OSU Department of Linguistics
Spring 2002
W. Detmar Meurers

LING 795K (11390-7)
Mon. + Wed. 11^{00}–12^{48}
200 Pomerene Hall

Corpora and Linguistic Knowledge
(or: a rationalist perspective on corpora)

The individual sessions

1. *Mon, 1. April*: Organizational meeting
   - Background reading:
     - Historical background: McEnery and Wilson (1996, ch. 1)
     - Overview: Garside et al. (1997, ch 1), McEnery and Wilson (1996, ch 2)
     - Hands-on: Wolters (2002)
2. *Wed, 3. April*: Corpus annotation efforts in Tübingen (Erhard Hinrichs)
   - Preparation for class: Hinrichs et al. (2002)
3. *Mon, 8. April*: On using corpora for theoretical linguistics (Detmar Meurers)
   - Preparation for class: Meurers (2002)
   - Overheads: http://ling.osu.edu/~dm/02/spring/795K/meurers-4up.pdf
4. *Wed, 10. April*: Tokenization (Markus Dickinson)
   - Preparation for class: Grefenstette and Tapanainen (1994)
   - Overheads: http://ling.osu.edu/~dm/02/spring/795K/tokenization-4up.pdf
5. *Mon, 15. April*: POS-tagging overview (Xiaofei Lu)
   - Preparation for class: Leech (1997); Brill (2000)
   - Overheads: http://ling.osu.edu/~dm/02/spring/795K/tagging-overview-4up.pdf
6. *Wed, 17. April*: POS-tagging tagsets (Kyuchul Yoon)
   - Preparation for class: (Cloeren 1999; Manning and Schütze 1999, pp. 139–145)
   - Overheads: http://ling.osu.edu/~dm/02/spring/795K/tagsets-4up.ps
7. *Mon, 22. April*: POS-tagging technology (Anton Rytting)
   - Preparation for class: (Cloeren 1999; Manning and Schütze 1999, pp. 139–145)
   - Overheads: http://ling.osu.edu/~dm/02/spring/795K/rytting_slides_6.pdf
8. *Wed, 24. April*: — "" —
9. *Mon, 29. April*: Morphological analysis (Mike Daniels)
   - Preparation for class: (Minnen et al. 2001; Antworth 1995)

- Overheads: http://ling.osu.edu/~dm/02/spring/795K/morphslides-4.pdf

10. *Wed, 1. May*: — "" —

11. *Mon, 6. May*: Corpus encoding issues: From vertical format to XML (Stacey Bailey)

    - Preparation for class:
      - A Gentle Introduction to SGML. http://www-tei.uic.edu/orgs/tei/sgml/teip3sg/index.html
        Some background XML's parent (of sorts), SGML. The details of defining things in SGML differ some from the XML rules, but this introduction gives some idea of where XML is coming from. (Not to mention that notable examples of corpora - such as the BNC - are marked up using SGML.) Feel free to skim.
      - Flynn, P. The XML FAQ. http://www.ucc.ie/xml/
        A list of frequently asked questions and answers about XML. Useful for general XML answers, but (while relevant) none of the questions are specific to corpora design and processing.
      - Ide, N. (2000). The XML Framework and its Implications for Corpus Access and Use. http://www.cs.vassar.edu/~ide/papers/xml-lrec00.ps
        A short (4-page) overview of specific XML features that are useful to corpus encoding.
    - Overheads: http://ling.osu.edu/~dm/02/spring/795K/xml-4up.pdf

12. *Wed, 8. May*: Sentence segmentation and shallow parsing (Wesley Davidson)

    - Preparation for class: (Palmer and Hearst 1997)

13. *Mon, 13. May*: — "" —

14. *Wed, 15. May*: Treebanks overview and the Penn treebank (Jason Casden)

    - Preparation for class: (Leech and Eyes 1997; Marcus et al. 1993, 1994)
    - Overheads: http://ling.osu.edu/~dm/02/spring/795K/casden-treebank-4up.pdf

15. *Mon, 20. May*: Non-English treebanks: Chinese (Lei Xu) and brief comments on German (Detmar)

    - Preparation for class: (Skut et al. 1997)
    - Overheads: http://ling.osu.edu/~dm/02/spring/795K/xu-4up.pdf

16. *Wed, 22. May*: Querying a syntactically annotated corpus (Nathan Vaillette)

    - Preparation for class: (Pito 1994; Rohde 2001; Corley et al. 2001; McKelvie 2001; König and Lezius 2000)
    - Overheads: http://ling.osu.edu/~dm/02/spring/795K/tq-slides-4up.pdf

    Mon, 27. May: *Memorial day*

17. *Wed, 29. May*: Annotation error detection and correction, part 1: POS tagging (Anna Feldman)

- Preparation for class: (van Halteren 2000; Oliva 2001; Hirakawa et al. 2000; Eskin 2000; Müller and Ule 2001)
- Overheads: http://ling.osu.edu/~dm/02/spring/795K/may29-4up.pdf

