguistics, pages 209-214, 1991.
- [Carpenter(1992)] Bob Carpenter. The Logic of Typed Feature Structures. Cambridge University Press, Cambridge, England, 1992.

Reference

- [Darroch and Ratcli(1972)] J. N. Darroch and D. Ratcli. Generalized iterative scaling for log-linear models. In Annals of Mathematical Statistics, volume 43, pages 1470{1480. Institute of Mathematical Statistics, 1972.
- [Erbach(1990)] Gregor Erbach. Syntactic processing of unknown words. IWBS Report 131, IBM, Stuttgart, 1990.
- [Fouvry(2003)] Frederik Fouvry. Lexicon acquisition with a large-coverage unication-based grammar. In Companion to the 10th of EACL, pages 87{90, ACL, Budapest, Hungary, 2003.
- [Oostdijk(2000)] Nelleke Oostdijk. The spoken dutch corpus: Overview and rst evaluation. In Proceedings of Second International Conference on Language Resource and Evaluation (LREC), pages 887{894, 2000.
- [Pollard and Sag(1994)] Carl J. Pollard and Ivan A. Sag. Head-Driven Phrase Structure Grammar. University of Chicago Press, Chicago, Illinois, 1994.
- [Ratnaparkhi(1996)] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In Eric Brill and Kenneth Church, editors,
- Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 133{142, Somerset, New Jersey, 1996.
- [Uszkoreit(2002)] Hans Uszkoreit. New chances for deep linguistic processing. In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, 2002.