EXPLOITING LINGUISTIC DATA IN MACHINE TRANSLATION

Ondřej Bojar

ÚFAL
ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

STUDIES IN COMPUTATIONAL AND THEORETICAL LINGUISTICS

Ondřej Bojar

EXPLOITING LINGUISTIC DATA IN MACHINE TRANSLATION

Published by Institute of Formal and Applied Linguistics
as the 7th publication in the series
Studies in Computational and Theoretical Linguistics.

Reviewers: Chris Callison-Burch, Johns Hopkins University
Jan Cuřín, IBM Czech Republic

This book has been printed with the support of the project MSM0021620838 of The Ministry of Education of the Czech Republic.

ISBN 978-80-904175-8-8

Contents

Acknowledgement

I would like to express my gratitude to the Institute of Formal and Applied Linguistics (ÚFAL) for excellent support and to my colleagues for all the stimulating discussions. It is impossible to name everyone who influenced my research, so here is an abbreviated list: the head of our department Jan Hajič, my thesis supervisor Vláďa Kuboň, and all the numerous friends and colleagues at ÚFAL; those I met at informatics in Hamburg (Wolfgang Menzel, Michael Daum and others), at the Programming Systems Lab and CoLi in Saarbrücken (Ralph Debusmann, Marco Kuhlmann, Ivana and Geert-Jan Kruijff, Valia Kordoni and many others); the very influential team of Hermann Ney at RWTH Aachen University (Richard Zens, Saša Hasan and many others again); Philipp Koehn's MT team at the Johns Hopkins University summer workshop, MT Marathons and in Edinburgh (Chris Dyer, Hieu Hoang, Phil Blunsom and Adam Lopez to name a few) as well as the very warm and inspiring groups in Melbourne: the LT group of Steven Bird, Tim Baldwin and many others; and the Mercury team (Ralph Becket, Julien Fischer and others). Also, I do not wish to forget all the short random friendly encounters with members of our community at summer schools, workshops or conferences.

And last but not least, this book would never have come into being without the support of my greater family, my parents, and my wife Pavla.

The work reported in this book was also partially supported by the following grants: FP6-IST-5-034291-STP (EuroMatrix), MSM0021620838, MŠMT ČR LC536, GA405/06/ /0589, GAČR 201/05/H014, GAAV ČR 1ET201120505, GAUK 351/2005, and Collegium Informaticum GAČR 201/05/H014.

1

Introduction

Computational linguistics and natural language processing (NLP) try to formally capture and model the complexity of how people communicate using a natural language. The field has implications in many aspects of the society: linguistic theories are sometimes used as a basis when prescribing what is an appropriate and correct usage of an expression, they predict how a message is perceived by a human recipient and justify which information should be included in language textbooks, dictionaries or lexicons. Applications are built to speed up human processing of text (such as finding relevant documents, answering questions, translating from one language to another) or attempt to turn the computer into a real partner able to share knowledge and obey commands issued in a natural language.

1.1 Relation between Theory, Applications and Data

Both linguistic theories and NLP applications rely heavily on language data, which include raw examples of language expressions (written sentences in books, newspapers, sentences uttered in a dialog, recorded or broadcasted) as well as more or less formalized data *about* the language itself (such as style guides or dictionaries). On the one hand, examples of language usage can validate linguistic theories (by testing predictions on real data) and on the other hand, linguistic theories provide a framework for creating derived language resources like the above mentioned lexicons and dictionaries. Thus, the theory is tested indirectly, by applying and using a derived resource in a practical task. NLP applications are related to data even more tightly simply because the application has some input and output data. Moreover, many NLP applications need to consult varying amounts of language data in order to be able to achieve their goal.

In this book, we study the mutual relationship between a linguistic theory, an NLP application and language data. We focus on one particular theory, the theory of Functional Generative Description (FGD), one particular type of derived language data, namely valency dictionaries, and on one particular NLP application, namely machine translation (MT). Whenever possible, we try to include references to relevant alternatives.

1.2 How Theory Should Help

The general belief is that having an established theory as a background of an NLP application should bring an advantage to the design of the application: the description of the algorithm could be shorter because it builds on top of notions defined in the theory, decisions that have to be made should be more local and thus easier to meet and finally, such an application should produce outputs of a predictable quality. In short, a good theory should constrain the internal structure of applications to their advantage.

There is a similar relation between the theory and language data: a good theory describes which features of unprocessed language data are significant for a particular task. A theory provides a view on unprocessed data. Given a task and following the theory, we can "compress" raw language data by ignoring all but relevant features. Dictionaries are an excellent example of such compression: instead of scanning large texts and looking at many occurrences of a word to understand the meaning and correct ways of using it in context we just read a short (formal) description.

In an NLP application such as MT, there is always someone who has to do the difficult job. In the extreme case, all the intelligence is contained in a "dictionary", i.e. the "dictionary" provides the expected output of the application for every possible input. More realistically, we can expect to know at least *parts* of the output from the top of our head but we have to correctly glue them together to create a complete answer. The more or the better training data we have, the simpler the application can be.

To sum up, a theory provides guidelines on how to build linguistic applications and how to look at language data. If all goes well, such a theoretical background will simplify the design and facilitate better performance at the same time.

1.3 Structure of the Book

This study consists of two major parts: the first one is devoted to lexical acquisition (Chapter 2) and the second one to machine translation (Chapters 3 and 4), linked as follows:

One of the key components in the theory of our choice, FGD (briefly introduced in Section 2.2), is the valency theory which predicts how an element in a grammatically well formed sentence can or must be accompanied by other elements. The prediction primarily depends on the sense of the governing word and it is best captured in a lexicon. The motivation to build such lexicons comes often from applications: some applications simply require a lexicon to e.g. produce an output text, while some only benefit from them by improving accuracy or increasing coverage. Finally, a syntactic lexicon is always a valuable reference for human users of the language. However, the development of lexicons is costly and therefore we focus on the question of automatic suggestion of entries based on available textual data. In short, Chapter 2 explores the theory of FGD and the journey from raw language data in a text to a compressed formalized representation in a lexicon.

In Chapter 3 we pick an NLP application, the task of machine translation (MT) in particular, to study how the theory lends itself to practical employment. After a brief review of various approaches to MT, we follow up on FGD and describe our system of syntax-based machine translation. The full complexity of the system is outlined, but the main focus is given only to our contribution, syntactic transfer. Nevertheless, we implement the whole pipeline of the MT system and we are able to evaluate MT quality using an established automatic metric.