18. *Mon, 3. June*: Annotation error detection and correction, part 2: Structural Annotation (Anna Feldman)
    - Preparation for class: (Brants and Skut 1998; Hinrichs et al. 2002)
    - Overheads: http://ling.osu.edu/~dm/02/spring/795K/june3-4up.pdf
19. *Wed, 5. June*: Meta-discussion on the use of corpora for theoretical linguistics
    - Preparation for class: (Borsley 2002; Borsley and Ingham 2002; Bayer et al. 1998; Meurers 2002)

Topics covered

1. Tokenization (Manning and Schütze 1999, pp. 124–131; Grefenstette and Tapanainen 1994; Grefenstette 1999; Sproat et al. 1996; Palmer 1997, 2000)
2. Word level annotation:
    - part of speech (Leech 1997; Brill 2000)
        - taggers (Manning and Schütze 1999, ch. 9 & 10; Brants 2000b)
        - tagsets
            * Which and how many tags? (Elworthy 1995; Brants 1995; Kübler and Wagner 2000; Dienes and Oravecz 2000; Tufiş et al. 2000; Teufel et al. 1996)
            * Tagsets for different languages: e.g., German (Thielen and Schiller 1996), Hungarian (Váradi and Oravecz 1999), Slovene (Džeroski et al. 2000)
            * tagset comparison: (Teufel 1995; Déjean 2000; Atwell et al. 1994, 2000a,b)
    - morphological analysis: general (Trost not dated; Sproat 2000), PC-Kimmo (Antworth 1995), Morphix (German) (Finkler and Neumann 1986, 1989; Neumann et al. 1997; Neumann and Mazzini 1999), Morphy (German) (Lezius et al. 1998), English (Minnen et al. 2001)
3. Background on corpus encoding issues: XML
    - (Brew 2000)
    - http://www.cogsci.ed.ac.uk/~jeanc/corpus-linguistics/
    - http://www.ltg.ed.ac.uk/xml2001/materials.html
    - Collections of interesting resources:
        - http://www.sil.org/computing/routledge/simons/text.html
        - http://xml.coverpages.org/ni2002-02-19-a.html
    - Encoding schemes:
        - TEI http://etext.virginia.edu/TEI.html

- CES http://www.cs.vassar.edu/CES/
- TUSNELDA http://www.sfb441.uni-tuebingen.de/c1/tusnelda-guidelines.html
- A Gentle Introduction to SGML. http://www-tei.uic.edu/orgs/tei/sgml/teip3sg/index.html
  Some background XML's parent (of sorts), SGML. The details of defining things in SGML differ some from the XML rules, but this introduction gives some idea of where XML is coming from. (Not to mention that notable examples of corpora - such as the BNC - are marked up using SGML.) Feel free to skim.
- Flynn, P. The XML FAQ. http://www.ucc.ie/xml/
  A list of frequently asked questions and answers about XML. Useful for general XML answers, but (while relevant) none of the questions are specific to corpora design and processing.
- Ide, N. (2000). The XML Framework and its Implications for Corpus Access and Use. http://www.cs.vassar.edu/~ide/papers/xml-lrec00.ps
  A short (4-page) overview of specific XML features that are useful to corpus encoding.
- The Cover Pages at http://www.oasis-open.org/cover/sgml-xml.html
  A reference site for all things XML.
- Simons, G. The Nature of Linguistic Data and the Requirements of a Computing Environment for Linguistic Research. http://www.sil.org/computing/routledge/simons/text.html
- The XML Specification itself from W3C at http://www.w3.org/TR/REC-xml

4. Syntactic annotation:

   - shallow annotation
     - sentence boundary detection / clause identification:
       - (Palmer 1994; Palmer and Hearst 1997)
       - CONLL-2001 Shared Task (Tjong Kim Sang and Déjean 2001). Good web page with general information and on-line papers from workshop and beyond: http://lcg-www.uia.ac.be/conll2001/clauses/
     - chunking:
       - CoNLL-2000 Shared Task (Tjong Kim Sang and Buchholz 2000). Good web page with general information and on-line papers from workshop and beyond: http://lcg-www.uia.ac.be/conll2000/chunking/
       - (Abney 1991, 1996, 1997; Skut and Brants 1998; Hinrichs et al. 2000c; Müller and Ule 2002)
     - topological fields: (Braun 1999; Müller and Ule 2001, 2002; Müller 2002)
   - full syntactic annotation: treebanks
     - overview: (Leech and Eyes 1997; Carstensen et al. 2001)
     - different treebanks
       - Penn Treebank[1], (Marcus et al. 1993, 1994)

