Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

Využití lingvistických dat ve strojovém překladu

Ondřej Bojar

Praha, 2008

Ústav formální a aplikované lingvistiky
I-3 Matematická lingvistika

Disertační práce byla vypracována v rámci doktorského studia na Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy v Praze v letech 2003 až 2008.

Uchazeč:	RNDr. Ondřej Bojar
Školitel:	RNDr. Vladislav Kuboň, Ph.D. Ústav formální a aplikované lingvistiky (ÚFAL)
Školicí pracoviště:	Ústav formální a aplikované lingvistiky (ÚFAL) Matematicko-fyzikální fakulta Univerzita Karlova v Praze Malostranské náměstí 25, 118 00 Praha 1
Oponenti:	Ing. Alexandr Rosen, Ph.D. Ústav teoretické a komputační lingvistiky Filozofická fakulta Univerzita Karlova v Praze Celetná 13, 110 00 Praha 1
	RNDr. Jan Cuřín, Ph.D. IBM Česká republika, spol. s r.o. Voice Technologies and Systems V Parku 2294/4, 148 00 Praha 4 – Chodov
Předsedkyně OR I-3	Prof. PhDr. Jarmila Panevová, DrSc.

Autoreferát byl rozeslán dne 1. srpna 2008.

Obhajoba se koná dne 3. září. 2006 v 11:00 hodin před komisí pro obhajoby disertačních prací oboru I-3 Matematická lingvistika na Matematicko-fyzikální fakultě UK, Malostranské nám. 25, Praha 1, místnost S1 (4. patro).

S doktorskou disertační prací je možno seznámit se na studijním oddělení doktorského studia, Ke Karlovu 3, Praha 2.

Charles University in Prague
Faculty of Mathematics and Physics

Exploiting Linguistic Data in Machine Translation

Ondřej Bojar

Prague, 2008

Institute of Formal and Applied Linguistics
I-3 Mathematical Linguistics

The results comprised in the thesis were obtained within the candidate's doctoral studies at the Faculty of Mathematics and Physics, Charles University in Prague (MFF UK) during the years 2003–2008.

Candidate:	RNDr. Ondřej Bojar
Supervisor:	RNDr. Vladislav Kuboň, Ph.D. Institute of Formal and Applied Linguistics (ÚFAL)
Department:	Institute of Formal and Applied Linguistics (ÚFAL) Faculty of Mathematics and Physics Charles University in Prague Malostranské náměstí 25, 118 00 Praha 1
Opponents:	Ing. Alexandr Rosen, Ph.D. Institute of Theoretical and Computational Linguistics Faculty of Philosophy & Arts Charles University in Prague Celetná 13, 110 00 Praha 1
	RNDr. Jan Cuřín, Ph.D. IBM Česká republika, spol. s r.o. Voice Technologies and Systems V Parku 2294/4, 148 00 Praha 4 – Chodov
I-3 Board Chair	Prof. PhDr. Jarmila Panevová, DrSc.

The summary was disseminated on August 1, 2008.

The thesis defense will be held on September 3, 2008 at 11.00 a.m. in the building of MFF UK, Malostranské nám. 25, Praha 1, room S1 (4^{th} floor).

The thesis is available for perusal at the Study and Students' Affairs Division, Doctoral Study, Ke Karlovu 3, Praha 2.

1 Introduction

This thesis 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. The text 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), they are linked as follows:

One of the key components of FGD 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. Such lexicons are assumed to be central components of various NLP applications but their development is costly. In Chapter 2, we explore the possibility of automatic suggestion of lexicon entries based on corpus data.

In Chapter 3, we study how the theory of FGD lends itself to practical employment in MT. After a brief review of various approaches to MT, we describe and evaluate our system of syntax-based machine translation. The valency lexicons discussed in Chapter 2 are included in our system to a limited extent only but we observe far more important problems than their lack of coverage.

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 of the thesis.

2 Extracting Verb Valency Frames

Verb valency frames formally describe the potential of a verb to combine with other elements in the sentence. When analyzing an input sentence, the knowledge of the verb frame allows to resolve ambiguity at various levels. When generating text from some deep representation, the valency frame of the verb is used to choose the appropriate morphemic form (e.g. the preposition and case) of a modifier and thus to guarantee the grammaticality of the output, e.g. Ptáček and Žabokrtský (2006).

2.1 Layers of Language Description and Valency Theory

FGD as implemented in the Prague Dependency Treebank (PDT, Hajič *et al.* (2006)) defines three layers of language representation called **morphological** (or m-layer), **analytical** (a-layer, corresponds to surface syntax) and **tectogrammatical** (t-layer, corresponds to deep syntax: only words bearing meaning have corresponding

nodes) to annotate an original text (the wordform, w-layer, where even typographical errors are stored verbatim, e.g. no space between *do* and *lesa*), see Figure 1.

In FGD, (verb) **valency frames** are defined at the t-layer only and describe formal requirements on the immediate dependents of the verbal t-node (Panevová, 1980; Hajič *et al.*, 2006). A brief summary of the key definitions is available in the thesis.

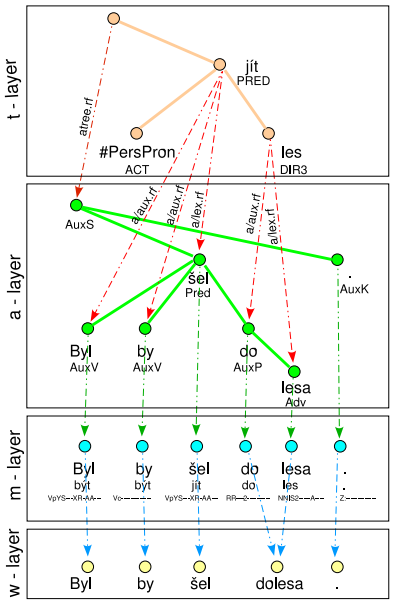


Figure 1: Layers of annotation as implemented in PDT.

2.2 VALLEX

VALLEX (Lopatková *et al.*, 2008) is a valency lexicon of Czech verbs developed in the framework of FGD. At the topmost level, VALLEX is a list of **verb entries**. The verb is characterized by its **headword lemma** (including a reflexive particle *se* or *si*, if appropriate) equipped with verb aspect (perfective, imperfective, biaspectual). Every verb entry includes one or more **valency frames** of the verb. Every valency frame consists of a set of **valency slots** characterizing complementations of the verb. Each slot describes the type of the syntactico-semantic relation between the verb and its complementation (by means of a **tectogrammatical functor**, such as Actor ACT, Patient PAT, Direction $DIR1$; see FGD) as well as all allowed surface realizations (**morphemic forms**) of the verb complementation (e.g. the required preposition and case or the subordinating conjunction for dependent clauses).

We use the term **verb lemma** to denote the infinitive of the verb, excluding a possible reflexive particle and homograph distinction, e.g. *odpovídat* is the verb lemma for the verbs *odpovídat* and *odpovídat se*.

