The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009

EDITORIAL BOARD

Published twice a year by Charles University in Prague

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: pbml@ufal.mff.cuni.cz

ISSN 0032-6585

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009

CONTENTS

Articles

Reviews

Notes

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 5–19

Event Structure in Russian: Semantic Roles, Aspect, Causation

Elena Paducheva

1. Decompositional semantic representations

More than three decades ago the idea of DECOMPOSITIONAL SEMANTIC REPRESENTATION (DSR) of a word was put forward (by Ch. Fillmore, Ju. Apresjan, A.Wierzbicka, J.McCawley, G.Lakoff, R.Jackendoff e.a.). The language under analysis in this paper is Russian but the problems are, to a great extent, independent of language. An example of semantic decomposition from Apresjan 1974, p. 108:

> A *dogonjaet* B (A *catches up* B) =
> 'A and B move in one direction, A is behind B, the distance between A and B diminishes'.

A bit later GRAMMATICALLY ORIENTED DSRs came into being, aiming at explaining morphosyntactic behavior of a word – structures uniting information about TAXONOMY, SEMANTIC ROLES, ASPECT and CAUSATION (Dowty 1979, Wierzbicka 1980). "Since verbs individuate and name events <…>, theories of predicate decomposition are often taken to be theories of the basic EVENT TYPES." (Levin, Rappaport Hovav 2005: 70).

An example from Fillmore 1970 – why *hit* and *break* behave differently:

(1) a. The boy *broke* the window with a ball; b. The boy *hit* the window with a ball.
(2) a. The window *broke*; b. *The window *hit*.

The answer is that *break* is a change of state verb, while *hit* belongs to a class of verbs involving contact: *hit* and *break* are verbs of different VERB CLASSES.

Two different semantic classifications of verbs are widely known.

1. There are traditional lexical classes – let's call them THEMATIC classes (see Wierzbicka 1987 on English speech act verbs; Levin 1993 on English verbs; about Russian verbs see Babenko 2001, Švedova 2007). Thematic classification distinguishes: verbs of MOVEMENT, EXISTENCE, PHYSICAL IMPACT, TREATMENT, CREATION, PERCEPTION,

Please cite this article as: Elena Paducheva, Event Structure in Russian: Semantic Roles, Aspect, Causation. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 5–19.

COGNITION, SPEECH, EMOTION, VOLITION, POSSESSION, PHYSIOLOGY; verbs of SOUND, etc.

2. On the other hand, there are Vendler's ASPECTUAL classes (STATES, ACTIVITIES, ACCOMPLISHMENTS, ACHIEVEMENTS), see Vendler 1967, Dowty 1979, Wierzbicka 1980, Jackendoff 1991, Paducheva 1996, Filip 1999 and many others. Vendler's classes have grammatical relevance; so it stands to reason to call them (taxonomic) CATEGORIES (T-CATEGORIES).

Thematic and category classifications are independent of one another.

In Dowty 1979 and many other postvendlerian classifications accomplishments and achievements are split into agentives and non-agentives. Only then do we arrive at an important category ACTION, missing among Vendler's classes: agentive accomplishments and agentive achievements are called ACTIONS (we have *napisat'* <*pis'mo*> 'write a letter', *vyigrat'* <*gonki*> 'win <the race>', etc.). Non-agentive achievements (*prostudit'sja* 'catch cold') are called HAPPENINGS; non-agentive accomplishments (*rastajat'* 'thaw') are called TELIC PROCESSES. Non-agentive activities (*kipet'* 'boil') are called NON-TELIC PROCESSES.

Agentivity has direct aspectual correlations. Cf. the verb *okružat'* 'surround' – when agentive, it is an accomplishment, when non-agentive, it is a state:

(3) a. Mal'čik pokazyvaet belogvardejcam fokusy, i, poka te smotrjat ego vystuplenie, krasnye *okružajut* stanciju i potom zanimajut ee. 'The boy presents tricks to the white guardians, and while they are watching the performance the reds *surround* the station and then occupy it' (example from National Corpus of Russian, `http://www.ruscorpora.ru`).

b. Daču *okružajut* lesa 'Forests *surround* the dacha'.

The role of the T-category in lexical semantics is similar to that of part of speech in grammar.

Meaning is flexible and context dependent; REGULAR POLYSEMY (Apresjan 1974) is widespread. Thus, not only MEANING but also MEANING CHANGE must be accounted for with the help of DSRs.