[1] http://www.cis.upenn.edu/~treebank/home.html

        * German (treatment of freer word order, good tools): NEGRA[2] (Skut et al. 1997, 1998; Brants et al. 1999) and the annotate tool (http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html), TIGER[3], (Dipper et al. 2001), Verbmobil (Stegmann et al. 2000; Hinrichs et al. 2000b,a)
        * Chinese Treebanks (tokenization issues, etc.): The Penn Chinese Treebank Project (http://morph.ldc.upenn.edu/ctb/)
        * Prague Dependency Treebank (Hajičová et al. 1998; Böhmová et al. 1999; Böhmová et al. to appear)
        * French (Abeillé and Clément 1999), Turkish (Oflazer et al. 1999)
    - Querying a (structually annotated) corpus: (Pito 1994; Rohde 2001; Keller et al. 1999; Corley et al. 2001; McKelvie 2001; König and Lezius 2000; Kallmeyer 2000; Steiner 2001; Brew 1999)

5. Annotation error detection and correction

    - POS tags:
        - (van Halteren 2000)
        - detection using manually written rules and automatic/manual correction: (Oliva 2001; Oliva and Petkevič 2001)
        - automatic correction after detection using
            * "anomaly detection" of outliers: (Eskin 2000)
            * deep parser: (Hirakawa et al. 2000)
            * shallow parser/topol.fields: (Müller and Ule 2001)
    - structural annotation: (Brants and Skut 1998; Brants 2000a; Hinrichs et al. 2002)

6. Use of corpora

    - discussion on use in theoretical lingusitics: (Meurers 2002; Borsley 2002; Borsley and Ingham 2002; Bayer et al. 1998; Hoard 1998)
    - some theoretical linguistic papers using corpus input: (Zamparelli 1998; Ehrich 2001)
    - quantitative issues: Kurz (2000)
    - acquiring lexical information (Lapata 1999; Lapata et al. 1999)
    - parser evaluation: (Carroll et al. 1999)

General and organizational things

- Course email list: 795k@ling.osu.edu
- Course web page: http://ling.osu.edu/~dm/2002/spring/795K/
- Corpora and corpus annotation tools page: http://ling.osu.edu/~dickinso/corpus.html

[2] http://www.coli.uni-sb.de/sfb378/negra-corpus/
[3] http://www.ims.uni-stuttgart.de/projekte/TIGER/

- Office hours: Wednesdays, 14^{15}–15^{15} in 201a Oxley, or by appointment (dm@ling.osu.edu)

This is a research seminar, i.e., each participant is expected to take an active role as a researcher. More concretely, each participant is expected to

a) explore and present one of the topics:
   - research the topic, starting from provided references
   - present topic to class using overheads which are discussed with me during the office hours the week before
   - prepare and discuss a practical exercise for all participants; for example: exploring the use of a particular tool or corpus, searching particular phenomena, finding tagging errors, ...

b) actively participate in the class discussion, the practical exercises, and the reading assignments

c) prepare a short (10 page) term paper using LaTeX which either
   - motivates and describes a theoretically relevant linguistic pattern, how it can be searched for, and what conclusions can be drawn from the findings, or
   - describes own computational work creating or improving annotations, or
   - deals with another course related topic we can agree on

   The finished paper has to be turned in as a ps or pdf file via email by July 15 (graduating seniors: June 3). There will be no extensions after this date.

References

Abeillé, Anne, Thorsten Brants and Hans Uszkoreit (eds.) (2000). *Proceedings of the Second Workshop on Linguistically Interpreted Corpora (LINC-00)*, Luxembourg.

Abeillé, Anne and Lionel Clément (1999). A tagged reference corpus for French. In Uszkoreit et al. (1999), pp. 13–20. http://wwwtalana.linguist.jussieu.fr/~lionel/papiers/linc99.ps.

Abney, Steven (1991). Parsing by Chunks. In Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing*, Dordrecht: Kluwer Academic Publishers.

Abney, Steven (1996). Partial Parsing via Finite-State Cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*.

Abney, Steven (1997). Partial Parsing via Finite-State Cascades. *Natural Language Engineering* 2(4), 337–344.

Antworth, Evan L. (1995). *User's Guide to PC-KIMMO Version 2*. SIL International, Dallas, TX. http://www.sil.org/pckimmo/v2/doc/guide.html OSU local copy: file:/home/dm/resources/papers/pckimmo-doc/guide.pdf.

Atwell, Eric, George Demetriou, John Hughes, Amanda Schiffrin, Clive Souter and Sean Wilcock (2000a). A comparative evaluation of modern English corpus grammatical annotation schemes. *International Computer Archive of Modern and Medieval English (ICAME). Issue on Computers in English Linguistics* http://www.hit.uib.no/icame/ij24/atwell.pdf OSU local copy: file:/home/dm/resources/papers/atwell-et-al-icamejournal2000.ps.

Atwell, Eric, George Demetriou, John Hughes, Amanda Schiffrin, Clive Souter and Sean Wilcock (2000b). Comparing linguistic interpretation schemes for English corpora. In Abeillé et al. (2000). http://www.comp.leeds.ac.uk/eric/coling2000linc.ps OSU local copy: file:/home/dm/resources/papers/atwell-et-al-coling2000linc.ps.