The first version of VALLEX 1.0 was publicly released in 2003 and contained over 1,400 verb entries. The set of covered verbs was extended to about 2,500 verb entries in VALLEX 1.5, an internal version released in 2005.

VALLEX 1.5 covers around 66% of verb occurrences in the Czech National Corpus; 23% of verb occurrences belong to a few frequent auxiliary verbs, esp. *být, bývat (to be)*. The remaining 10% occurrences belong to verbs with low corpus frequency. The distribution of verbs closely follows Zipf's law and there are about 28k additional verbs needed just to cover our particular corpus. An automated method of lexical extraction would save a lot of labour.

VALEVAL (Bojar *et al.*, 2005) evaluated the inter-annotator agreement of annotating verb occurrences with VALLEX 1.0 frames. The level 75% of pairwise agreement we achieved is no worse than results for other languages, but a better match is certainly desirable. VALEVAL experiment provided VALLEX developers with a valuable feedback and also lead to the creation of "Golden VALEVAL" corpus, the collection of sentences where three annotators agreed on the frame of the verb. Golden VALEVAL contains 108 verbs in 7804 sentences (72±26 sentences per verb), annotated with a single VALLEX frame that was used in the sentence.

2.3 Learning Task and Evaluation Metrics

Our learning task is to provide a test verb lemma v_t with a hypothesized frame set H. For the purpose of evaluation of our learning methods, we always choose a known v_t from a dictionary that provides the "golden frame set" G for comparison.

Methods of frame extraction are usually evaluated in terms of precision and recall of either frames as wholes or of individual frame elements (slots). We report on the frame-based **precision** and **recall**:

$$\text{P}(H,G) = \frac{|H \cap G|}{|H|} \quad \text{R}(H,G) = \frac{|H \cap G|}{|G|} \tag{1}$$

To better account for the fine structure of VALLEX frames, our main focus lies in a novel metric, **frame edit distance** (FED) proposed in Benešová and Bojar (2006): FED represents the minimum number of edit operations (insert, delete, replace) necessary to convert a hypothesized frame into the correct frame. FED is extended to **entry similarity** (ES) that compares whole sets of frames:

$$ES(H,G) = 1 - \frac{\min FED(G,H)}{FED(G,\emptyset) + FED(H,\emptyset)} \tag{2}$$

ES attempts to capture how much of lexicographic labour has been saved thanks to the contribution of the automatic frame-generation procedure. If the system did not suggest anything ($H = \emptyset$), the ES is 0%. If the system suggested exactly all the golden frames ($H = G$ and thus $FED(G,H) = 0$), the ES achieves 100%. With this explanation in mind, we use the term **expected saving** (ES) as a synonym to "entry similarity".

2.4 Direct Methods of Learning VALLEX Frames

This section is devoted to the description and comparison of three rather direct methods of frame extraction. An additional method PatternSearch is described in Section 2.5.

Word-Frame Disambiguation (WFD) Inspired by frame-disambiguation techniques by Semecký (2007), we train a classifier on training examples of all known verbs ignoring their lemmas (i.e. pretending that all annotated verb occurrences belong to the same verb). The classifier is then used to predict the most likely frame to each occurrence of the test verb v_t. Finally, all suggested frames are collected and returned as the hypothesized frame set H for v_t.

Deep Syntactic Distance (DSD) Like WFD, DSD uses frame-labelled occurrences of training verbs to predict frames to occurrences of the test verb v_t. In WFD, the prediction is handled by a classifier with features extracted from surface-syntactic neighbourhood of the verb occurrence. In DSD, we explicitly consider the underlying deep-syntactic layer and estimate how difficult it is to believe that an occurrence of v_t has the same frame as an occurrence of a known verb. In a sense, DSD is the "nearest neighbour" variant of WFD.

Method	Options	Fit Frame Count	Avg ES	Avg Prec	Avg Rec
WFD		no	21.4±4.7	4.1±1.4	26.9±11.1
Baseline	1×ACT-PAT	no	27.7±4.9	45.7±21.9	9.7±6.8
Baseline	2×ACT-PAT	no	38.8±4.9	22.8±11.0	9.7±6.8
Decomp		no	43.0±1.5	4.2±2.1	4.3±2.0
DSD	Penalize, ReqObl	no	43.1±8.1	7.9±6.5	14.2±11.3
Baseline	3×ACT-PAT	no	43.7±3.6	15.2±7.3	9.7±6.8
Baseline	avg×ACT-PAT	no	45.3±4.6	5.9±2.7	9.7±6.8
Baseline	4×ACT-PAT	no	46.8±3.2	11.4±5.5	9.7±6.8
DSD	Penalize, ReqObl	CLUST	62.2±9.3	11.7±8.0	11.7±8.0
Decomp		SIMPLE/CLUST	64.5±3.6	4.5±2.0	4.5±2.0
Baseline	expected×ACT-PAT	SIMPLE	65.3±3.8	9.7±6.8	9.7±6.8
WFD		CLUST	66.0±3.1	13.4±8.6	13.4±8.6
WFD		SIMPLE	67.8±1.1	12.7±3.3	12.6±3.3

Table 1: Evaluation of direct frame suggestion methods.

Learning Frames by Decomposition (Decomp) Both WFD and DSD assume frames are opaque units and rely on a similarity between verb occurrences. In Decomp, we decompose frames into independent "frame components" (e.g. "frame-has-PAT", "ADDR-is-obligatory", "PAT-expressed-as-acc"). We train separate classifiers to predict which frame components are present in the frame of a particular occurrence of v_t. The hypothesized frame set H is constructed by recombining the suggested components.

Post-processing of Suggested Frame Sets ES is very sensitive to any difference in the number of frames expected vs. proposed. Instead of trying to predict the correct cardinality of the frame set of a test verb v_t based e.g. on the frequency of v_t in a corpus or on the number of translation equivalents of v_t to a foreign language, we use two methods that modify a suggested frame set to *match the expected* number of frames for each verb, thus allowing the methods to peek at the test data partly.

If fewer frames than expected were hypothesized, additional baseline frames (*ACT.obl.nom PAT.obl.acc*) are added to reach the expected count. If more frames were hypothesized, only those with a high support (SIMPLE) or the centroids of automatically generated clusters (CLUST) are taken into account.

Empirical Evaluation of Direct Methods The methods were evaluated on VALEVAL verbs and frame sets from VALLEX 1.0. In every fold we pick one tenth of verb lemmas as the test verbs. The remaining 9/10s of the verbs and their VALEVAL occurrences are available to the methods for training. Every method has to produce a frame set for every test verb based on unlabelled occurrences in the VALEVAL corpus. The results are in Table 1.

The column "Fit Frame Count" specifies whether the framesets were post-processed to match the expected (correct) number of frames (SIMPLE or CLUST). Our "Baseline" method is to suggest one to four copies of a frame with two obligatory slots: *ACT.obl.nom PAT.obl.acc*.