2. «Lexicographer» – a semantic database of Russian verbs and a theory of event structure

I'll speak about decompositional semantic representations contained in the Database of Russian verbs «Lexicographer»: `http://www.rusling.narod.ru` (see Kustova, Paducheva 1994, Kustova 2004, Paducheva 2004); main researchers – Galina Kustova, Elena Paducheva, Raisa Rozina, Elena Xasina. The database is conceived as a realization of a certain THEORY OF EVENT STRUCTURE.

The lexical entry in the DB «Lexicographer» is exemplified by the lexeme VYTERET' 1.2 'wipe' (the term LEXEME is here used to mean a word taken in one of its meanings, as in Mel'čuk 1974, Apresjan 1974).

The lexical entry of a verb in the database is divided into several domains. The domains are: Argument structure, T-Category, Decomposition, Thematic class, Aspect, Legend.

Let's begin with the ARGUMENT STRUCTURE of VYTERET' 1.2, see Table 1.

VYTERET' 1.2
'wipe dry <the dishes, one's hands>': *X vyter Y* (*Z-om*) '*X* wiped *Y* (with *Z*)'

Variable	Morphosyntax	Rank	Semantic role	Thematic class
X	Subject	Center	Agent	Person
Y	Object	Center	Patient	physical entity: with a surface
(Z)	Instrumental	Periphery	Instrument	physical entity
W	—-	Off Screen	Theme	liquid / substance

Table 1. Argument structure for vyteret' *1.2.*

A verb describes an event. Each participant of the event is represented by a VARIABLE – a Latin letter, which functions as a Name: a participant is called this name in the Decomposition. This is the 1st column. The second column – MORPHOSYNTACTIC REALIZATION, i.e. syntactic POSITION of the participant (Subject, i.e. Nominative case; Object, i.e. Accusative; Other cases; prepositional phrases – PPs). The third column is called COMMUNICATIVE RANK (Croft 1991, Testelec 2001: 420). Three ranks are distinguished: **Center** (for participants occupying syntactic positions of Subject and Object); **Periphery** (for Instrumental case and Prepositional Phrases); and **Off Screen**. This last rank is ascribed to a participant that is not projected to the surface – as is the case with the participant W in the Argument structure of *vyteret'* 1.2. (Participant W shows itself in the lexeme *vyteret'* 1.1, which will appear later). The 4th column – **Semantic role** (Agent, Patient, Theme, etc.) The 5th column – **Thematic class** (person, physical object, body part, etc.; additional semantic specifications can be added, such as, e.g., "sharp edge" for the participant Instrument in the lexical entry for the verb *cut*).

NB the notion of diathesis: DIATHESIS is a correspondence between roles and their morphosyntactic realizations, see Mel'čuk, Xolodovič 1970. Causative alternation, for example, is a change of diathesis. Basically, diathesis is a role-POSITION and a role-rank correspondence. Participant W without morphosyntax (see Table 1) is a kind of riddle – this riddle will be solved when we come down to the lexeme *vyteret'* 1.1 and address diatheses.

T-CATEGORY has already been spoken about. The central domain in the lexical entry is DECOMPOSITION. Decomposition of a verb in the DB «Lexicographer» does not purport to be an exhaustive description of its lexical meaning. It is a SCHEMATIC decomposition: it represents exhaustively only GRAMMATICALLY RELEVANT (or, somewhat broader, STRUCTURALLY RELEVANT) aspects of the verb's meaning.

Decomposition is given not for a word but for a lexeme. The verb *vyteret'* 'wipe' has three lexemes: *vyteret'* 1.2 (about the dishes), *vyteret'* 1.1 (about the dust) and *vyteret'* 2 (about clothes on knees and elbows).

Lexicographer type semantic decomposition (LSD) of a lexeme is a sequence of syntactically independent semantic components: each component is, basically, a predication. Decomposition is a kind of scenario describing the event in question.

Components are divided into CATEGORIAL and THEMATIC.

See an example of Lexicographer type semantic decomposition in Table 2.