Atwell, Eric, John Hughes and Clive Souter (1994). AMALGAM: Automatic Mapping Among Lexico-Grammatical Annotation Models. In Judith Klavans and Philip Resnik (eds.), *Proceedings of The Balancing Act - Combining Symbolic and Statistical Approaches to Language, Workshop in conjunction with the 32nd Annual Meeting of the Association for Computational Linguistics*. New Mexico State University, Las Cruces, New Mexico, USA. http://www.scs.leeds.ac.uk/nlp/papers/atwell+hughes+souter94acl.ps.Z OSU local copy: file:/home/dm/resources/papers/atwell-hughes-souter-acl94.ps.

Bayer, Samuel, John Aberdeen, John Burger, Lynette Hirschman, David Palmer and Marc Vilain (1998). Theoretical and computational linguistics: toward a mutual understanding. In Lawler and Dry (1998), pp. 231–255.

Böhmová, Alena, Jan Hajič, Eva Hajičová and Barbora Hladká (to appear). The Prague Dependency Treebank: Three-Level Annotation Scenario. In Anne Abeillé (ed.), *Treebanks: Building and using syntactically annoted corpora*, Dordrecht: Kluwer Academic Publishers.

Böhmová, Alena, Jarmila Panevová and Petr Sgall (1999). Syntactic Tagging Procedure for the Transition from the Analytic to the Tectogrammatical Tree Structure. In *Proceedings of the Second Workshop on Text, Speech, Dialogue*. Mariánské Lázně, Czech Republic, pp. 34–38.

Borsley, Robert D. (2002). Review of S. Hunston and G. Francis, Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English, Amsterdam: John Benjamins, 2000. *Lingua* 112, 235–241. OSU local copy: file:/home/dm/resources/papers/borsley-02.pdf.

Borsley, Robert D. and Richard Ingham (2002). Grow your own linguistics? On some applied linguists' views of the subject. *Lingua* 112, 1–6. OSU local copy: file:/home/dm/resources/papers/borsley-ingham-02.pdf.

Brants, Thorsten (1995). Tagset reduction without information loss. In *Proceedings of the 33nd Annual Meeting of the Association for Computational Linguistics (ACL 95)*. Cambridge, MA: MIT. http://www.coli.uni-sb.de/~thorsten/publications/Brants-ACL95.ps.gz OSU local copy: file:/home/dm/resources/papers/brants-acl95.ps.

Brants, Thorsten (2000a). Inter-Annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. Athens, Greece. http://www.coli.uni-sb.de/~thorsten/publications/Brants-LREC00.ps.gz OSU local copy: file:/home/dm/resources/papers/brants-lrec00.ps.

Brants, Thorsten (2000b). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP 2000)*. Seattle, WA. http://www.coli.uni-sb.de/~thorsten/publications/Brants-ANLP00.pdf.

Brants, Thorsten and Wojciech Skut (1998). Automation of Treebank Annotation. In *Proceedings of New Methods in Language Processing (NeMLaP-98)*. Syndey. http://www.coli.uni-sb.de/~thorsten/publications/Brants-Skut-NeMLaP98.ps.gz

Brants, Thorsten, Wojciech Skut and Hans Uszkoreit (1999). Syntactic Annotation of a German Newspaper Corpus. In *Proceedings of the ATALA Treebank Workshop*. Paris, France, pp. 69–76. http://www.coli.uni-sb.de/~thorsten/publications/Brants-ea-ATALA99.ps.gz

Braun, Christian (1999). Flaches und robustes Parsen deutscher Satzgefüge. Diplomarbeit, Fachbereich Computerlinguistik, Universität des Saarlandes.

Brew, Chris (1999). An extensible visualization tool to aid treebank exploration. In Uszkoreit et al. (1999), pp. 49–55. http://www.ltg.ed.ac.uk/~chrisbr/styling-trees.ps.

Brew, Chris (2000). XML and linguistic markup. Course material OSU Department of Linguistics and ELSNET European Summer School, http://ling.osu.edu/~cbrew/xml.html.

Brill, Eric (2000). Part-of-Speech Tagging. In Dale et al. (2000). http://www.netLibrary.com/ebook_info.asp?product_id=47610.

Carroll, John, Guido Minnen and Ted Briscoe (1999). Corpus Annotation for Parser Evaluation. In Uszkoreit et al. (1999), pp. 35–41.

Carstensen, Kai-Uwe, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde and Hagen Langer (eds.) (2001). *Computerlinguistik und Sprachtechnologie: eine Einführung*. Heidelberg, Berlin: Spektrum, Akademischer Verlag. Additional on-line information: http://www.ifi.unizh.ch/groups/CL/CLBuch/.

Cloeren, Jan (1999). Tagsets. In van Halteren (1999), chap. 4, pp. 37–54.