We observe that baseline methods generally perform better than our frame-suggestion techniques both in case the methods do not access the expected number

of frames as well as when they do. It is only WFD (CLUST and SIMPLE) that insignificantly outperforms the baseline.

2.5 PatternSearch: Guessing Verb Semantic Class

As seen above, direct methods of frame suggestion averaged over all verbs do not bring much improvement over the baseline. In this section, we tackle frame suggestion indirectly: we first automatically guess, which verbs belong to a particular semantic and then suggest the most typical frame of that class to them. In this preliminary experiment published in Benešová and Bojar (2006), we focus on one class: so-called **verbs of communication** rendering the situation when "a speaker conveys information to a recipient".

Automatic Identification of Verbs of Communication We search a corpus for verbs accompanied by: (1) a noun in one of the following cases: genitive, dative or accusative (to approximate the "recipient" slot) and (2) a dependent clause introduced by one of the set of characteristic subordinating conjunctions (*že*, *aby*, *ať*, *zda* or *jestli*) (to approximate the slot of "information").

Sorting all verbs by the descending number of occurrences of the tested pattern, we obtain a ranking of verbs according to their "communicative character". For all possible cut-off thresholds, Figure 2 plots the **true positive rate** (correctly recognized verbs of communication) against the **true negative rate** (correctly recognized verbs without a communication sense). The left chart compares the performance against three golden standards (VALLEX 1.0, VALLEX 1.5 and translations of English verbs in the Communication frame and derived frames in FrameNet 1.2.[1]), the right chart gives further details on the contribution of different subordinating conjunctions.

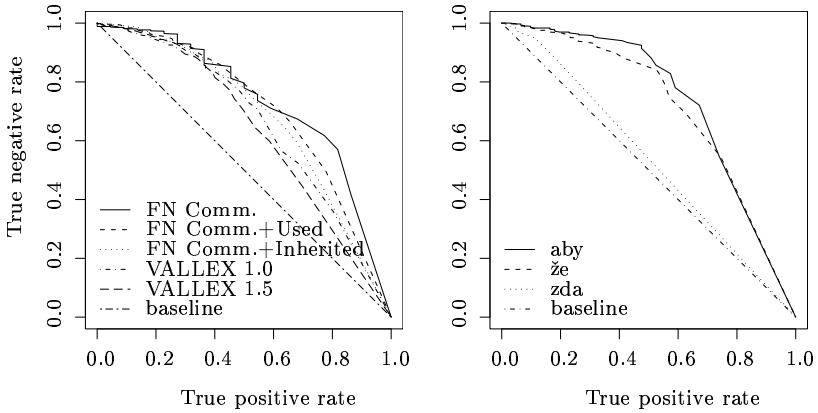


Figure 2: Verbs of communication as suggested by the pattern V+N234+subord, evaluated against VALLEX and FrameNet (left) and evaluated against VALLEX 1.0 for the three main subordinating conjunctions (*aby*, *že*, *zda*) independently (right).

[1] `http://framenet.icsi.berkeley.edu/`

The closer the curve lies to the upper right corner, the better the performance. With an appropriate threshold, about 40% to 50% of verbs of communication are identified correctly while 20% of non-communication verbs are falsely marked. We obtain about the same performance level for both VALLEX and FrameNet-based evaluation. This confirms that our method is not too tightly tailored to the classification introduced in VALLEX.

Application to Frame Suggestion For all verbs occurring frequently enough in the typical pattern, we propose the most typical "communication frame" consisting of ACT, ADDR and PAT (all obligatory). Every verb of communication can have some additional senses not noticed by our method but at least the communication frame should be suggested correctly.

Suggested frames	*ES* **[%]**
Specific frame for verbs of communication, default for others	38.00 ± 0.19
Baseline 1: ACT(1)	26.69 ± 0.14
Baseline 2: ACT(1) PAT(4)	37.55 ± 0.18
Baseline 3: ACT(1) PAT(4) ADDR(3,4)	35.70 ± 0.17
Baseline 4: Two identical frames: ACT(1) PAT(4)	39.11 ± 0.12

Table 2: Expected saving when suggesting frame entries automatically.

Table 2 displays the *ES* as reported in Benešová and Bojar (2006) of four various baselines and the result obtained by our method. We can slightly improve over Baseline 2 if we first identify verbs of communication automatically and assign ACT PAT ADDR with appropriate subordinating conjunctions to them, leaving other verbs with ACT PAT only. This confirms our assumption that verbs of communication have a typical three-slot frame and also that our method managed to identify some of the verbs correctly.

2.6 Discussion

Reasons of the failure of our direct methods include **lack of semantic information** (only PatternSearch that used verb classes in VALLEX gave somewhat promising results), no treatment of **deletability of modifiers** (i.e. the fact that even obligatory modifiers are often not present in the sentence), relatively **limited fine-tuning of features and training data** (by carefully selecting which training and test occurrences we consider, the noise in predicting frames to verb occurrences could be greatly reduced), and **lack of manual intervention** in the rather complex lexicographic process.

Ideally, the lexicons considered in this chapter would improve NLP applications. To better understand practical needs of NLP applications, we now experiment with a syntax-based (Chapter 3) and a phrase-based (Chapter 4) machine translation system.

3 Machine Translation via Deep Syntax

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. 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 it uses some formal representation (partially) describing the meaning of the sentence. We examine both of the approaches further below.

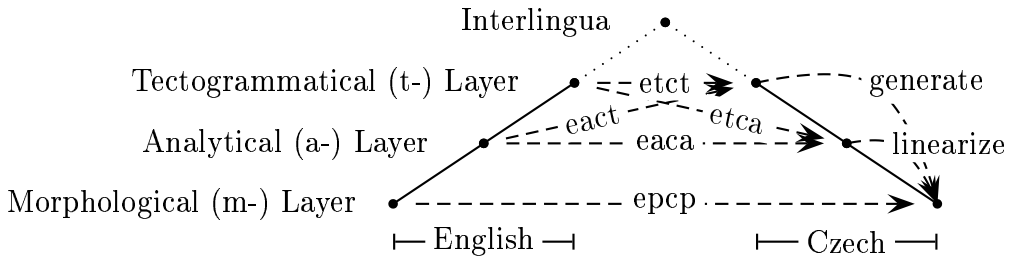


Figure 3: Direct and shallow methods of MT. Abbreviations explained in Section 3.3.

3.1 Synchronous Tree Substitution Grammar

Synchronous Tree Substitution Grammars (STSG, e.g. Čmejrek (2006)) capture the basic assumption of syntax-based MT that a valid translation of an input sentence can be obtained by local structural changes of the input syntactic tree (and translation of node labels) while there exists a derivation process common to both languages. Some training sentences may violate this assumption because human translators do not always produce literal translations but we are free to ignore such sentences in the training.