VYTERET' 1.2
'wipe dry (the dishes /one's hands)': *X wiped Y* =

K0	**Initial state** | before t < MS *Y* was in a state: *Y had W on its surface*
K1	**ipso facto** *the state of Y was not normal*
K2	–
K3	–
K4	**Activity** | at t < MS X acted with the Goal in mind
K5	**Manner of action** | *X acted upon Y*; **ipso facto** *upon W* (: with the help of Z)
K6	**Causation** | K4 was causing K7
K7	**Process in Object** | simultaneous with activity; has limit: *W was being removed from the surface of Y*
K8	**Result** | new state of Y came about & holds at the MS: *Y has no W on its surface*
K9	**Entailment** | *the state of Y is normal*
K10	**Implication** | *there is no W on the surface of Y;* **ipso facto** *W does not exist*

Table 2. Decomposition of vyteret' *1.2.*

Abbreviations and comments. MS – moment of speech (in the context of an utterance MS can be replaced by some other moment of reference). *Result* (of the activity of the Agent) is a state that corresponds to the Goal of the Agent, once it is reached. (So *Goal* need not be explicated – it coincides with the Result.) Result may correspond to the final state (= LIMIT) of a telic process *in* the Object (or *with* the Object; namely, a process which the Object participates in).

The domain LEGEND shows how different lexemes of a word are related to one another. Each lexical entry begins with EXAMPLES and ends with a COMMENTARY.

3. Event structure: taxonomy and semantic roles

3.1. Categories

Decompositions obey a certain FORMAT – different verb classes have different decomposition formats (DFs): all verbs of the same category have the same DF.

Verbs of Action are characterized by the following configuration of components:

(1) K4. **Activity** | X acted with the Goal in mind
K6. **Causation** | this caused
K8. **Result** | new state came about & holds at the MS.

This configuration is present in the decomposition of such verbs as *vyteret'* 'wipe', *razrezat'* 'cut <the water melon>', *vystirat'* 'wash', *postroit'* 'build', *pokrasit'* 'paint <the roof>', *svarit'* 'boil <an egg>', *vykopat'* 'dig out' etc.

There are different kinds of actions. Their decomposition formats differ from one another. But configuration (1) is present in all formats for actions.

3.2. Thematic classes

Category components constitute the CATEGORY FRAME of the decomposition. Thematic components are inserted in different places of the category frame. If we replace, e.g., the concrete state *sleep* – by its natural hyperonym PHYSIOLOGICAL STATE we are able to identify *razbudit'* as a verb belonging to the thematic class PHYSIOLOGY verbs. For *vyteret'* 1.2 'wipe' its thematic class TREATMENT is substantiated by the following configuration:

(2) K0. **Initial state** | the (functional) state of Y was not normal /desirable
K8. **Result** | the (functional) state of Y is normal /desirable.

Other verbs of treatment – *žarit'* 'stew', *varit'* 'boil', *gladit'*, 'iron'. Decompositions provide a semantic basis both for category and thematic classification of verbs.

3.3. Meaning shifts

– how can they be presented as operations on LSDs.

3.3.1. Deagentivization, a CATEGORY SHIFT

(3) a. Ivan *razbudil* menja grubym pinkom [*razbudil* 'woke up' – **action**]
Ivan$_{NOM}$ *wake*$_{PAST}$ me$_{ACC}$ rude$_{INSTR}$ kick$_{INSTR}$
'Ivan *woke* me up with a rude kick.'

b. Zvonok v dver' *razbudil* menja [*razbudil* 'woke up' – **happening**]
ringing$_{NOM}$ in door *wake*$_{PAST}$ me$_{ACC}$
'The ringing of the doorbell *woke* me up.'

Templates (#3a) and (#3b) below present two abbreviated LSDs of the verb *razbudit'* (corresponding to its different lexemes; T-category of the lexeme and thematic classes of the participants are given in brackets; components in parenthesis are optional).

(#3a) X *razbudil* Y [action : ordinary] =
K0. **Initial state** | before t < MS Y was in a state: *Y slept*
K4. **Activity** | at t < MS X acted with the Goal in mind [X is a person]
K5. (**Manner of action** | acted upon Y: applying Z)
K6. **Causation** | this was causing [causation as a process] / caused [causation as event]
K7. (**Process in Object** | synchronous; telic)
K8. **Result** | new state of Y came about & holds at the MS: *Y does not sleep*
K9, K10. **Entailment, Implication** |—

(#3b) X *razbudil* Y [happening] =
K0. **Initial state** | before t < MS Y was in a state: *Y slept*
K4. **Causer** | X took place [X is an event]
K5. (**Manner of action** |—)
K6. **Causation** | this caused [causation as event]
K8. **Effect** | new state of Y came about & holds at the MS: *Y does not sleep*
K9. **Entailment** |—
K10. **Implication** | this is bad for Y

The difference between action and happening lexemes consists in that:

1. In the template of a causative verb of action the Causer (see component K4) is the activity of the goal-setting Agent: 'X [person] acted with the Goal in mind', so component K8 is called "Result"; while in the template of a verb of happening the Causer is an event: 'X [event] took place' and what is caused is the effect.
2. Component Manner of action, though optional, is present in the semantics of *razbudit'*-action. In the template of a happening the parameter Manner of action loses its sense.