Corley, Steffan, Martin Corley, Frank Keller, Matthew W. Crocker and Shari Trewin (2001). Finding Syntactic Structure in Unparsed Corpora: The Gsearch Corpus Query System. *Computers and the Humanities* 35(2), 81–94. OSU local copy: file:/home/dm/resources/papers/corley-et-al-01.pdf.

Dale, Robert, Hermann Moisl and Harold Somers (eds.) (2000). *Handbook of Natural Language Processing*. New York: Marcel Dekker. http://www.netLibrary.com/ebook_info.asp?product_id=47610.

Déjean, Hervé (2000). How to Evaluate and Compare Tagsets? A Proposal. In Gavrilidou et al. (2000). http://lcg-www.uia.ac.be/lcg/ps/dejean.lrec2000.ps.gz OSU local copy: file:/home/dm/resources/papers/dejean-00.ps.

Dienes, Péter and Csaba Oravecz (2000). Bottom-up tagset design from maximally reduced tagset. In Abeillé et al. (2000), pp. 42–47. http://www.coli.uni-sb.de/~dienes/dior2000.ps.gz OSU local copy: file:/home/dm/resources/papers/dienes-oravecz-linc00.ps.

Dipper, Stefanie, Thorsten Brants, Wolfgang Lezius, Oliver Plaehn and George Smith (2001). The TIGER treebank. http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/linc2001-abstract-tiger.pdf.

Džeroski, Sašo, Tomaž Erjavec and Jakub Zavrel (2000). Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. In Gavrilidou et al. (2000), pp. 1099–1104. http://nl.ijs.si/et/Bib/LREC00/lrec-tag.ps OSU local copy: file:/home/dm/resources/papers/resources/papers/dzeroski-erjavec-zavel-lreg00.ps.

Ehrich, Veronika (2001). Was *nicht müssen* und *nicht können* (nicht bedeuten können: Zum Skopus der Negation bei den Modalverben des Deutschen. In Reimar Müller and Marga Reis (eds.), *Modalität und Modalverben im Deutschen*, Hamburg: Helmut Buske Verlag, vol. 9 of *Linguistische Berichte, Sonderheft*, pp. 140–176.

Elworthy, David (1995). Tagset Design and Inflected Languages. In *Proceedings of the ACL-SIGDAT Workshop*. Dublin. http://arXiv.org/abs/cmp-lg/9504002.

Eskin, Eleazar (2000). Automatic Corpus Correction with Anomaly Detection. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*. Seattle, Washington. http://www.cs.columbia.edu/~eeskin/papers/treebank-anomaly-naacl00.ps OSU local copy: file:/home/dm/resources/papers/eskin-00.ps.

Finkler, Wolfgang and Günter Neumann (1986). *MORPHIX – Ein hochportabler Lemmatisierungsmodul für das Deutsche*. Memo 8, Universität des Saarlandes, FB Informatik, KI-Labor, Saarbrücken.

Finkler, Wolfgang and Günter Neumann (1989). MORPHIX – A Fast Realization of a Classification-Based Approach to Morphology. In Harald Trost (ed.), *4. Österreichische Artificial-Intelligence-Tagung. Wiener Workshop - Wissensbasierte Sprachverarbeitung. Proceedings*, Springer, no. 176 in Informatik Fachberichte, pp. 11–19.

Garside, Roger, Geoffrey Leech and Tony McEnery (eds.) (1997). *Corpus annotation: linguistic information from computer text corpora*. Harlow, England: Addison Wesley Longman Limited.

Gavrilidou, Maria, George Carayannis, Stella Markantonatou, Stelios Piperidis and Gregory Steinhauer (eds.) (2000). *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-00)*, Athens.

Grefenstette, Gregory (1999). Tokenization. In van Halteren (1999), chap. 9, pp. 117–133.

Grefenstette, Gregory and Pasi Tapanainen (1994). What is a word, what is a sentence? Problems of tokenization. In *Third Conference on Computational Lexicography and Text Research (COMPLEX-94)*. Budapest, Hungary. http://www.xrce.xerox.com/publis/mltt/mltt-004.ps OSU local copy: file:/home/dm/resources/papers/grefenstette-tapanainen-94.ps.

Hajičová, Eva (ed.) (2001). *Proceedings of the Third Wokshop on Linguistically Interpreted Corpora (LINC-01)*, Leuven, Belgium.

Hajičová, Eva, Jarmila Panevova and Petr Sgall (1998). Language resources need annotations to make them reusable: The Prague Dependency Treebank. In *Proceedings of the First Conference on Linguistic Resources*. Granada, Spain, pp. 713–718.

Hinrichs, Erhard, Julia Bartels, Yasuhiro Kawata, Valia Kordoni and Heike Telljohann (2000a). The Tübingen Treebanks for Spoken German, English, and Japanese. In Wahlster (2000), pp. 552–576.