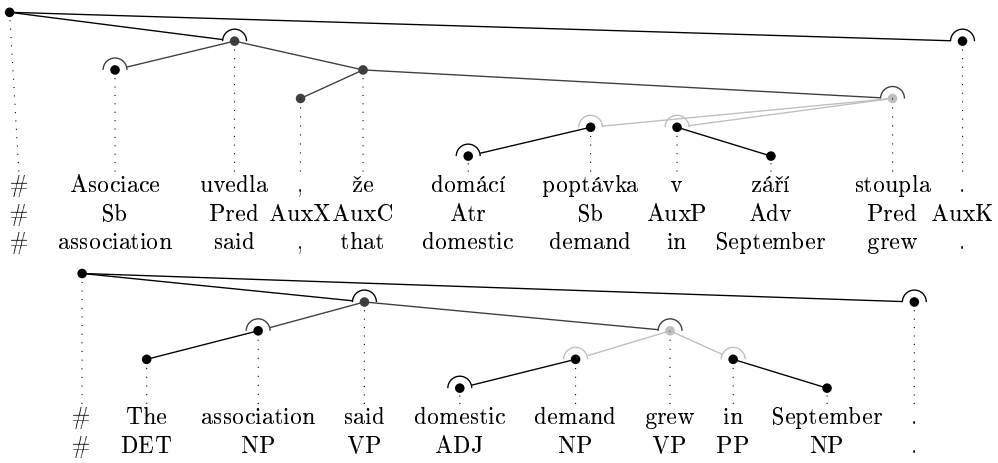


Figure 4: A sample pair of analytical trees synchronously decomposed into treelets. Linguistic annotation is provided for illustration purposes only.

As illustrated in Figure 4, STSG describe the tree transformation process using the basic unit of a **treelet pair** and the basic operation of **tree substitution**. Both source and target trees are decomposed into treelets that fit together. Each treelet

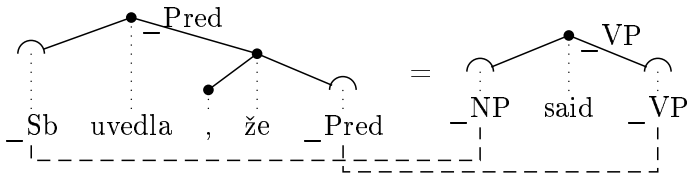


Figure 5: A sample analytical treelet pair.

can be considered as representing the minimum translation unit. A treelet pair such as depicted in Figure 5 represents the structural and lexical changes necessary to transfer local context of a source tree into a target tree.

Each node in a treelet is either **internal** (• , constitutes treelet internal structure and carries a lexical item) or **frontier** (◠, represents an open slot for attaching another treelet). Frontier nodes are labelled with **state labels** (such as "_Sb" or "_NP"), as is the root of each treelet. A treelet can be attached at a frontier node only if its root state matches the state of the frontier. A **treelet pair** also describes the **mapping** of the frontier nodes. A pair of treelets is always attached synchronously at a pair of matching frontier nodes.

3.2 STSG in Machine Translation

Our goal is to translate a source sequence of words s_1 into a target sequence of words $\hat{s_2}$, where $\hat{s_2}$ is the most likely translation out of all possible translations s_2:

$$\hat{s_2} = \underset{s_2}{\operatorname{argmax}}\, p(s_2 \mid s_1) \tag{3}$$

Breaking the search into independent steps of parsing, tree transfer and generation, we search for the most likely **synchronous derivation** $\hat{\delta}$ that constructs the source tree T_1 and the target tree T_2, and we take its target-side projection:

$$\hat{T_2} = \underset{T_2}{\operatorname{argmax}}\, p(T_2 \mid T_1) \doteq \operatorname{target}(\hat{\delta}) = \operatorname{target}\Big(\underset{\delta \in \Delta(T_1)}{\operatorname{argmax}}\, p(\delta)\Big) \tag{4}$$

A derivation δ consists of a sequence of treelet pairs. When searching for $\hat{\delta}$, we consider all decompositions of T_1 into a set of treelets $t_1^0, \ldots, t_1^k$, we expand each treelet t_1^i into a treelet pair $t_{1:2}^i$ using a treelet pair dictionary and evaluate the probability of the synchronous derivation $\delta = \{t_{1:2}^0, \ldots, t_{1:2}^k\}$.

Following Och and Ney (2002), we further extend Eq. 4 into a general log-linear framework that allows us to include various features or **models**:

$$\hat{\delta} = \underset{\delta \in \Delta(T_1)}{\operatorname{argmax}} \exp\Big(\sum_{m=1}^{M} \lambda_m h_m(\delta)\Big) \tag{5}$$

Each of the M models $h_m(\delta)$ provides a different score aimed at predicting how good the derivation δ is. The weighting parameters λ_m, $\sum_1^M \lambda_m = 1$, indicate the relative importance of the various features and they are tuned on an independent dataset. See the thesis for a detailed descriptions of the models we use as well as the training procedure based on a parallel treebank.

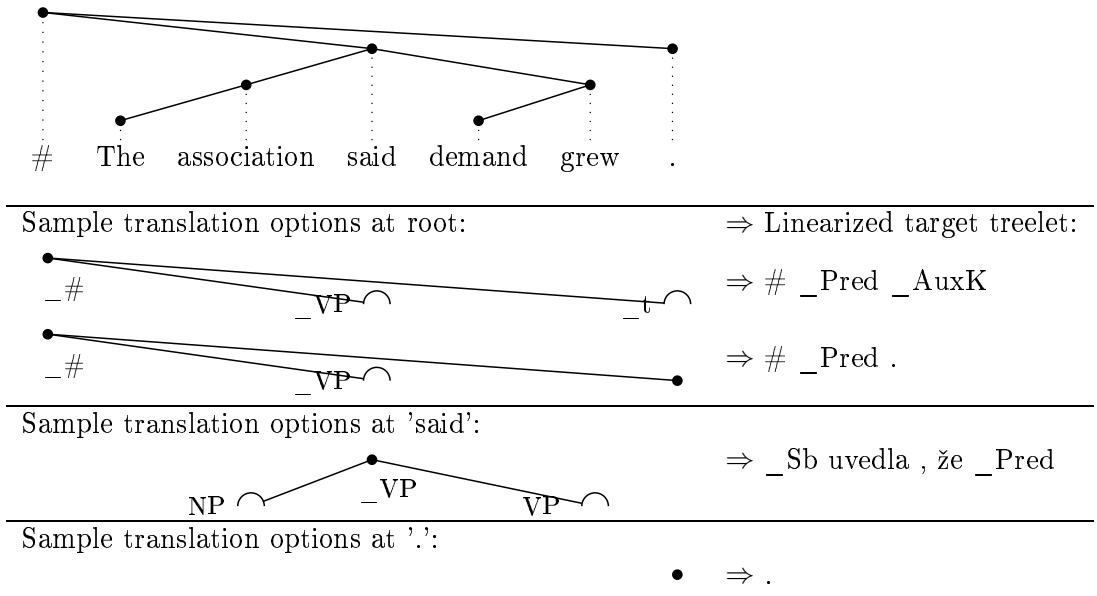


Figure 6: Sample translation options for translating an English a-tree to a Czech a-tree. For conciseness, the target tree is linearized (i.e. structure omitted).