Chapter 4 is devoted to a contrast experiment: we aim at English-to-Czech MT leaving the framework of FGD aside and using a rather direct method. We briefly summarize the state-of-the-art approach, so-called phrase-based statistical machine translation, including an extension to factored MT where various linguistically motivated aspects can be explicitly captured. Then we demonstrate how to use factors to improve morphological coherence of MT output and compare the performance of the direct approach with the syntax-based system from Chapter 3.

We conclude by Chapter 5, providing a broad survey of documented utility of lexicons in NLP and summarizing our observations and contributions.

3

Machine Translation via Deep Syntax

In the previous chapter we studied methods of automated lexical acquisition. Resulting syntactic lexicons can serve as a resource for various NLP applications. In order to better empirically understand the applicability of lexicons, we now focus on a single practical task, namely machine translation (MT). After a brief review of approaches to MT (Section 3.1), we describe a syntax-based MT system. In theory, this is the approach where deep syntactic lexicons could be later used.

3.1 The Challenge of Machine Translation

Machine translation (MT) is an intriguing task. Researchers have hoped in automated text translation since the era of John von Neumann and Alan Turing (see Hutchins (2005) or the IBM press release in 1954[1]), and the field has seen both spectacular failures[2] as well as surge of activity and success. For a review including a summary of issues that an MT system has to overcome see e.g. Dorr *et al.* (1998).

While fully automatic high-quality MT is still far beyond our reach, restricted settings often allowed to create highly successful applications such as computer tools aiding human translation (e.g. translation memories, see Lagoudaki (2006)), closed--domain fully automatic systems (Chevalier *et al.*, 1978), or tentative machine translation to enable at least a partial access to information in a foreign text (e.g. web services Babelfish[3] or Google Translation[4]).

In essence, the task of MT is to correctly reuse pieces of texts previously translated by humans to translate sentences never seen so far.[5] Some methods follow the line very tightly, not being able to produce any word or expression not seen in some training text, while some methods (most notably all rule-based or dictionary-based ones) operate with a very distilled representation of words and their translations. In the latter setup, training texts as well as a broad world knowledge were processed

---

[1] `http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html`

[2] Failure to meet expectations causing a decline in funding for a decade (ALPAC, 1966; Hutchins, 2003) or failure to produce any working system in the EUROTRA project (Oakley, 1995; Hutchins, 1996). Note however, that there are quite conflicting objectives in MT research and even a failing project can bring a very significant progress in theoretical understanding or language modelling, see Rosen (1996) for a discussion.

[3] `http://babelfish.altavista.com/`

[4] `http://translate.google.com/`

[5] Human translators proceed well beyond this boundary, trying to understand the described situation based on other information sources and e.g. to enrich the translation with all explanation necessary for the reader.

by human experts, so there is no well defined set of training data and no direct link between the data and the system.[6] Further serious empirical questions arise as we start to investigate what the best "piece" of a sentence to reuse might be, as discussed below.

3.1.1 Approaches to Machine Translation

One of the key distinctions between various MT systems is the level of linguistic analysis employed in the system, see the MT triangle by Vauquois (1975) in Figure 3.1. Roughly speaking, an MT system is "direct" or "shallow" if it operates directly with words in source and target languages and it is "deep" if is uses some formal representation (partially) describing the meaning of the sentence. We examine both of the approaches further below.

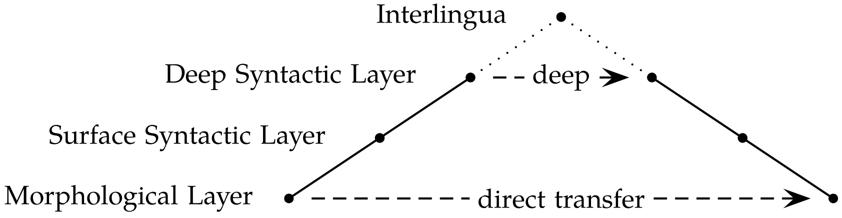

Figure 3.1: Vauquois' triangle of approaches to machine translation.

Another distinction is made between "rule-based" and "statistical" (or "stochastic" or "data-driven") systems. In rule-based systems, all the implementation work is done by human experts, in statistical systems, humans design a probabilistic model describing the process of translation and use large amounts of data to train the model.

To an extent, we do not consider the difference between "rule-based" and "statistical" approaches being too big. In both cases, there has to be someone who does some data abstraction at some point. In hand-crafted rule-based systems, the abstraction happens as human translators learn the two languages and formally describe the rules of translation. In data-driven systems, the abstraction according to the specification of the model happens either at a pre-processing phase (collecting statistics) or on the fly when searching for sentences similar to the one that is to be translated (example-based methods). Moreover, many rule-based systems rely on large linguis-

---

[6] Some researchers argue that human experts may not have used any training parallel texts at all when implementing the transfer rules. Still, while learning the two languages, they have at least discussed real-life situations in the two languages with others, if not read a foreign language textbook.

tic resources such as translation dictionaries anyway and in such cases, automated creation of such resources is highly desirable (see Chapter 2).

Direct (Shallow) MT

Introduced by King (1956) and applied by Brown *et al.* (1988), shallow MT systems treat words in a input sentence as more or less atomic units and attempt a direct conversion of the input sequence of atomic units into the output sequence of atomic units.

For instance, the Czech sentence *Dobré ráno* can be translated to English *Good morning* using a simple word-to-word translation dictionary. The linguistic inadequacy of the direct approach becomes apparent if we consider a similar sentence *Dobrý večer (Good evening)*. A completely uninformed system wastefully needs two new entries in the dictionary (*Dobrý* for *Good* and *večer* for *evening*) because it has no idea that both *Dobré* and *Dobrý* are just two morphological variants of the same word. In order to reverse the translation direction, some additional information has to be provided to make the system correctly choose between *Dobrý* and *Dobré* for *Good*.

In short, direct approaches start with little or no linguistic theory and introduce further extensions to the process of translation only when necessary. As we will see in Chapter 4, such systems can still deliver surprisingly good results, and more so once some (limited) linguistic knowledge is implemented into the design of the system.

Deep Syntactic MT

First machine translation systems as well as prevailing commercial MT systems to date (e.g. SYSTRAN) incorporate principles from various linguistic theories from the very beginning.