Hinrichs, Erhard, Julia Bartels, Yasuhiro Kawata, Valia Kordoni and Heike Telljohann (2000b). The VerbMobil Treebanks. In Ernst G. Schukat-Talamazzini and Werner Zühlke (eds.), *KONVENS-2000 Sprachkommunikation*. Ilmenau, Germany: VDE-Verlag, pp. 107–112. http://www.coli.uni-sb.de/~kordoni/papers/treebanks.pdf OSU local copy: file:/home/dm/resources/papers/hinrichs-et-al-konvens-00.pdf.

Hinrichs, Erhard, Sandra Kübler, Frank H. Müller and Tylman Ule (2002). A Hybrid Architecture for Robust Parsing of German. In *Third International Conference Language Resources and Evaluation (LREC-02)*. Las Palmas, Canary Islands, Spain. OSU local copy: file:/home/dm/resources/papers/hinrichs-et-al-lrec02.pdf.

Hinrichs, Erhard W., Sandra Kübler, Valia Kordoni and Frank H. Müller (2000c). Robust Chunk Parsing for Spontaneous Speech. In Wahlster (2000), pp. 163–182.

Hirakawa, Hideki, Kenji Ono and Yumiko Yoshimura (2000). Automatic Refinement of a POS Tagger Using a Reliable Parser and Plain Text Corpora. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*. ICCL, Saarbrücken, Germany.

Hoard, James E. (1998). Language understanding and the emerging alignment of linguistics and natural language processing. In Lawler and Dry (1998), pp. 170–230.

Kallmeyer, Laura (2000). A query tool for syntactically annotated corpora. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong, China, pp. 190–198. http://www.sfb441.uni-tuebingen.de/a1/Publikationen/emnlp2000.ps.

Keller, Frank, Martin Corley, Steffan Corley, Matthew W. Crocker and Shari Terwin (1999). Gsearch: A Tool for Sytnactic Investigation of Unparsed Corpora. In Uszkoreit et al. (1999), pp. 56–63.

König, Esther and Wolfgang Lezius (2000). A description language for syntactically annotated corpora. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*. Saarbrücken, Germany, pp. 1056–1060. http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/coling2000.pdf.

Kübler, Sandra and Andreas Wagner (2000). Evaluating POS Tagging under Sub-optimal Conditions. Or: Des Meticulousness Pay? In *Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications (ACIDCA'2000)*. Monastir, Tunisia. http://www.sfs.uni-tuebingen.de/~kuebler/papers/acidca.ps.

Kurz, Daniela (2000). A Statistical Account on Word Order Variation in German. In Abeillé et al. (2000). http://www.coli.uni-sb.de/~kurz/coling00.html OSU local copy: file:/home/dm/resources/papers/kurz-linc00.ps.

Lapata, Maria (1999). Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*. College Park, MA, pp. 397–404. http://www.cogsci.ed.ac.uk/~mlap/acl99_final.ps.gz.

Lapata, Maria, Scott McDonald and Frank Keller (1999). Determinants of Adjective-Noun Plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*. Bergen, pp. 30–36. http://www.coli.uni-sb.de/~keller/papers/eacl99.html.

Lawler, John M. and Helen Aristar Dry (eds.) (1998). *Using Computers in Linguistics: a practical guide*. London and New York, NY: Routledge.

Leech, Geoffrey (1997). Grammatical Tagging. In Garside et al. (1997), chap. 2, pp. 19–33.

Leech, Geoffrey and Elizabeth Eyes (1997). Syntactic annotation: Treebanks. In Garside et al. (1997), pp. 34–52.

Lezius, Wolfgang, Reinhard Rapp and Manfred Wettler (1998). A Freely Available Morphological Analyzer, Disambiguator and Context Sensitive Lemmatizer for German. In *Proceedings of COLING-ACL Conference*. Montreal, Canada.

Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, M., G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz and B. Schasberger (1994). The Penn treebank: Annotating predicate argument structure. ftp://ftp.cis.upenn.edu/pub/treebank/doc/arpa94.ps.gz.

Marcus, M., Beatrice Santorini and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330. ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz.

McEnery, Tony and Andrew Wilson (1996). *Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh, UK: Edinburgh University Press.

McKelvie, David (2001). XMLQUERY 1.5 manual. Web page. http://www.cogsci.ed.ac.uk/~dmck/xmlstuff/xmlquery/index.html.

Meurers, Walt Detmar (2002). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua* to appear, draft at http://ling.osu.edu/~dm/papers/meurers-02.html.

Minnen, Guido, John Carroll and David Pearce (2001). Applied morphological processing of English. *Natural Language Engineering* 7(3), 207–223. http://www.cogs.susx.ac.uk/lab/nlp/carroll/abs/01mcp.html.

Müller, Frank H. (2002). The German Chunk-Style-Book. http://ling.osu.edu/~dm/02/spring/795K/mueller-chunks.ps.

Müller, Frank H. and Tylman Ule (2001). Satzklammer annotieren und Tags korrigieren: Ein mehrstufiges Top-Down-Bottom-Up-System zur flachen, robusten Annotierung von Sätzen im Deutschen. In *Proceedings der GLDV-Frühjahrstagung 2001*. pp. 235–244. http://www.sfb441.uni-tuebingen.de/a1/Publikationen/GLDV2001-mueller.pdf.