The search space of all possible decompositions of input tree multiplied by all possible translations of source treelets is too large. The current version of our decoder implements a beam search inspired by the strategy of phrase-based decoder Moses (Koehn *et al.*, 2007). While Moses constructs partial hypotheses in a left-to-right fashion (picking source phrases in arbitrary order), our partial hypotheses are constructed top-to-bottom as the source tree T_1 is covered.

The first step is the construction of "translation options". For each input node $x \in T_1$, all possible treelets rooted at x are examined and if a translation of a treelet is found, it is stored as one of the translation options for x. Figure 6 illustrates sample translation options for three input nodes: "#", "said", and ".".

Figure 7 illustrates the second and main step, i.e. the gradual expansion of a hypothesis using translation options constructed in the first step. Once all input nodes are covered (and thus no frontiers are left in the partial output), the output hypothesis is returned. In practice, we beam-search the space of derivations, studying only σ best-scoring partial hypotheses of the same number of covered input nodes.

Methods of Back-off As expected, and also pointed out by Čmejrek (2006), the additional structural information boosts data-sparseness problem. Many source treelets in the test corpus were never seen in our training data. To tackle the problem, our decoder utilizes a sequence of back-off models, i.e. a sequence of several methods of target treelet construction and probability estimation. Each subsequent model is based on less fine-grained description of the input treelet and constructs the target treelet on the fly from independent components (see the thesis for a detailed description). The order and the level of detail of the back-off methods is fixed but easily customizable in a configuration file.

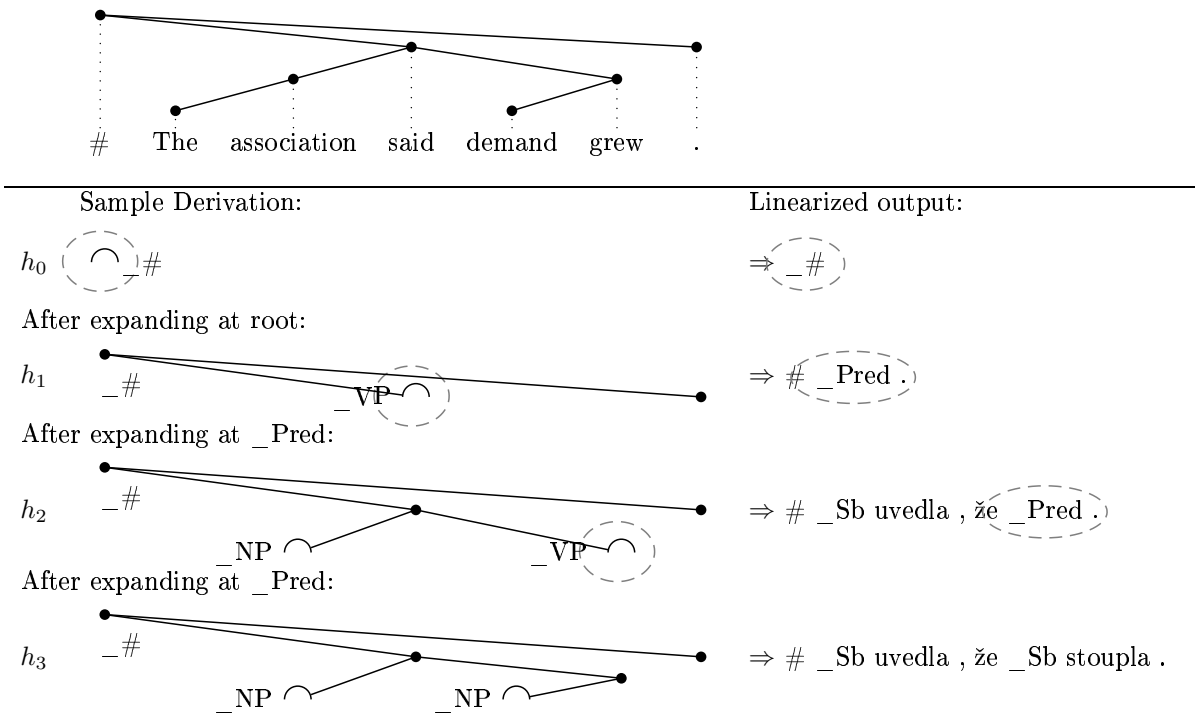


Figure 7: Top-down hypothesis expansion using translation options from Figure 6. Dashed circles indicate where treelet pairs are attached at each step.

3.3 Empirical Evaluation of STSG Translation

In an end-to-end evaluation, we try to cover a wide range of experimental settings when translating from English to Czech, as illustrated in Figure 3.

Our main focus is the translation from the English t-layer to the Czech t-Layer (etct). The general applicability of STSG to any dependency trees allows us to test the same model also for analytical translation (eaca) or across the layers (etca and eact). To a certain extent, our tree-based decoder can simulate a direct approach to MT (phrase-based decoding, as discussed in Chapter 4 below) if we replace the dependency structure of an a-tree with a simple left-to-right chain of words ("linear tree"). The results obtained using this approach are labelled "epcp". Our phrase-based approximation epcp is bound to work worse because we prohibit any phrase reorderings.

Table 3 reports the BLEU scores of several configurations of our system, higher scores suggest better MT quality. The values in the column "LM Used" indicate the type of language model used in the experiment. An n-gram model can be applied to the output sequence of words. For setups where the final sequence of words is constructed using the generation component by Ptáček and Žabokrtský (2006) with no access to a language model, we use at least a binode LM to improve output tree coherence.

3.4 Discussion

At the first sight, our preliminary results support common worries that with a more complex system it is increasingly difficult to obtain good results. However, we are well aware of many limitations of our current experiments.

Method of Transfer	LM Used	BLEU
epcp	n-gram	10.9±0.6
eaca	n-gram	8.8±0.6
epcp	none	8.7±0.6
eaca	none	6.6±0.5
etca	n-gram	6.3±0.6
etct factored, preserving structure	binode	5.6±0.5
etct factored, preserving structure	none	5.3±0.5
eact, no output factors	binode	3.0±0.3
etct, vanilla STSG (no factors), all node attributes	binode	2.6±0.3
etct, vanilla STSG (no factors), all node attributes	none	1.6±0.3
etct, vanilla STSG (no factors), just t-lemmas	none	0.7±0.2

Table 3: English-to-Czech BLEU scores for syntax-based MT evaluated on DevTest dataset of ACL 2007 WMT shared task.