For an input sentence represented as a string of words, some symbolic representation is constructed, possibly in several steps. This symbolic representation, with the exception of a hypothetical Interlingua, remains language-dependent, so a transfer step is necessary to adapt the structure to the target language. The translation is concluded by generating target-language string of words from the corresponding symbolic representation.

In the following, we focus on one particular instance of this symbolic representation, namely the framework of FGD (see Section 2.2). We experiment primarily English-to-Czech translation via the t-layer (deep) and compare it to transfer at the a-layer (surface syntax). Previous research within the same framework but limited to rather surface syntax includes the system APAČ (Kirschner and Rosen, 1989).

Other examples of a deep syntactic representation, in essence very similar to FGD, include Mel'čuk (1988), Microsoft logical form (Richardson *et al.*, 2001) or the ideas spread across the projects PropBank (Kingsbury and Palmer, 2002), NomBank (Meyers *et al.*, 2004) and Penn Discourse Treebank (Miltsakaki *et al.*, 2004). MT systems are also being implemented in less dependency-oriented formalisms such as the DELPH-IN initiative (Bond *et al.*, 2005) for HPSG (Pollard and Sag, 1994). See e.g. Oepen *et al.*

(2007) and the cited papers for a recent overview of the LOGON project that combines various formalisms of deep syntactic representation.

3.1.2 Advantages of Deep Syntactic Transfer

The rationale to introduce additional layers of formal language description such as the tectogrammatical (t-) layer in FGD is to bring the source and target languages closer to each other. If the layers are designed appropriately, the transfer step will be easier to implement because (among others):

- t-structures of various languages exhibit less divergences, fewer structural changes will be needed in the transfer step.
- t-nodes correspond to auto-semantic words only, all auxiliary words are identified in the source language and generated in the target language using language-dependent grammatical rules between t- and a- layers.
- t-nodes contain word lemmas, the whole morphological complexity of either of the languages is handled between m- and a- layers.
- the t-layer abstracts away word-order issues. The order of nodes in a t-tree is meant to represent information structure of the sentence (topic-focus articulation). Language-specific means of expressing this information on the surface are again handled between t- and a- layers.

Overall, the design of the t-layer aims at reducing data sparseness so less parallel training data should be sufficient to achieve same coverage.

Moreover, the full definition of the t-layer includes explicit annotation of phenomena like co-reference to resolve difficult but inevitable issues of e.g. pronoun gender selection. As tools for automatic tectogrammatical annotation improve, fine nuances could be tackled.

3.1.3 Motivation for English→Czech

This study focuses on translation from English to Czech. Apart from personal reasons, our choice has two advantages: both languages are well studied and there are available language data for both of the languages.

Table 3.1 summarizes some of the well known properties of Czech language[7]. Czech is an inflective language with rich morphology and relatively free word order. However, there are important word order phenomena restricting the freedom. One of the most prominent examples are clitics, i.e. pronouns and particles that occupy a very specific position within the whole clause. The position of clitics is rather rigid and global within the sentence. Examples of locally rigid structure include (non-recursive) prepositional phrases, coordination and to some extent also the internal

---

[7] Data by Nivre *et al.* (2007), Zeman (http://ufal.mff.cuni.cz/˜zeman/projekty/neproj), Holan (2003), and Bojar (2003). Consult Kruijff (2003) for empirical measurements of word order freeness.

	Czech	English
Morphology	rich $\geq$ 4,000 tags $\geq$ 1,400 actually seen	limited 50 used
Word order	free with rigid global phenomena	rigid
Known dependency parsing results		
Labelled edge accuracy	80.19%	89.61%
Unlabelled edge accuracy	86.28%	90.63%

Table 3.1: Properties of Czech compared to English.

order of noun phrases. Other elements, such as the predicate, subject, objects or other modifiers of the verb may be nearly arbitrarily permuted. Such permutations correspond to the topic-focus articulation of the sentence. Formally, the topic-focus articulation is expressed as the order of nodes at the t-layer.

Moreover, like other languages with relatively free word order, Czech allows non--projective constructions (crossing dependencies). Only about 2% of edges in PDT are non-projective, but this is enough to make nearly a quarter (23.3%) of all the sentences non-projective. While in theory there is no upper bound on the number of gaps (Holan *et al.*, 2000; Kuhlmann and Möhl, 2007) in a Czech sentence (see Figure 3.2), Debusmann and Kuhlmann (2007) observe that 99% of sentences in PDT contain no more than one gap and are well-nested, which makes them parsable by Tree-Adjoining Grammars (TAG, Joshi *et al.* (1975), see also the review by Joshi *et al.* (1990)). Note that other types of texts may exhibit more complex sentence structure.

3.1.4 Brief Summary of Czech-English Data and Tools

Table 3.2 summarizes available Czech monolingual and Czech-English parallel corpora, including the available annotation. We use the tools listed in Table 3.3 to automatically add any further layers of annotation and to generate plaintext from the deep representation.

A new version of Prague Czech-English Dependency Treebank (PCEDT 2.0) is currently under development. PCEDT 2.0 will not only be about twice the size of PCEDT 1.0, but more importantly the annotation at both Czech and English t-layers will be manual. This will allow to collect reliable estimates of structural divergence at the t-layer and train deep-syntactic transfer models on highly accurate data.

3.2 Synchronous Tree Substitution Grammar

Synchronous Tree Substitution Grammars (STSG) were introduced by Hajič *et al.* (2002) and formalized by Eisner (2003) and Čmejrek (2006). They capture the basic

Summary

This study explores the mutual relationship between linguistic theories, data and applications. We focus on one particular theory, Functional Generative Description (FGD), one particular type of linguistic data, namely valency dictionaries and one particular application: machine translation (MT) from English to Czech.

First, we examine methods for automatic extraction of verb valency dictionaries based on corpus data. We propose an automatic metric for estimating how much lexicographers' labour was saved and evaluate various frame extraction techniques using this metric.

Second, we design and implement an MT system with transfer at various layers of language description, as defined in the framework of FGD. We primarily focus on the tectogrammatical (deep syntactic) layer.

Third, we leave the framework of FGD and experiment with a rather direct, "phrase-based" MT system. Comparing various setups of the system and specifically treating target-side morphological coherence, we are able to significantly improve MT quality and out-perform a commercial MT system within a pre-defined text domain.