Müller, Frank H. and Tylman Ule (2002). Annotating topological fields and chunks – and revising POS tags at the same time. In *Proceedings of COLING*. http://ling.osu.edu/~dm/02/spring/795K/mueller-ule.ps.

Neumann, G., R. Backofen, J. Baur, M. Becker and C. Braun (1997). An Information Extraction Core System for Real World German Text Processing. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*. Washington, D.C. http://www.dfki.de/~neumann/publications/new-ps/smes-anlp97.ps.gz OSU local copy: file:/home/dm/resources/papers/neumann-et-al-smes-anlp97.ps.

Neumann, Günter and Giampaolo Mazzini (1999). *Domain-adaptive Information Extraction*. Technical report, DFKI, Saarbrücken. http://www.dfki.de/~neumann/smes/smes.ps.gz OSU local copy: file:/home/dm/resources/papers/neumann-mazzini-99.ps.

Oflazer, Kemal, Dilek Z. Hakkani-Tür and Gökhan Tür (1999). Design for a Turkish Treebank. In Uszkoreit et al. (1999), pp. 28–34.

Oliva, Karel (2001). The Possibilities of Automatic Detection/Correction of Errors in Tagged Corpora: A Pilot Study on a German Corpus. In Václav Matoušek, Pavel Mautner, Roman Mouček and Karel Taušer (eds.), *Text, Speech and Dialogue. 4th International Conference, TSD 2001, Zelezna Ruda, Czech Republic, September 11-13, 2001, Proceedings*. Springer, vol. 2166 of *Lecture Notes in Computer Science*, pp. 39–46.

Oliva, Karel and Vladimír Petkevič (2001). On the Need of *Linguistic* Linguistic Interpretation of Corpora. In Hajičová (2001). Abstract at http://wwwling.arts.kuleuven.ac.be/sle2001/abstracts/web-emp-oliva.htm.

Palmer, D. and M. Hearst (1997). Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics* 23(2), 241–269. http://crow.ee.washington.edu/people/palmer/papers/cl97.ps OSU local copy: file:/home/dm/resources/papers/palmer-hearst-cl97.ps.

Palmer, David D. (1994). SATZ – An Adaptive Sentence Segmentation System. M.s. thesis, University of California, Berkeley. http://crow.ee.washington.edu/people/palmer/papers/ms-thesis.ps OSU local copy: file:/home/dm/resources/papers/palmer-94.ps.

Palmer, David D. (1997). A Trainable Rule-Based Algorithm for Word Segmentation. In Philip R. Cohen and Wolfgang Wahlster (eds.), *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Somerset, New Jersey: Association for Computational Linguistics, pp. 321–328. http://crow.ee.washington.edu/people/palmer/papers/acl97.ps OSU local copy: file:/home/dm/resources/papers/palmer-acl97.ps.

Palmer, David D. (2000). Tokenisation and Sentence Segmentation. In Dale et al. (2000), pp. 11–35. http://www.netLibrary.com/ebook_info.asp?product_id=47610.

Pito, Richard (1994). TGREPDOC. Manual page for tgrep. http://mccawley.cogsci.uiuc.edu/corpora/tgrep.pdf.

Rohde, Doug (2001). Tgrep2. The next-generation search engine for parse trees. Version 1.02. Web page. http://www-2.cs.cmu.edu/~dr/Tgrep2/.

Skut, Wojciech and Thorsten Brants (1998). Chunk Tagger – Statistical Recognition of Noun Phrases. In *Proceedings of the ESSLLI Workshop on Automated Acquisition of Syntax and Parsing*. Saarbrücken, Germany. http://www.coli.uni-sb.de/~thorsten/publications/Skut-Brants-ESSLLI-Parsing98.ps.gz

Skut, Wojciech, Thorsten Brants, Brigitte Krenn and Hans Uszkoreit (1998). A Linguistically Interpreted Corpus of German Newspaper Text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*. Saarbrücken, Germany. http://www.coli.uni-sb.de/~thorsten/publications/Skut-ea-ESSLLI-Corpus98.ps.gz

Skut, Wojciech, Brigitte Krenn, Thorsten Brants and Hans Uszkoreit (1997). An Annotation Scheme for Free Word Order Languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*. Washington, D.C. http://www.coli.uni-sb.de/~thorsten/publications/Skut-ea-ANLP97.ps.gz OSU local copy: file:/home/dm/resources/papers/skut-ea-anlp97.ps.

Sproat, Richard (2000). Lexical Analysis. In Dale et al. (2000), pp. 37–57. http://www.netLibrary.com/ebook_info.asp?product_id=47610.

Sproat, Richard, Chilin Shih, William Gale and Nancy Chang (1996). A stochastic finite-state wordsegmentation algorithm for Chinese. *Computational Linguistics* 22(3), 377–404.