BLEU Favours n-gram LMs. BLEU is known to favour methods employing n-gram based language models. Empirical evidence can be observed in Table 3: an n-gram LM gained 2 BLEU points for both "eaca" and "epcp".

Cumulation of Errors. All components in our setup deliver only the single best candidate. Any errors will therefore accumulate over the whole pipeline. This primarily hurts the "etct" scenario where most tools are employed.

Conflict of Structures. Our current heuristic treelet extraction crucially depends on the quality of both English and Czech trees as well as on the node alignment between them. A single error in any of the rigid sources may prevent the extraction of a treelet pair, not to mention natural divergence between the sentence and its translation. We thus lose a significant portion of training data.

Combinatorial Explosion. In the current implementation, target-side treelets are fully built during the preparatory phase of translation option generation. Uncertainty in the many t-node attributes leads to too many treelets with insignificant variations while e.g. different lexical choices are pushed off the stack. While vital for final sentence generation (see Table 3), fine-grained t-node attributes should be produced only when all key structural, lexical and form decisions have been made.

Sentence Generation Tuned for Manual Trees. The rule-based generation system (Ptáček and Žabokrtský, 2006) was designed to generate Czech sentences from full-featured manual Czech tectogrammatical trees. Our system produces t-trees with many errors and omissions due to the pipeline of automatic annotation tools as well as due to the noise caused by our STSG transfer.

More Free Parameters. Last but not least, the more complex the setup is ("etct" being our most complicated design), the more free parameters have to be configured in the system (the choice of e.g. a parser, a tagger, a method of treelet extraction as well as the many options the components have).

Despite not reflected in the error-bar figures in Table 3, which describe the variance due to randomness in input data, we suggest that the variance or rather the room for improvement due to the sub-component selection and configuration is much greater for more complex scenarios. It is an open software engineering and

management question, which of the free parameters or which of the methods should be further studied.

Another drawback of the complex model is the abundance of model parameters (λ_m in the log-linear model, Section 3.2), so our parameter optimization method fails to converge and we stick to a default: all models equally important.

4 Improving Morphological Coherence in Phrase-Based Machine Translation

The previous chapter was devoted to a study of a deep-syntactic MT system and one of its components, tree-to-tree transfer, in particular. Completely reversing our research priorities, we now tackle the task of MT in a direct end-to-end fashion, employing very little of linguistic analysis.

State-of-the-art empirical results in MT are currently achieved by phrase-based systems for many language pairs. Known limitations of phrase-based MT include worse quality when translating to morphologically rich languages as opposed to translating from them, and worse grammatical coherence of longer sentences. We participated in the 2006 summer engineering workshop at Johns Hopkins University that attempted to tackle these problems by introducing separate **factors** in MT input and/or output to allow explicit modelling of the underlying language structure. The support for factored translation models was incorporated into the Moses open-source MT system (Koehn *et al.*, 2007). Our contribution to the workshop was the design of factors improving English-to-Czech translation.

4.1 Overview of Factored Phrase-Based MT

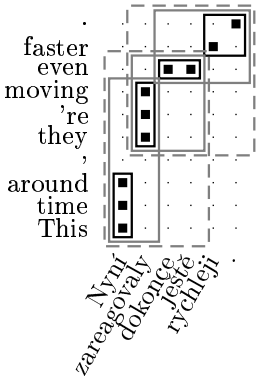
In statistical MT (SMT), the goal is to translate a source (foreign) language sentence $f_1^J = f_1 \dots f_j \dots f_J$ into a target language (Czech) sentence $c_1^I = c_1 \dots c_j \dots c_I$. In **phrase-based SMT** (e.g. Koehn (2004)), the assumption is made that the target sentence can be constructed by segmenting source sentence into K phrases[2], translating each phrase and finally composing the target sentence from phrase translations. See Figure 8 for an example of phrases automatically extracted from a word-aligned sentence pair. We denote the segmentation of the input sentence into K phrases as s_1^K. Among all possible target language sentences, we choose the sentence with the highest probability:

$$\hat{c}_1^{\hat{I}} = \underset{I,c_1^I,K,s_1^K}{\operatorname{argmax}} \{Pr(c_1^I | f_1^J, s_1^K)\} \tag{6}$$

Again, we use a log-linear combination of features: $\sum_m \lambda_m h_m(c_1^I, f_1^J, s_1^K)$.

In **factored phrase-based SMT**, source and target words f and c are represented as tuples of F and C **factors**, resp., each describing a different aspect of the word, e.g. its word form, lemma, morphological tag, role in a verbal frame.

[2]It should be noted that the term "phrases" refers merely to a sequence of words and is not related to linguistically grounded phrases from e.g. Chomskian grammars.



faster	=	rychleji
.	=	.
This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
faster .	=	rychleji .
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	...	...

Figure 8: Sample word alignment and sample phrases consistent with it (not all consistent phrases have been marked).

The process of translation consists of **decoding steps** of two types: **Mapping steps** translate a subset of source factors $S \subseteq \{1 \dots F\}$ into a subset of target factors $T \subseteq \{1 \dots C\}$ using the standard phrase-based model (see e.g. (Koehn, 2004)). **Generation steps** map a subset of target factors T_1 to a disjoint subset of target factors T_2, $T_{1,2} \subset \{1 \dots C\}$. A **translation scenario** is a fixed configuration describing which decoding steps to use in which order. See the thesis for details.

In addition to features for decoding steps, we include arbitrary number of **language models**[3] over subsets of target factors, $T \subseteq \{1 \dots C\}$. We currently use the standard n-gram language model.

While generation steps are used to enforce "vertical" coherence between "hidden properties" of output words, language models are used to enforce sequential coherence of the output.

4.2 Experiments with Factored Phrase-Based MT

Scenarios of Factored English→Czech Translation We experimented with the following factored translation scenarios.

The baseline scenario (labelled T for translation) is single-factored: input (English) lowercase word forms are directly translated to target (Czech) lowercase forms. A 3-gram language model (or more models based on various corpora) checks the stream of output word forms. The baseline scenario thus corresponds to a plain phrase-based SMT system:

English		Czech	
lowercase	⟶	lowercase	+LM
lemma		lemma	
morphology		morphology	

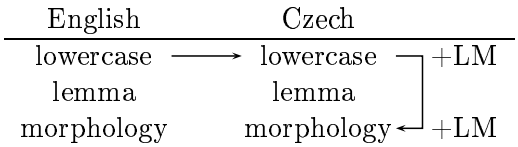
In order to check the output not only for word-level coherence but also for morphological coherence, we add a single generation step: input word forms are first

[3]This might be perceived as a non-standard use of the term, because the models may contain more than just word forms. More generally, these models represent a specific case of a probabilistic sequence model.

translated to output word forms and each output word form then generates its morphological tag.

Two types of language models can be used simultaneously: a (3-gram) LM over word forms and a (7-gram) LM over morphological tags.

We call this the "T+C" (translate and check) scenario:



More complicated scenarios (T+T+C and T+T+G) use additional linguistically-motivated decoding steps.