The concluding chapter provides a broader perspective on the utility of lexicons in various applications, highlighting the successful features.

Bibliography

Alfred V. Aho and Stephen C. Johnson. Optimal code generation for expression trees. *J. ACM*, 23(3):488–501, 1976. Cited on page 50

ALPAC. Language and Machines: Computers in Translation and Linguistics. Technical report, Automatic Language Processing Advisory Committee, 1966. Cited on page 37

Jurij Apresjan, Igor Boguslavsky, Leonid Iomdin, Alexander Lazursky, Vladimir Sannikov, Victor Sizov, and Leonid Tsinman. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. In *MTT 2003. First International Conference on Meaning-Text Theory*, pages 279–288, Paris, June 2003. Ecole Normale Superieure. Cited on page 85

B. T. Sue Atkins. Theoretical Lexicography and its Relation to Dictionary-making. In W. Frawley, editor, *Dictionaries: the Journal of the Dictionary Society of North America*, pages 4–43, Cleveland, Ohio, 1993. DSNA. Cited on page 19

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 86–90, San Francisco, California, 1998. Morgan Kaufmann Publishers. Cited on page 11, 27

Regina Barzilay and Kathleen R. McKeown. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3):297–328, September 2005. Cited on page 88

Sugato Basu. *Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments*. PhD thesis, Department of Computer Sciences, University of Texas, Austin, Texas, May 2005. Cited on page 34

Václava Benešová and Ondřej Bojar. Czech Verbs of Communication and the Extraction of their Frames. In *Text, Speech and Dialogue: 9th International Conference, TSD 2006*, volume LNAI 3658, pages 29–36. Springer Verlag, September 2006. Cited on page 17, 27, 31

Jeff A. Bilmes and Katrin Kirchhoff. Factored language models and generalized parallel backoff. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 4–6, Morristown, NJ, USA, 2003. Association for Computational Linguistics. Cited on page 79

Alexandra Birch, Miles Osborne, and Philipp Koehn. CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 80

Igor Boguslavsky, Leonid Iomdin, and Victor Sizov. Multilinguality in ETAP-3: Reuse of Lexical Resources. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources*, pages 1–8, Geneva, Switzerland, August 28 2004. COLING. Cited on page 35, 85

Ondřej Bojar and Jan Hajič. Extracting Translation Verb Frames. In Walther von Hahn, John Hutchins, and Christina Vertan, editors, *Proceedings of Modern Approaches in Translation Technologies, workshop in conjunction with Recent Advances in Natural Language Processing (RANLP 2005)*, pages 2–6. Bulgarian Academy of Sciencies, September 2005. Cited on page 35

Ondřej Bojar and Jan Hajič. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 143–146, Columbus, Ohio, June 2008. Association for Computational Linguistics. Cited on page 77

Ondřej Bojar and Zdeněk Žabokrtský. CzEng: Czech-English Parallel Corpus, Release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62, 2006. Cited on page 60, 75

Ondřej Bojar, Jiří Semecký, and Václava Benešová. VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics*, 83:5–17, 2005. Cited on page 10, 18

Ondřej Bojar, Evgeny Matusov, and Hermann Ney. Czech-English Phrase-Based Machine Translation. In *FinTAL 2006*, volume LNAI 4139, pages 214–224, Turku, Finland, August 2006. Springer. Cited on page 59, 72, 80

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. ELRA. Cited on page 43, 77

Ondřej Bojar. Towards Automatic Extraction of Verb Frames. *Prague Bulletin of Mathematical Linguistics*, 79–80:101–120, 2003. Cited on page 33, 34, 40

Ondřej Bojar. Problems of Inducing Large Coverage Constraint-Based Dependency Grammar for Czech. In *Constraint Solving and Language Processing, CSLP 2004*, volume LNAI 3438, pages 90–103, Roskilde University, September 2004. Springer. Cited on page 55

Ondřej Bojar. English-to-Czech Factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 81

Francis Bond and Sanae Fujita. Evaluation of a Method of Creating New Valency Entries. In *Proc. of Machine Translation Summit IX (MT Summit-2003)*, pages 16–23, New Orleans, Louisiana, September 2003. Cited on page 32, 35

Francis Bond, Stephan Oepen, Melanie Siegel, Ann Copestake, and Dan Flickinger. Open source machine translation with DELPH-IN. In *Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit*, pages 15–22, Phuket, Thailand, September 2005. Cited on page 39

Thorsten Brants. TnT - A Statistical Part-of-Speech Tagger . In *ANLP-NAACL 2000*, pages 224–231, Seattle, 2000. Cited on page 44

Peter F. Brown, J. Cocke, S. A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, Robert L. Mercer, and P. S. Roossin. A Statistical Approach to French/English Translation. In *Proc. of the Second International Conference on Theoretical and Methodological Issues is Machine Translation of Natural Languages; Panel 2: Paradigms for MT*, Pittsburg, PA, 1988. Carnegie Mellon University. Cited on page 39

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479, 1992. Cited on page 79

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 59, 76, 77, 86

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June 2008. Association for Computational Linguistics. Cited on page 77

Hiram Calvo, Alexander Gelbukh, and Adam Kilgarriff. Automatic Thesaurus vs. WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment. In *CICLING, 5th Int. Conf. on Intelligent Text Processing and Computational Linguistics*, Mexico City, 2005. Springer Verlag. Cited on page 86

Nicoletta Calzolari, Francesca Bertagna, Alessandro Lenci, Monica Monachini, et al. Standards and Best Practice for Multilingual Computational Lexicons & MILE (the Multilingual ISLE Lexical Entry), 2001. Available at `http://www.w3.org/2001/sw/BestPractices/WNET/ISLE_D2.2-D3.2.pdf`. Cited on page 19

Jean Carletta. Assessing agreement on classification task: The kappa statistics. *Computational Linguistics*, 22(2):249–254, 1996. Cited on page 10

Marine Carpuat and Dekai Wu. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 387–394, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. Cited on page 86

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, 2007. Cited on page 87

John Carroll, Guido Minnen, and Ted Briscoe. Can Subcategorisation Probabilities Help a Statistical Parser. In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora (WVLC-6)*, Montreal, Canada, 1998. Cited on page 84

Eugene Charniak. Immediate-Head Parsing for Language Models. In *Meeting of the Association for Computational Linguistics*, pages 116–123, 2001. Cited on page 49