Optionality of the Manner of action component in the semantics of the agentive *razbudit'* (as well as *otkryt'* 'open', *razbit'* 'break', *razrušit'* 'destroy') is responsible for the easiness with which these verbs acquire happening interpretation: happening is an event type with no volitional agent. Not so with *vyteret'* 'wipe': *wipe* has Manner of action as an obligatory component. Or take the verb *razrezat'* 'cut': cutting presupposes the use of an instrument with a sharp edge, specific movements on the part of the Agent and, thus, a volitional Agent.

In Levin, Rappaport Hovav 1995: 103 the opposition is introduced of VERBS OF MANNER <of action> (such as *lock, cut, sweep*) and VERBS OF RESULT (such as *close, break*, which specify only the resulting state). Verbs of manner (of action) specify the activity of the Agent; the Agent's intentions and evaluations, instruments s/he uses, etc. They do not deagentivize.

There is another type of non-agentive subject of a causative verb. This subject appears in the context of the event type called "Happening with the subject of responsibility":

(4) Vanja razbil maminu čašku <nečajanno> 'Vanja broke mummy's cup <inadvertently>'.

The Causer is not the subject X but something that happened to X **not because he wanted it**. The Causer is non-specified. Decomposition format for *razbit'* 'break <unvoluntary>':

(#4) *X razbil Y* [happening with the subject of responsibility] =
K0. **Initial state** | before t < MS Y was in a state*: Y functioned in a normal way*
K1. **Exposition** | *X was doing something in the vicinity of Y*
K4. **Causer** | something happened to X (: *X acquired or lost contact with Y*)
K6. **Causation** | this caused [causation as event]
K8. **Effect** | new state came about & holds at the MS: *Y is broken / doesn't function normally*
K9. **Entailment** |—
K10. **Implication** | X caused damage; X bears responsibility for the damage

Happenings tend to have negative consequences. If it is something that happened to a person this person is responsible for the damage. Note that implications are cancelable.

Such verbs as *prolit'* 'spill', *porvat'* 'tear', *rassypat'* 'scatter', *peregret'* 'overheat' have the same format as *razbit'* 'break <unvoluntary>'.

3.3.2. Combined CATEGORY and DIATHETIC SHIFT

(5) a. *zapolnit'* 1.1: X *zapolnil* Y Z-*om* 'X filled Y with Z' [action] –
Ja *zapolnil* kotel vodoj 'I *filled* the boiler with water'; Mat' *zapolnila* škafy saxarom, mukoj i drugim prodovol'stviem 'Mother *filled* the shelves with sugar, flour and other stuff.
b. *zapolnit'* 1.2: Z *zapolnil* Y 'Z filled Y' [process] –
Voda *zapolnila* kotel 'Water *filled* the boiler'. Bezobraznye natjurmorty *zapolnili* inter'ery naspex postroennyx kvartir 'Ghastly still-lifes *filled* the interiors of quickly built apartments'.

Compare argument structures of *zapolnit'* 1.1 and *zapolnit'* 1.2.

Variable	Morphosyntax	Rank	Semantic role	Thematic class
X	Subject	Center	Agent	Person
Y	Object	Center	Location	container/physical object: has volume
Z	Instrumental case	Periphery	Theme	Mass

Table 3. Argument structure of zapolnit' *1.1 'X filled Y with Z'*

Variable	Morphosyntax	Rank	Semantic role	Thematic class
Z	Subject	Center	Theme	Mass
Y	Object	Center	Location	container/physical object: has volume

Table 4. Argument structure of zapolnit' *1.2 'Z filled Y'*

Two changes take place: 1) change of diathesis (Agent X goes Off screen and the Theme Z occupies the Subject position – in the Center); 2) a category shift: from action to process.

3.3.3. Combined DIATHETIC and THEMATIC SHIFT (a verb changes diathesis & thematic class)

(6) a. *vyteret'* pot so lba 'wipe sweat from the forhead' [*vyteret'* 1.1, REMOVAL; ANNIHILATION];
b. *vyteret'* posudu 'wipe the dishes' [*vyteret'* 1.2, thematic class – TREATMENT].