Table 4 summarizes estimated translation quality of the various scenarios.

	BLEU
T+T+G	13.9±0.7
T+T+C	13.9±0.6
T+C	13.6±0.6
Baseline: T	12.9±0.6

Table 4: BLEU scores of various translation scenarios.

The good news is that multi-factored models always outperform the baseline T. Unfortunately, the more complex multi-factored scenarios do not bring any significant improvement over T+C. Our belief is that this effect is caused by search errors: with multi-factored models, more hypotheses get similar scores and future costs of partial hypotheses might be estimated less reliably. With a limited stack size (not more than 200 hypotheses of the same number of covered input words), the decoder may more often find sub-optimal solutions. Moreover, the more steps are used, the more model weights have to be tuned in the minimum error rate training. Considerably more tuning data might be necessary to tune the weights reliably.

Granularity of Part-of-Speech and Impact of Additional Data In other experiments we focus on a good balance of detail in the morphological tag and data sparseness it creates. In a small data setting, fine-tuning of the morphological tag pays off; if more parallel data are available, full Czech morphological tags perform better.

4.3 Human Evaluation

The best system described in this chapter took part in an open MT evaluation campaign carried out during ACL 2007 Second Workshop on Statistical Machine Translation[4]. Table 5 reproduces the results from Callison-Burch *et al.* (2007) for English to Czech MT quality. The **adequacy** scale describes how well the translation conveys the original meaning, the **fluency** reflects how grammatically correct

[4] `http://www.statmt.org/wmt07/`

System	Adequacy	Fluency	Rank	Constituent
Our T+C (cu)	0.523	0.510	0.405	**0.440**
PC Translator (pct)	**0.542**	**0.541**	**0.499**	0.381
Single-Factored Moses (uedin)	0.449	0.433	0.249	0.258

Table 5: Human judgements of English→Czech MT at ACL WMT 2007.

System	Commentary (in-domain)	News (out-of-domain)
Our T+C (cu-bojar)	**71.4%**	63.4%
PC Translator	66.3%	**71.5%**
TectoMT (cu-tectomt)	48.8%	49.4%
Single-Factored Moses (uedin)	48.6%	50.2%

Table 6: Percentage of sentences where the system was ranked better than or equal to any other system (human judgements, ACL WMT08).

the MT output is and **rank** shows how often would human judges prefer to get output from that particular system compared to other systems. The **constituent rank** is a new scale introduced by Callison-Burch *et al.* (2007) that tries to simplify the task of ranking hypotheses by asking the judges to rank only randomly selected sections of sentences.

Our system improved over the phrase-based baseline (provided by University of Edinburgh, uedin) and got very close to a major English-Czech commercial MT system PC Translator by LangSoft (a rule-based system with a long history of development). Despite the comparison not being completely fair (PC Translator is a generic MT system while our system was trained and evaluated in the known domain of news commentaries), we consider the result very promising.

We also participated with a very similar setup in ACL 2008 WMT shared task[5] (Bojar and Hajič, 2008), the main difference being only larger parallel and monolingual training corpora. As documented in Table 6 (results from Callison-Burch *et al.* (2008)), the additional data allowed us to improve over PC Translator for in-domain setting (Commentary). In the generic domain of News, PC Translator performs better. For an illustration of our MT output, see Appendix A.

5 Concluding Discussion

In the last chapter of the thesis we attempt to provide a larger picture of the relation between linguistic data and applications. We survey available literature asking whether lexicons bring an improvement to various NLP applications. Not surprisingly, there is not a simple and conclusive answer to this question. We name several projects that perform very well without employing complicated lexicons, and we also mention various indisputable uses of lexicons.

[5] `http://www.statmt.org/wmt08/`

We see that lexicons in NLP applications can be successfully avoided if the intelligence is left to the human:

- Grammaticality is ensured by reusing a text produced by humans (sentence fusion in text summarization).
- Selection of the translation equivalent is based on the choice of a human in a similar context (MT).
- Overgeneration never hurts, if the output of the system is intersected with some man-made data (information extraction).

We guess that the reason for the relatively rare use of independently designed (manual or automatic) lexicons in NLP applications is the difficulty of adapting the formats and more importantly the difference in the types of decisions an application has to make and hints a lexicon can offer.

On the other hand, we mention several applications that build their own lexicons (or probabilistic tables), the features of which are very much influenced by linguistic insights incorporated in human lexicons.

Our belief is that linguistic theories provide an indispensable source of inspiration that is being slowly reflected in the design of applications. Any data produced by computational *linguists* remain difficult to reuse in practical NLP systems because they provide answers for questions the system is nowhere near to ask.

5.1 Contribution of the Thesis

The first part of the thesis (Chapter 2) examined automatic ways of constructing a valency dictionary, an important resource for various applications including rule-based or syntax-based MT. Several methods of frame extraction were designed and evaluated using a novel metric that gives a partial credit even for not quite complete frames by estimating the savings in a lexicographer's work.

The second part (Chapters 3 and 4) focused directly on linguistic data within the task of MT. First, we designed, implemented and evaluated a full-fledged syntax-based MT system. The generic engine was applied in various settings ranging from transfer at a deep syntactic layer to an approximation of an uninformed phrase-based translation. The results indicate that the best translation quality is still achieved by the most simple methods; the main reasons for this being the cumulation of errors, the loss in training data due to both natural and random syntactic divergence between Czech and English and finally a combinatorial explosion in the complex search space.

In Chapter 4 we moved to a relatively simple model of phrase-based MT and we improved its accuracy by adding a limited amount of linguistic information. While word lemmas and morphological tags can be successfully exploited by the phrase-based model thanks to their direct correspondence to the sequence of words achieving a better morphological coherence of MT output, the applicability of syntactic information remains an open research question.

The thesis contributes to the art of natural language processing and machine translation in particular by designing and evaluating:

- an automatic metric estimating the savings in a lexicographer's work;
- experiments with various methods for automatic deep valency frame acquisition based on corpus observations;
- a machine translation system with a deep syntactic transfer, including the evaluation of an end-to-end pipeline; the system can also be applied at a surface-syntactic layer;
- improved word-alignment techniques by preprocessing parallel texts, utilized in experiments reported here and fully described in Bojar *et al.* (2006);
- various configurations of factored phrase-based models for English-to-Czech translation improving target-side morphological coherence.

Moreover, we prepared and made the following data available to the research community:

- a Czech-English parallel corpus CzEng, two public releases (Bojar and Žabokrtský, 2006; Bojar *et al.*, 2008),
- manual Czech-English word-alignment data (Bojar and Prokopová, 2006), including an evaluation of inter-annotator agreement,
- Golden VALEVAL, word-sense disambiguation data from the VALEVAL experiment (Bojar *et al.*, 2005),
- a mildly cleaned-up collection of Czech-English translation dictionaries (Bojar and Prokopová, 2007).