Ciprian Chelba and Frederick Jelinek. Exploiting Syntactic Structure for Language Modeling. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 225–231, San Francisco, California, 1998. Morgan Kaufmann Publishers. Cited on page 80

Monique Chevalier, Jules Dansereau, and Guy Poulin. TAUM-METEO: description du système. Groupe TAUM, Université de Montréal. Montréal, Canada, 1978. Cited on page 37

David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. Cited on page 80

Silvie Cinková. From PropBank to EngValLex: Adapting PropBank-Lexicon to the Valency Theory of Functional Generative Description. In *Proceedings of LREC 2006*, pages 2170–2175, 2006. Cited on page 11

Martin Čmejrek, Jan Cuřín, and Jiří Havelka. Czech-English Dependency-based Machine Translation. In *EACL 2003 Proceedings of the Conference*, pages 83–90. Association for Computational Linguistics, April 2003. Cited on page 44, 56, 59, 63

Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *Proceedings of LREC 2004*, Lisbon, May 26–28 2004. Cited on page 43, 71

Martin Čmejrek. *Using Dependency Tree Structure for Czech-English Machine Translation*. PhD thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2006. Cited on page 41, 43, 44, 51, 52, 54, 60, 62

Trevor Cohn and Mirella Lapata. Large margin synchronous generation and its application to sentence compression. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 73–82, 2007. Cited on page 43

Michael Collins and Brian Roark. Incremental parsing with the perceptron algorithm. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 111, Morristown, NJ, USA, 2004. Association for Computational Linguistics. Cited on page 21

Michael Collins, Jan Hajič, Eric Brill, Lance Ramshaw, and Christoph Tillmann. A Statistical Parser of Czech. In *Proceedings of 37th ACL Conference*, pages 505–512, University of Maryland, College Park, USA, 1999. Cited on page 84

Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. Cited on page 80

Michael Collins. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191, 1996. Cited on page 44, 63

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995. Cited on page 23

Jan Cuřín. *Statistical Methods in Czech-English Machine Translation*. PhD thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2006. Cited on page 80

Adrià de Gispert, José B. Mariño, and Josep M. Crego. Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Eurospeech 2005*, pages 3185–3188, Lisbon, Portugal, September 2005. Cited on page 80

Ralph Debusmann and Marco Kuhlmann. Dependency grammar: Classification and exploration, 2007. Project report (CHORUS, SFB 378). Cited on page 41

George Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proc. of HLT*, 2002. Cited on page 64

Bonnie J. Dorr and Douglas A. Jones. Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *COLING*, pages 322–327, 1996. Cited on page 32

Bonnie J. Dorr and Olsen Broman Mari. Multilingual Generation: The Role of Telicity in Lexical Choice and Syntactic Realization. *Machine Translation*, 11(1–3):37–74, 1996. Cited on page 11, 27

Bonnie J. Dorr, Pamela W. Jordan, and John W. Benoit. A Survey of Current Paradigms in Machine Translation. Technical Report LAMP-TR-027, UMIACS-TR-98-72, CS-TR-3961, University of Maryland, College Park, December 1998. Cited on page 37, 58

İlknur Durgar El-Kahlout and Kemal Oflazer. Initial Explorations in English to Turkish Statistical Machine Translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 7–14, New York City, June 2006. Association for Computational Linguistics. Cited on page 80

Jason Eisner. Learning Non-Isomorphic Tree Mappings for Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Companion Volume*, pages 205–208, Sapporo, July 2003. Cited on page 41, 80

Marcello Federico and Mauro Cettolo. Efficient Handling of N-gram Language Models for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 88–95, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 58

Christiane Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998. Cited on page 11, 83

Charles J. Fillmore, Charles Wooters, and Collin F. Baker. Building a Large Lexical Databank Which Provides Deep Semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation*, Hong Kong, 2001. Cited on page 11, 29

Charles J. Fillmore. FrameNet and the Linking between Semantic and Syntactic Relations. In Shu-Cuan Tseng, editor, *Proceedings of COLING 2002*, pages xxviii–xxxvi. Howard International House, 2002. Cited on page 11, 29

Sanae Fujita and Francis Bond. A Method of Adding New Entries to a Valency Dictionary by Exploiting Existing Lexical Resources. In *The 9th International Conference on Theoretical and Methodological Issues in Machine Translation, TMI 2002*, pages 42–53, Keihanna, Japan, March 2002. Cited on page 85

Sanae Fujita and Francis Bond. A Method of Creating New Bilingual Valency Entries using Alternations. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources*, pages 41–48, Geneva, Switzerland, August 28 2004. COLING. Cited on page 35, 85

Sharon Goldwater and David McClosky. Improving statistical MT through morphological analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 676–683, Morristown, NJ, USA, 2005. Association for Computational Linguistics. Cited on page 80

Jan Hajič and Barbora Hladká. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*, pages 483–490, Montreal, Canada, 1998. Cited on page 71

Jan Hajič, Alexandr Rosen, and Hana Skoumalová. RUSLAN - systém strojového překladu z češtiny do ruštiny. Technical report, Výzkumný ústav matematických strojů, Prague, Czech Republic, 1987. Cited on page 5

Jan Hajič, Eva Hajičová, Milena Hnátková, Vladislav Kuboň, Jarmila Panevová, Alexandr Rosen, Petr Sgall, and Hana Skoumalová. MATRACE – MAchine TRAnslation between Czech and English. In *Proceedings of the IBM Academic Initiative Projects Seminar, Praha, November 1992*, pages 75–82, Praha, 1992. České vysoké učení technické. Cited on page 32

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68. Växjö University Press, November 14–15, 2003 2003. Cited on page 10

Jan Hajič. RUSLAN: an MT system between closely related languages. In *Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, pages 113–117. Association for Computational Linguistics, 1987. Cited on page 5

Jan Hajič. Complex Corpus Annotation: The Prague Dependency Treebank. In *Insight into Slovak and Czech Corpus Linguistics*, Bratislava, Slovakia, 2004. Jazykovedný ústav Ľ. Štúra, SAV. Cited on page 43

Jan Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, Prague, 2004. Cited on page 44, 67

Jan Hajič, Martin Čmejrek, Bonnie Dorr, Yuan Ding, Jason Eisner, Daniel Gildea, Terry Koo, Kristen Parton, Gerald Penn, Dragomir Radev, and Owen Rambow. Natural Language Generation in the Context of Machine Translation. Technical report, Johns Hopkins University, Center for Speech and Language Processing, 2002. NLP WS'02 Final Report. Cited on page 41