In the template of *vyteret'* 1.1, see Table 5, the participant W occupies the position of the Object, its semantic role is Theme, and the thematic class of *vyteret'* 1.1 is REMOVAL. Lexeme *vyteret'* 1.2 (see Table 6 = Table 1) is a derivate of *vyteret'* 1.1 (the derivation consists in the change of diathesis); the Object position is occupied by the participant Y, Location-Patient, participant W is Off stage, and the thematic class of *vyteret'* 1.2 is TREATMENT. This is how the change of diathesis results in a change of the thematic class.

(a) *vyteret' sljozy* 'wipe tears' (wipe 1.1) [REMOVAL; ANNIHILATION]

Variable	**Morphosyntax**	**Rank**	**Semantic role**	**Thematic class**
X	Subject	Center	Agent	Person
W	Object	Center	Theme	liquid / substance:
Y	s + Gen	Periphery	Location	physical entity: with surface
(Z)	Instrumental	Periphery	Instrument	physical entity

Table 5. Argument structure of vyteret' *1.1.*

(b) *vyteret' posudu* 'wipe the dishes' (wipe 1.2) [TREATMENT]

Variable	**Morphosyntax**	**Rank**	**Semantic role**	**Thematic class**
X	Subject	Center	Agent	Person
Y	Object	Center	Location-Patient	physical entity: with surface
(Z)	Instrumental	Periphery	Instrument	physical entity
W	—-	Off Screen	Theme	liquid / substance

Table 6. (= Table 1). Argument structure of vyteret' *1.2.*

This demonstrates the role of the parameter rank in the LSD. Object position expresses "aboutness": *wipe* 1.1 is ABOUT participant W, which is annihilated; so the thematic class of *wipe* 1.1 is ANNIHILATION ; *wipe* 1.2 is ABOUT participant Y (dishes), which changes its functional state, and the thematic class of *wipe* 1.2 is TREATMENT.

A COMMENTARY is needed here – W exists only while it is on Y; this fact explains annihilation component in the semantics of *wipe*: annihilation is a consequence of removal.

The same mechanism is responsible for the ambiguity of the verb *vymesti* 'sweep':

(7) a. *vymesti* dvor 'sweep up the yard' [*vymesti* 1.2, thematic class – TREATMENT];
b. *vymesti* musor 'sweep up litter' [*vymesti* 1.1, thematic class – REMOVAL];

The shift in example (7) is a kind of METONYMY: you may pay attention either to the yard (in the prominent Object position) or to sweepings in the yard. The same with the verb meaning 'wipe' in example (6) and many others verbs (cf. *ispravit'* 'correct'; *correct a document* [TREATMENT]; *correct a mistake* [ANNIHILATION], see Apresjan 1974: 206).

A similar relationship between diathesis and thematic class in the example from Fillmore 1977 about loading the truck with hay: in *load the hay* the thematic class of the verb *load* is MOVEMENT (of hay); in *load the truck* it is CHANGE OF STATE (of the truck). Thematic class of the verb depends on what participant occupies the position of the Object, i.e. is in the Center.

4. Event structure: aspect

It is a challenge for «Lexicographer» to predict, on semantic grounds, i.e. within the LSD, whether an agentive verb will behave as an accomplishment or achievement.

Accomplishments can undergo processualization – in the following sense. A derived Imperfective (Ipfv) of an accomplishment is also an accomplishment – but viewed in a SYNCHRONOUS PERSPECTIVE. Accomplishments describe a situation that has an internal limit in its development, and the limit is approached successively, step by step. This point can be illustrated by the following test.

(1) a. otkryval-otkryval [Ipfv], i otkryl [Pfv] [accomplishment];
b. *zamečal-zamečal [Ipfv], i zametil [Pfv] [achievement].

Usually, if both Manner of action component and the component «Process in the Object: simultaneous with the action of the Subject» are present in the LSD, then the event described by a verb can be looked upon from two perspectives, see the decomposition of *vyteret'* 1.2, Table 2: specified manner of action and simultaneity of the Subject's activity with the Process in the Object guarantees the progressive meaning of the derived imperfective of *vyteret'* 1.2.

A derived Ipfv of an achievement is either a perfective state, see example (2), or a tendency, see example (3) (note the absence of Manner of action specification):

(2) Ja *ponjal* 'I've understood' – Ja *ponimaju* 'I understand' [perfective state].
(3) John *vyigral* 'John won' – John *vyigryvaet* = 'most probably, John will win' [tendency].

On the other hand, there are several different semantic sources of instantaneousness (Paducheva 2004: 477–480), e.g., component 'Process in the Object: non-simultaneous with the activity'.