Stegmann, Rosmary, Heike Telljohann and Erhard W. Hinrichs (2000). *Stylebook for the German Treebank in VERBMOBIL*. Verbmobil-Report 239, Universität Tübingen, Tübingen, Germany. http://verbmobil.dfki.de/cgi-bin/verbmobil/htbin/decode.cgi/share/VM-depot/FTP-SERVER/vm-reports/report-239-00.ps.

Steiner, Ilona (2001). VIQTORIA (A Visual Query Tool for Syntactically Annotated Corpora). Talk at the Conference on Linguistic Data Structures. University of Tübingen. 22.-24. February 2001.

Teufel, Simone (1995). A Support Tool for Tagset Mapping. In *Proceedings of the SIGDAT Workshop at EACL 95*. Dublin. http://www.cogsci.ed.ac.uk/~simone/eacl95.ps.gz OSU local copy: file:/home/dm/resources/papers/teufel-95.

Teufel, Simone, Helmut Schmid, Hulrich Heid and Anne Schiller (1996). *EAGLES Study of the relation between Tagsets and Taggers*. Eagles document eag clwg tags/v. ftp://ftp.ilc.pi.cnr.it/pub/eagles/lexicons/tags.ps.gz OSU local copy: file:/home/dm/resources/papers/eagles-tags-96.ps.

Thielen, Christine and Anne Schiller (1996). Ein kleines und erweitertes Tagset fürs Deutsche. In Helmut Feldweg and Erhard W. Hinrichs (eds.), *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, Tübingen: Max Niemeyer Verlag, vol. 73 of *Lexicographica: Series maior*, pp. 215–226.

Tjong Kim Sang, Erik F. and Sabine Buchholz (2000). Introduction to the CoNLL-2000 Shared Task: Chunking. In *Proceedings of the Fourth Conference on Computational Language Learning (CoNLL-2000) and the Second Learning Language in Logic Workshop (LLL-2000)*. Lisbon, Portugal, pp. 127–132. http://lcg-www.uia.ac.be/conll2000/ps/12732tjo.ps OSU local copy: file:/home/dm/resources/papers/tjong-kim-sang-buchholz-conll00.ps.

Tjong Kim Sang, Erik F. and Hervé Déjean (2001). Introduction to the CoNLL-2001 Shared Task: Clause Identification. In Walter Daelemans and Rémi Zajac (eds.), *Proceedings of CoNLL-2001*. Toulouse, France, pp. 53–57. http://lcg-www.uia.ac.be/conll2001/ps/05357tjo.ps OSU local copy: file:/home/dm/resources/papers/tjong-kim-sang-dejean-conll01.ps.

Trost, Harald (not dated). Computational Morphology. On-line document http://www.ai.univie.ac.at/~harald/handbook.html.

Tufiş, Dan, Peter Dienes, Csaba Oravecz and Tamás Váradi (2000). Principled Hidden Tagset Design for Tiered Tagging of Hungarian. In Gavrilidou et al. (2000). http://www.coli.uni-sb.de/~thorsten/tnt/papers/lrec2000-tufis-ea.pdf OSU local copy: file:/home/dm/resources/papers/lrec2000-tufis-ea.pdf.

Uszkoreit, Hans, Thorsten Brants and Brigitte Krenn (eds.) (1999). *Proceedings of the Workshop on Linguistically Interpreted Corpora (LINC-99)*, Bergen, Norway. Association for Computational Linguistics.

van Halteren, Hans (ed.) (1999). *Syntactic Wordclass Tagging*. Dordrecht: Kluwer Academic Publishers.

van Halteren, Hans (2000). The Detection of Inconsistency in Manually Tagged Text. In Abeillé et al. (2000). OSU local copy: file:/home/dm/resources/papers/vanhalteren-linc00.ps.

Váradi, Tamás and Csaba Oravecz (1999). Morpho-syntactic ambiguity and tagset design for Hungarian. In Uszkoreit et al. (1999), pp. 8–12. http://www.inf.u-szeged.hu/~alexin/ILP/EACL99-Bergen.ps.gz OSU local copy: file:/home/dm/resources/papers/varadi-oravecz-99.ps.

Wahlster, Wolfgang (ed.) (2000). *Verbmobil: Foundations of Speech-to-Speech Translation*. Artificial Intelligence. Berlin: Springer.

Wolters, Maria (2002). Working with Corpora. Lecture Notes for ESSLLI 2002, Trento. http://groups.yahoo.com/group/workingwithcorpora/files/main.pdf OSU local copy: file:/home/dm/resources/papers/wolters-esslli02.pdf.

Zamparelli, Roberto (1998). A theory of kinds, partitives and of/z possessives. In Chris Wilder and Artemis Alexiadou (eds.), *Possessors, Predicates and Movement in the Determiner Phrase*, Amsterdam: John Benjamins Publishing Co. http://www.unibg.it/dsfc/pers/zamparelli/ling/parkind.A4.pdf.