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4, 2006. Cited on page 5, 7, 10, 11

Keith Hall. k-best Spanning Tree Parsing. In *Proceedings of the ACL 2007*, Prague, Czech Republic, June 2007. Cited on page 62

Dana Hlaváčková and Aleš Horák. VerbaLex - New Comprehensive Lexicon of Verb Valencies for Czech. In *Computer Treatment of Slavic and East European Languages*, pages 107–115, Bratislava, Slovakia, 2006. Slovenský národný korpus. Cited on page 9, 11

Dana Hlaváčková, Aleš Horák, and Vladimír Kadlec. Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. In *Text, Speech and Dialogue: 9th International Conference, TSD 2006*, volume LNAI 3658. Springer Verlag, September 2006. Cited on page 5, 84

Tomáš Holan, Vladislav Kuboň, Karel Oliva, and Martin Plátek. On Complexity of Word Order. *Special Issue on Dependency Grammar of the journal TAL (Traitement Automatique des Langues)*, 41(1):273–300, 2000. Cited on page 41, 42

Tomáš Holan. K syntaktické analýze českých(!) vět. In *MIS 2003*. MATFYZPRESS, January 18–25, 2003 2003. Cited on page 40, 87

Albert Sydney Hornby. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, 3 edition, 1974. Cited on page 32

Liang Huang, Kevin Knight, and Aravind Joshi. Statistical Syntax-Directed Translation with Extended Domain of Locality. In *Proc. of 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, MA, 2006. Cited on page 50, 62

John Hutchins. The state of machine translation in Europe. In *Expanding MT horizons: proceedings of the Second Conference of the Association for Machine Translation in the Americas*, pages 198–205, Montreal, Quebec, Canada, October 1996. Cited on page 37

John Hutchins. ALPAC: the (in)famous report. In S. Nirenburg, H. Somers, and Y. Wilks, editors, *Readings in machine translation.*, pages 131–135. The MIT Press, Cambridge, Mass., 2003. Originally published in MT News International 14, June 1996, 9-12. Cited on page 37

John Hutchins. The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954. Expanded version of AMTA-2004 paper. `http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf`, November 2005. Cited on page 37

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. Toward an MT System without Pre-Editing — Effects of New Methods in ALT-J/E. In *Proceedings of MT Summit III*, pages 101–106, 1991. Cited on page 35, 85

Ray Jackendoff. *Semantic Structures*. The MIT Press, 1990. Cited on page 11, 27

Aravind K. Joshi, Leon S. Levy, and Masako Takahashi. Tree adjunct grammars. *J. Comput. Syst. Sci.*, 10(1):136–163, 1975. Cited on page 41, 43

Aravind K. Joshi, K. Vijay Shanker, and David Weir. The Convergence of Mildly Context-Sensitive Grammar Formalisms. Technical Report MS-CIS-90-01, University of Pennsylvania Department of Computer and Information Science, 1990. Cited on page 41

George Karypis. CLUTO - A Clustering Toolkit. Technical Report #02-017, University of Minnesota, Department of Computer Science, November 2003. Cited on page 25

Adam Kilgarriff. Language is never ever ever random. *Corpus Linguistics and Linguistic Theory*, 1(2):263–276, 2005. Cited on page 29

Gilbert W. King. Stochastic Methods of Mechanical Translation. *Mechanical Translation*, 3(2):38–39, 1956. Cited on page 39

Paul Kingsbury and Martha Palmer. From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain, 2002. Cited on page 39

Paul Kingsbury, Martha Palmer, and Mitch Marcus. Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*, 2002. Cited on page 11

Paul Kingsbury. Verb clusters from PropBank annotation. Technical report, University of Pennsylvania, Philadelphia, PA, 2004. Cited on page 32

Karin Kipper, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press / The MIT Press, 2000. Cited on page 11

Karen Kipper-Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005. Cited on page 11, 32

Zdenek Kirschner and Alexandr Rosen. Apac - an experiment in machine translation. *Machine Translation*, 4(3):177–193, 1989. Cited on page 39

Václav Klimeš. *Analytical and Tectogrammatical Analysis of a Natural Language*. PhD thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2006. Cited on page 44, 62

Jan Kocek, Marie Kopřivová, and Karel Kučera, editors. *Český národní korpus - úvod a příručka uživatele*. FF UK - ÚČNK, Praha, 2000. Cited on page 43

Philipp Koehn and Hieu Hoang. Factored Translation Models. In *Proc. of EMNLP*, 2007. Cited on page 57

Philipp Koehn and Kevin Knight. Empirical Methods for Compound Splitting. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 187–193, Morristown, NJ, USA, 2003. Association for Computational Linguistics. Cited on page 80

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 50

Philipp Koehn. Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In Robert E. Frederking and Kathryn Taylor, editors, *AMTA*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124. Springer, 2004. Cited on page 50, 68, 70, 71

Philipp Koehn. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, Barcelona, Spain, 2004. Cited on page 59

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, 2005. Cited on page 67

Anna Korhonen. Subcategorization Acquisition. Technical Report UCAM-CL-TR-530, University of Cambridge, Computer Laboratory, Cambridge, UK, February 2002. Cited on page 17

Květoslava Králíková and Jarmila Panevová. ASIMUT - A Method for Automatic Information Retrieval from Full Textts. *Explicite Beschreibung der Sprache und automatische Textbearbeitung*, XVII, 1990. Cited on page 84

Geert-Jan M. Kruijff. 3-Phase Grammar Learning. In *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*, 2003. Cited on page 40

Marco Kuhlmann and Mathias Möhl. Mildly context-sensitive dependency languages. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 160–167, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 41

Elina Lagoudaki. Translation Memories Survey 2006: Users' perceptions around TM use. In *Proceedings of the ASLIB International Conference Translating and the Computer 28*, London, UK, November 2006. Cited on page 37

Beth C. Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL, 1993. Cited on page 13, 27, 32

Kenneth C. Litkowski. Computational Lexicons and Dictionaries. In Keith Brown, editor, *Encyclopedia of Language and Linguistics (2nd ed.)*. Elsevier Publishers, Oxford, 2005. `http://www.clres.com/online-papers/ell.doc`. Cited on page 83, 88