Take the verb *brosit'* 'throw', which lexicalizes causation of movement *by an initial impulse*: the activity of the Agent gives rise to a process that takes place when the activity is already over; this is so called BALLISTIC MOVEMENT (Wierzbicka 1988: 365, Rappaport Hovav 2008). Similar temporary delay of the Process in the object characterizes such events as *vzorvat'* 'explode', *otravit'* 'poison', *ubit'* 'kill'.

5. Event structure: causation

The last facet of event structure is causation. Table 2 seems to imply that causation is an indispensable component in semantic decompositions. Now what about decausativization? Sentence (1b) is said to be the result of decausativization (causative alternation) of (1a):

(1) a. Vanja *razbil* okno
Vanja$_{NOM}$ break$_{PAST}$ window$_{ACC}$
'Vanja broke the window'
b. Okno *razbilos'*
window$_{NOM}$ break$_{SJA.PAST}$
'The window broke'

See Haspelmath 1993, Levin, Rappaport Hovav 1995. Semantically, decausativization in Russian and English is very similar. Syntactically, decausativization in English is a semantic derivation, while in Russian decausative is one of many possible interpretations of the *sja*-form of a verb.

I take it for granted that in Russian derived decausatives exist only for those verbs that are either non-agentive in their primary use (such as *utomit'*, *rasstroit'*) or can undergo *deagentivization* (such as *razbudit'*, *razbit'*), see examples (3), (4) in section 3.

I argue that decausativization resembles passivization: the subject leaves its position in the Center and moves to the Periphery – wherefrom it can afterwards be deleted. For example.

(4) a. Bystraja ezda *utomila* moju lošad' 'fast ride *tired* my horse';
b. Moja lošad' *utomilas'* ot bystroj ezdy 'my horse *got tired* of fast ride'.

(#4.1) Y *utomil* X-a 'Y tired X' =
K0. **Initial state** | before t < MS X was in a state: *normal*
K4. **Causer** | at t event Y took place
K6. **Causation** | this caused
K8. **Effect** | (new state of X came about &) holds at the MS: *X is tired*
K8,9. **Entailment & Implication** |—

(#4.2) X *utomilsja* (*ot* Y-*a*) = 'X became tired (because of Y)'
K0. **Initial state** | before t < MS X was in a state: *normal*
K1. **Periphery causer** | at t event Y took place
K2. **Background causation** | this caused
K4. **New state** | new state of X came about & holds at the MS: *X is tired*
K9. **Entailment** |—
K10. **Implication** | Causer is not relevant

Transition from template (#4.1) to (#4.2) represents decausativization as a change of diathesis. In a diathetic shift participants change their syntactic positions and, consequently, COMMUNICATIVE RANKS.

In (#4.1), with a causative verb *utomit'*, the Causer occupies the position of the grammatical Subject – the first line K4 of the zone Center. In (#4.2) the Causer becomes a peripheral participant – so the two components – Causer and Causation – move from the Center to the Background. Thus, in (#4.2) the first line in the Center, K4 belongs to the participant Theme, which has now acquired the highest rank – Subject.

The Periphery causer and Background causation component are **optional**: they are included in the LSD of a verb in the context **of a sentence** on the condition that the syntactic position of the Periphery causer is filled by a PP. If there is no background Causer in the sentence – then there are no causal components in the meaning of the decausative. In fact, a non-obligatory participant cannot be Off-screen. In the presence of the Periphery causer the Implication is blocked.

Thus, «Lexicographer» can provide a derived verb of happening with a decomposition lacking causative component. Non-derived event types with no causation component also exist. They are represented by such verbs as *pojavit'sja* 'appear', *isčeznut'* 'disappear'.

6. Conclusion

The DB «Lexicographer» has proved to be a source of event structure representations containing information about thematic class, argument structure, aspect and causation. It is a source of explanations, predictions and generalizations (such as compatibility and non-compatibility with time adverbials). At the same time, LSDs can be used for description of meaning shifts of different kind. Here are my main points.