Many suggestions on how to further improve or extend our methods were mentioned throughout the thesis. We plan to continue our research by further attempts to combine successful simple models with linguistically-informed methods.

References

For publications co-authored by Ondřej Bojar see the separate listing below.

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.

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.

Martin Čmejrek. *Using Dependency Tree Structure for Czech-English Machine Translation.* PhD thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2006.

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.

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.

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.

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

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

Jan Ptáček and Zdeněk Žabokrtský. Synthesis of Czech Sentences from Tectogrammatical Trees. In *Proc. of TSD*, pages 221–228, 2006.

Jiří Semecký. *Verb Valency Frames Disambiguation*. PhD thesis, Charles University, Prague, 2007.

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

Publications by Ondřej Bojar

Refereed

1. O. Bojar and J. 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.
2. O. Bojar, M. Janíček, Z. Žabokrtský, P. Češka, and P. 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.
3. O. Bojar, S. Cinková, and J. Ptáček. Towards English-to-Czech MT via Tectogrammatical Layer. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, Bergen, Norway, December 2007.
4. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. 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.
5. O. 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.
6. O. Bojar and Z. Žabokrtský. CzEng: Czech-English Parallel Corpus, Release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62, 2006.
7. V. Benešová and O. 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.
8. O. Bojar and M. Prokopová. Czech-English Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1236–1239. ELRA, May 2006.
9. O. Bojar, E. Matusov, and H. Ney. Czech-English Phrase-Based Machine Translation. In *FinTAL 2006*, volume LNAI 4139, pages 214–224, Turku, Finland, August 2006. Springer.
10. O. Bojar, J. Semecký, and V. Benešová. VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics*, 83:5–17, 2005.
11. M. Lopatková, O. Bojar, J. Semecký, V. Benešová, and Z. Žabokrtský. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In V. Matoušek, P. Mautner, and T. Pavelka, editors, *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings*, volume LNAI 3658, pages 99–106. Springer Verlag, September 2005.
12. O. Bojar, C. Brom, M. Hladík, and V. Toman. The Project ENTs: Towards Modelling Human-like Artificial Agents. In P. Vojtáš, M. Bieliková, B. Charron-Bost, and O. Sýkora, editors, *SOFSEM 2005 Communications*, pages 111–122. Society for Computer Science, January 2005.
13. O. Bojar, P. Homola, and V. Kuboň. Problémy recyklování systému automatického překladu. In P. Vojtáš, editor, *ITAT 2005 Information Technologies – Applications and Theory*, pages 335–344, Košice, Slovakia, September 2005. University of P. J. Šafařík.
14. O. Bojar, P. Homola, and V. Kuboň. Problems Of Reusing An Existing MT System. In *IJCNLP 2005 - Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts*, pages 181–186, October 2005.
15. O. Bojar and J. Hajič. Extracting Translation Verb Frames. In W. von Hahn, J. Hutchins, and C. 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.
16. O. Bojar. Budování česko-anglického slovníku pro strojový překlad. In P. Vojtáš, editor, *ITAT 2005 Information Technologies – Applications and Theory*, pages 201–211, Košice, Slovakia, September 2005. University of P. J. Šafařík.
17. O. Bojar, P. Homola, and V. Kuboň. An MT System Recycled. In *Proceedings of MT Summit X*, pages 380–387, September 2005.
18. O. 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.
19. O. Bojar. Czech Syntactic Analysis Constraint-Based, XDG: One Possible Start. *Prague Bulletin of Mathematical Linguistics*, 81:43–54, 2004.
20. O. Bojar. Automated Extraction of Lexico-Syntactic Information. In J. Šafránková, editor, *WDS'04 Proceedings of Contributed Papers: Part I - Mathematics and Computer Sciences*, pages 211–217, Prague, June 15–18 2004. Charles University, Matfyzpress.
21. O. Bojar. Towards Automatic Extraction of Verb Frames. *Prague Bulletin of Mathematical Linguistics*, 79–80:101–120, 2003.
22. O. Bojar. Building Subcorpora Suitable for Extraction of Lexico-Syntactic Information. In *Proceedings of the Student Session, ESSLLI*, August 2003.

Other

1. O. Bojar and A. Lopez. Tree-based Translation. Handout for MT Marathon Tutorial, May 2008.
2. O. Bojar and M. Prokopová. Czech-English Machine Translation Dictionary. Technical report, ÚFAL MFF UK, Prague, Czech Republic, April 2007.
3. O. Bojar. Strojový překlad: zamyšlení nad účelností hloubkových jazykových analýz. In *MIS 2006*, pages 3–13, Josefův Důl, Czech Republic, January 2006. MATFYZ-PRESS.
4. P. Koehn, M. Federico, W. Shen, N. Bertoldi, O. Bojar, C. Callison-Burch, B. Cowan, C. Dyer, H. Hoang, R. Zens, A. Constantin, C. Moran, and E. Herbst. Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding. Technical report, Johns Hopkins University, Center for Speech and Language Processing, 2006.
5. O. Bojar, J. Semecký, S. Vasishth, and I. Kruijff-Korbayová. Processing noncanonical word order in Czech. In *Proceedings of Architectures and Mechanisms for Language Processing, AMLaP 2004*, pages 91–91, Université de Provence, September 16-18 2004.
6. O. Bojar, C. Brom, M. Hladík, M. Vejlupek, V. Toman, and D. Voňka. ENTI – Simulátor přirozeného prostředí lidského světa. In *MIS 2003*, pages 3–14. MATFYZPRESS, January 18–25, 2003 2003.
7. O. Bojar. AX - Systém pro automatizovanou extrakci lexikálně-syntaktických údajů. In *MIS 2003*, pages 15–24. MATFYZPRESS, January 18–25 2003.

A Sample Translation Output

Source text, WMT 08 News Test

New Russia-Ukraine gas row fears

A fresh gas price dispute is brewing between Ukraine and Russia, raising the risk that Russian exports of the fuel to western Europe may be affected. Most of Russia's gas exports to the European Union (EU) are piped through Ukraine and any row between the two nations is keenly watched. Kiev has warned that if Moscow raises the price it has to pay for the gas it will charge Russia higher transit fees. A previous dispute between the two last year reduced supplies to EU states.

Moses T+C, LM from SYN2006 BLEU 11.93%

Nové Rusko - Ukrajina plynu obává řádek.

A čerstvé ceny plynu bublají spor mezi Ukrajinou a Ruskem, zvýší riziko, že ruský vývoz paliva do západní Evropy, může být ovlivněn. Většina ruských vývozů plynu do evropské unie (EU) jsou píšťala přes Ukrajinu a každý řádek mezi oběma národy je naléhavě střežen. Kyjev již varoval, že pokud Moskva zvyšuje cenu, která se má platit za plyn bude účtovat vyšší tranzitní poplatky v Rusku. A předchozí spor mezi dvěma v loňském roce snížené dodávky pro státy EU.