Yang Liu, Qun Liu, and Shouxun Lin. Log-linear models for word alignment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 459–466, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. Cited on page 35, 85

Markéta Lopatková and Jarmila Panevová. Recent developments of the theory of valency in the light of the Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistic*, pages 83–92. Veda Bratislava, Slovakia, 2005. Cited on page 8, 10, 12, 86

Markéta Lopatková, Zdeněk Žabokrtský, and Václava Benešová. Valency Lexicon of Czech Verbs VALLEX 2.0. Technical Report 34, UFAL MFF UK, 2006. Cited on page 13

Markéta Lopatková, Zdeněk Žabokrstký, and Karolína Skwarska. Valency Lexicon of Czech Verbs: Alternation-Based Model. In *Proceedings of LREC 2006*, pages 1728–1733. ELRA, 2006. Cited on page 13

Markéta Lopatková, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves*. Univerzita Karlova v Praze, Nakladatelství Karolinum, Praha, 2008. In cooperation with Karolína Skwarska, Eduard Bejček, Klára Hrstková, Michaela Nová and Miroslav Tichý. Cited on page 10, 13

Markéta Lopatková. Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *Prague Bulletin of Mathematical Linguistics*, 79–80:37–60, 2003. Cited on page 11

Adam Lopez and Philip Resnik. Word-Based Alignment, Phrase-Based Translation: What's the Link? In *Proc. of 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 90–99, Boston, MA, August 2006. Cited on page 85

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT/EMNLP 2005*, October 2005. Cited on page 44, 62

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In *Proceedings of HLT 2002*, 2002. Cited on page 88

Igor A. Mel'čuk. *Dependency Syntax - Theory and Practice*. Albany: State University of New York Press, 1988. Cited on page 39, 85

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The NomBank Project: An Interim Report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. Cited on page 39

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep., 2006. Cited on page 5, 11, 56

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank. In *Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004*, 2004. Cited on page 39

Guido Minnen, John Carroll, and Darren Pearce. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223, 2001. Cited on page 44, 72

Raymond J. Mooney. Learning for Semantic Interpretation: Scaling Up without Dumbing Down. In James Cussens and Sašo Džeroski, editors, *Learning Language in Logic, LLL'99*, volume LNAI 1925, pages 57–66, Berlin Heidelberg, 2000. Springer-Verlag. Cited on page 88

Thai Phuong Nguyen and Akira Shimazu. Improving Phrase-Based SMT with Morpho-Syntactic Analysis and Transformation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*, pages 138–147, August 2006. Cited on page 80

Sonja Nießen and Hermann Ney. Toward hierarchical models for statistical machine translation of inflected languages. In *Proceedings of the workshop on Data-driven methods in machine translation*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics. Cited on page 80

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, 2007. Cited on page 40

Brian Oakley. To do the right thing for the wrong reason, the Eurotra experience. In *MT Summit V Proceedings*, Luxembourg, July 1995. Cited on page 37

Franz Josef Och and Hermann Ney. A Comparison of Alignment Models for Statistical Machine Translation. In *Proceedings of the 17th conference on Computational linguistics*, pages 1086–1090. Association for Computational Linguistics, 2000. Cited on page 52

Franz Josef Och and Hermann Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *ACL*, pages 295–302, 2002. Cited on page 46, 69

Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003. Cited on page 72

Franz Joseph Och. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, RWTH Aachen University, 2002. Cited on page 48, 49

Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7 2003. Cited on page 58, 63, 69

Franz Joseph Och. Statistical Machine Translation: Foundations and Recent Advances. Tutorial at MT Summit 2005, September 2005. Cited on page 87

Stephan Oepen, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. Towards Hybrid Quality-Oriented Machine Translation. On Linguistics and Probabilities in MT. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, Skövde, Sweden, 2007. Cited on page 39

Karel Oliva. A Parser for Czech Implemented in Systems Q. *Explicite Beschreibung der Sprache und automatische Textbearbeitung*, XVI, 1989. Cited on page 5

Karel Pala and Pavel Ševeček. Valence českých sloves. In *Sborník prací FFBU*, pages 41–54, Brno, 1997. Cited on page 11, 32

Karel Pala and Pavel Smrž. Building Czech Wordnet. *Romanian Journal of Information, Science and Technology*, 7(1-2):79–88, 2004. Cited on page 11

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005. Cited on page 27

Jarmila Panevová. *Formy a funkce ve stavbě české věty [Forms and functions in the structure of the Czech sentence]*. Academia, Prague, Czech Republic, 1980. Cited on page 7, 20

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, 2002. Cited on page 59

Carl J. Pollard and Ivan A. Sag. *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago, 1994. Cited on page 39

Maja Popović and Hermann Ney. Improving Word Alignment Quality using Morpho-Syntactic Information. In *Proceedings of COLING 2004*, Geneva, Switzerland, August 23–27 2004. Cited on page 80

Paul Procter, editor. *Longman Dictionary of Contemporary English*. Longman Group, Essex, England, 1978. Cited on page 32

Jan Ptáček and Zdeněk Žabokrtský. Synthesis of Czech Sentences from Tectogrammatical Trees. In *Proc. of TSD*, pages 221–228, 2006. Cited on page 5, 10, 44, 60, 61, 62, 64

J. Ross Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, 1986. Cited on page 23

J. Ross Quinlan. Bagging, boosting, and c4.5. In *Proceedings, Fourteenth National Conference on Artificial Intelligence*, 1996. Cited on page 15

J. Ross Quinlan. Data Mining Tools See5 and C5.0, 2002. `http://www.rulequest.com/see5-info.html`. Cited on page 23, 32

Christopher Quirk and Arul Menezes. Dependency Treelet Translation: The Convergence of Statistical and Example-Based Machine-Translation? *Machine Translation*, 20(1):43–65, 2006. Cited on page 80

Chris Quirk, Arul Menezes, and Colin Cherry. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 271–279. Association for Computational Linguistics, 2005. Cited on page 62

Adwait Ratnaparkhi. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania, May 1996. Cited on page 44, 71

Stephen D. Richardson, William B. Dolan, Arul Menezes, and Monica Corston-Oliver. Overcoming the Customization Bottleneck Using Example-Based MT. In *Proceedings of the workshop on Data-driven methods in machine translation*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics. Cited on page 39

Stefan Riezler and III John T. Maxwell. Grammatical Machine Translation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 248–255, Morristown, NJ, USA, 2006. Association for Computational Linguistics. Cited on page 64