1. Format of definition can be looked upon as an approach to formalization of the notion of taxonomic category, or aspectual class. Thus, LSD predicts the category. Thematic class of a verb was demonstrated to be deducible from its LSD and dependent on the verb's diathesis in a predictable way.
2. One remark about semantic-syntactic interface. The main point in Levin, Rappaport Hovav 2005 is that morphosyntax of participants (argument realization) is deducible from semantic decomposition. As for the set of semantic roles of a verb, it IS determined by its semantic decomposition, while perspective, i.e. distribution of communicative ranks among participants, seems, at least to a certain degree, to be independent of semantic role. Communicative ranks seem to provide independent input information for the rules that determine argument realization. All the attempts to construct hierarchy of semantic roles that would determine their morphosyntactic realization (nine different hierarchies are enumerated in Liutikova e.a. 2006) have failed so far. It seems to be the case that, at least in some cases information about ranks should be the input of the rules of morphosyntax. Take, for example the verb *kišet'* 'swarm', which has

two diatheses: Location at the Periphery, which is its due place (as in *Besschetnoe kolichestvo zver'ja kišit v lesax i dolinax*) and Location in the Subject position (as in *Strana opjat' kišit špionami*). The second is seven times more widespread and is to be recognized as the basic one.

3. There are several parameters that characterize the meaning of a verb: Category, Thematic class, Argument structure, or Diathesis. It turns out that these very parameters undergo change in the course of semantic derivation. In many cases meaning difference between lexemes can be looked upon as a difference in the value of these parameters. Example with the verb meaning 'wipe' (lexemes *vyteret'* 1.1 and 1.2) demonstrates change of Diathesis and Thematic class (TREATMENT *vs.* REMOVAL); in Fillmore's example with hay loading – MOVEMENT *vs.* CHANGE OF STATE.
   Example with the lexemes of the verb *zapolnit'* 'fill' demonstrates change of Category (lexeme *zapolnit'* 1.1, action and *zapolnit'* 1.2, process) and change of diathesis (*Ja zapolnil kotel vodoj – Voda zapolnila kotel*), while their thematic class remains unchanged – CONTACT WITH THE SURFACE.
4. Several types of causation are to be distinguished: foreground causation (as a process and as an event) and background causation. A separate case is pseudo-causation: IPSO FACTO, i.e. entailment. The verb *zapolnit'* 1.2 'fill', process, demonstrates an event structure described with the help of a causative verb but with causation missing. There are two processes that constitute the event of filling Y with Z. One is the process in Z – it moves; another is the process in Y – it becomes filled with Z. The second process is not caused by the first (as is the case with ordinary actions): these two processes are just different ways of looking at one and the same event (situation). In «Lexicographer» this kind of relationship is described by means of a connector IPSO FACTO. This is a kind of entailment relation, but an entailment relation "at the heart" of decomposition. So it deserves special attention. Movement is more essential for what is going on, but it is not movement that measures the event (and licenses the form of Pfv) but the volume of the boiler. In «Lexicographer» pseudo-causation is used in description of rank shifts.[1]

[1]I am grateful to Barbara Partee, Galina Kustova and two anonymous reviewers for comments and suggestions.

References

Apresjan Ju.D. 1974. Leksičeskaja semantika. Moskva: Nauka.

Babenko L.G. 1999. Tolkovyj slovar' russkix glagolov. Moskva: AST-PRESS.

Croft W. A. 1991. Syntactic Categories and Grammatical Relations: The cognitive organization of information. Chicago: Univ. of Chicago Press.

Dowty D. R. 1979. Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ. Dordrecht (Holland): Reidel, 1979.

Filip H. 1999. Aspect, Eventuality, Types and Nominal Reference. N. Y– L.: Garland publishing.

Fillmore Ch. J. 1977. The case for case reopened // Syntax and Semantics. Vol. 8. N. Y. etc., 59–81.

Haspelmath M. 1993. More on the typology of the inchoative / causative alternations // B. Comrie, M. Polinsky (eds). Causation and Transitivity. Amsterdam; Philadelphia: John Benjamins, 1993.

Jackendoff R. S. 1990. Semantic Structures. Cambridge etc.: MIT Press, 1990.

Kustova G. I. 2004. Tipy proizvodnyx značenij i mexanizmy jazykovogo rasširenija. M.: JaSK, 2004.

Kustova G. I., Paducheva E. V. 1994. Slovar' kak leksičeskaja baza dannyx // Voprosy jazykoznanija, № 4, 96–106.

Levin B. 1993. English Verb Classes and Alternations: A preliminary investigation. Chicago: Chicago UP.

Levin B., Rappaport Hovav M. 1995. Unaccusativity: At the syntax-lexical semantics interface. Cambridge, Mass.: MIT Press.

Liutikova e.a. . 2006. Struktura sobytija i semantika glagola v karachaevo-balkarskom jazyke. M.: IMLI RAN.

Mel'čuk I.A. 1974. Opyt teorii lingvističeskix modelej "Smysl $\Leftrightarrow$ Tekst". Moskva: Nauka.