Alexandr Rosen, Eva Hajičová, and Jan Hajič. Derivation of underlying valency frames from a learner's dictionary. In *Proceedings of the 14th conference on Computational linguistics*, pages 553–559, Nantes, France, 1992. Association for Computational Linguistics. Cited on page 32

Alexandr Rosen. In Defense of Impractical Machine Translation Systems. `http://utkl.ff.cuni.cz/~rosen/public/pognan.ps.gz`, 1996. Cited on page 37

Pavel Rychlý and Pavel Smrž. Manatee, Bonito and Word Sketches for Czech. In *Proceedings of the Second International Conference on Corpus Linguisitcs*, pages 124–131, 2004. Cited on page 16, 29, 34

Anoop Sarkar and Daniel Zeman. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics (Coling 2000)*, Saarbrücken, Germany, 2000. Universität des Saarlandes. Cited on page 33

Sabine Schulte im Walde. *Experiments on the Automatic Induction of German Semantic Verb Classes*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2003. Published as AIMS Report 9(2). Cited on page 33

Jiří Semecký and Petr Podveský. Extensive Study on Automatic Verb Sense Disambiguation in Czech. In Petr Sojka, Ivan Kopecek, and Karel Pala, editors, *TSD*, volume 4188 of *Lecture Notes in Computer Science*, pages 237–244. Springer, 2006. Cited on page 10

Jiří Semecký. *Verb Valency Frames Disambiguation*. PhD thesis, Charles University, Prague, 2007. Cited on page 21

Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands, 1986. Cited on page 5

Hana Skoumalová. *Czech syntactic lexicon*. PhD thesis, Univerzita Karlova, Filozofická fakulta, 2001. Cited on page 11, 32

David A. Smith and Jason Eisner. Minimum-Risk Annealing for Training Log-Linear Models. In *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL), Companion Volume*, pages 787–794, Sydney, July 2006. Cited on page 58

David A. Smith and Jason Eisner. Quasi-Synchronous Grammars: Alignment by Soft Projection of Syntactic Dependencies. In *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 23–30, New York, June 2006. Cited on page 62

Zoltan Somogyi, Fergus Henderson, and Thomas Conway. Mercury: An Efficient Purely Declarative Logic Programming Language. In *Proceedings of the Australian Computer Science Conference*, pages 499–512, Glenelg, Australia, February 1995. Cited on page 58

Mark Stevenson. *Word Sense Disambiguation: The Case for Combinations of Knowledge Sources*. CSLI Publications, 2003. Cited on page 13, 19, 86

Markéta Straňáková-Lopatková and Zdeněk Žabokrtský. Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, volume 3, pages 949–956. ELRA, 2002. LN00A063. Cited on page 5, 9

David Talbot and Miles Osborne. Modelling Lexical Redundancy for Machine Translation. In *Proc. of COLING and ACL 2006*, pages 969–976, Sydney, Australia, 2006. Cited on page 80

Bernard Vauquois. La traduction automatique à Grenoble. Document de linguistique quantitative 24. Dunod, Paris., 1975. Cited on page 38

Jean Véronis. Sense tagging: does it make sense? In *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech*. Peter Lang, 2003. Published version of a paper presented at the Corpus Linguistics'2001 Conference, Lancaster, U.K. Cited on page 10, 27

Piek Vossen. Introduction to EuroWordNet. *Computers and the Humanities, Special Issue on EuroWordNet*, 32(2–3), 1998. Cited on page 11

Yuk Wah Wong and Raymond Mooney. Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 960–967, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Cited on page 88

William A. Woods, Lawrence A. Bookman, Ann Houston, Robert J. Kuhns, Paul Martin, and Stephen Green. Linguistic knowledge can improve information retrieval. Technical Report TR-99-83, Sun Labs, December 1999. Cited on page 84

Dekai Wu and Hongsing Wong. Machine translation with a stochastic grammatical channel. In *COLING-ACL*, pages 1408–1415, 1998. Cited on page 80

Kenji Yamada and Kevin Knight. A decoder for syntax-based statistical mt. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 303–310, Morristown, NJ, USA, 2002. Association for Computational Linguistics. Cited on page 80

Mei Yang and Katrin Kirchhoff. Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages. In *EACL 2006*, April 2006. Cited on page 79

Zdeněk Žabokrtský and Ondřej Bojar. TectoMT, Developer's Guide. Technical Report TR-2008-39, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, December 2008. Cited on page 44, 62

Zdeněk Žabokrtský and Markéta Lopatková. Valency Frames of Czech Verbs in VALLEX 1.0. In *Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference*, pages 70–77, May 6, 2004 2004. Cited on page 11

Zdeněk Žabokrtský, Václava Benešová, Markéta Lopatková, and Karolina Skwarská. Tektogramaticky anotovaný valenční slovník českých sloves. Technical Report TR-2002-15, ÚFAL/CKL, Prague, Czech Republic, 2002. Cited on page 11

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, USA, 2008. Cited on page 63, 78

Zdeněk Žabokrtský. *Valency Lexicon of Czech Verbs*. PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague, 2005. Cited on page 8

Daniel Zeman. Can Subcategorization Help a Statistical Parser? In *Proceedings of the 19th International Conference on Computational Linguistics (Coling 2002)*, Taibei, Tchaj-wan, 2002. Zhongyang Yanjiuyuan (Academia Sinica). Cited on page 84

Richard Zens, Oliver Bender, Saša Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuqi Zhang, and Hermann Ney. The RWTH Phrase-based Statistical Machine Translation System. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October 2005. Cited on page 50, 68

Yi Zhang, Timothy Baldwin, and Valia Kordoni. The Corpus and the Lexicon: Standardising Deep Lexical Acquisition Evaluation. In *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, pages 152–159, Prague, Czech Republic, June 2007. Cited on page 17, 35

Le Zhang. Maximum Entropy Modeling Toolkit for Python and C++. `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`, 2004. Cited on page 21, 23

George Kingsley Zipf. The Meaning-Frequency Relationship of Words. *Journal of General Psychology*, 3:251–256, 1945. Cited on page 9, 13, 29

Andreas Zollmann, Ashish Venugopal, and Stephan Vogel. Bridging the Inflection Morphology Gap for Arabic Statistical Machine Translation. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics, 2006. Cited on page 80

List of Figures

List of Tables

Index