Mel'čuk I. A., Xolodovič A. A. 1970. K teorii grammatičeskogo zaloga // Narody Azii i Afriki. № 4, 111–124.

Paducheva E. V. 1996. Semantičeskie issledovanija: Semantika vremeni i vida v russkom jazyke. Semantika narrativa. M.: Jazyki russkoj kul'tury.

Paducheva E. V. 2003. Is there an "anticausative" component in the semantics of decausatives? Journal of Slavic Linguistics, v. 11, N 1, 173–198.

Paducheva E. V. 2004. Dinamičeskie modeli v semantike leksiki. M.: Jazyki slavjanskoj kul'tury.

Rappaport Hovav M.2008. Lexicalized meaning and the internal temporal structure of events //Susan Rothstein (ed.) Theoretical and Crosslinguistic Approaches to the Semantics of Aspect. Amsterdam: John Benjamins, 13-42.

Švedova N.Ju. 2007. Russkij semantičeskij slovar'. T.IV. Glagol. Moskva: Azbukovnik.

Testelec Ja.G. 2001. Vvedenie v obščij sintaksis. Moskva: RGGU.

Vendler Z. 1967. Linguistics in Philosophy. Ithaca, N. Y.: Cornell Univ. Press.
Wierzbicka A. 1980. Lingua mentalis. Sydney etc.: Acad. Press.
Wierzbicka A. 1987. English Speech Act Verbs: A Semantic Dictionary. Sydney etc.: Acad. Press.
Wierzbicka A. 1988. The Semantics of Grammar. Amsterdam; Philadelphia: John Benjamins.
Zaliznjak Anna A., Levontina I. B., Šmelev A. D. 2005. Ključevye idei russkoj jazykovoj kartiny mira. M.: JaSK.

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 21–61

A Contrastive Lexical Description of Basic Verbs

Examples from Swedish and Czech

Silvie Cinková

Abstract

This paper aims at a lexical description of frequent uses of frequent lexical verbs in Swedish on the background of Czech, with some implications for the lexical description of such verb uses verbs in general. It results in a draft of a production lexicon of Swedish frequent verbs for advanced Czech learners of Swedish, with focus on their uses as light verbs.

The introductory sections (1 and 2) discuss semantic shifts in highly frequent lexical verbs, whose most literal or 'primary' uses express motion, location, or physical control; e.g. *stand, put, go, hold*. These verbs are called *basic verbs*, which is a term coined by Viberg (Viberg, 1990) that suggests that they typically denote events belonging to *basic level categories* described by Lakoff (Lakoff, 1987). The 'literalness' of verb uses is judged according to how much they are the ones speakers pick first to illustrate the meaning of that given verb (*cognitive salience*, a term coined by Hanks in (Hanks, forthcoming). Hanks pointed out an interesting discrepancy between the cognitive salience and the actual frequency of a given verb usage in large corpora. This discrepancy is extremely significant in basic verbs. Some of their uses exhibit such a low cognitive salience, that they are not even noticed by native speakers. This has consequences in second-language acquisition. Foreign learners, even the advanced ones, often lack competence in using the most frequent lexical verbs of the second language in their most frequent patterns.

Basic verbs often act as light verbs. Sections 3 to 7 are dedicated to light verbs and light verb constructions. Section 8 discusses the morphosyntactic variability in predicate nouns (i.e. the nominal components of light verb constructions) and their possible semantic impact on the entire light verb construction.

Different aspects of polysemy of basic verbs are dealt with by contrasting Swedish examples to Czech in Section 9. Special attention is paid to uses of basic verbs that denote relations between abstract entities. Section 10 focuses on grammaticalizing uses of lexical verbs. It gives a Swedish example *of context-induced reinterpretation* – an interesting semantic shift that often leads to grammaticalization.

Please cite this article as: Silvie Cinková, A Contrastive Lexical Description of Basic Verbs: Examples from Swedish and Czech. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 21–61.

All the aspects of basic verbs discussed in Sections 1–10 are integrated in a structure of a Swedish-Czech lexicon, which captures verbs and predicate nouns in two respective interlinked parts. Sections 11-14 give its detailed decription.

1. Introduction

Probably every human language operates with a set of very frequent lexical verbs that are primarily perceived as verbs denoting location, motion, or physical control over something; e.g. *stand, go, put, keep, get*, etc. Their literal meanings (meanings not derived by metaphorical transfers) are very *cognitively salient* (Hanks, forthcoming); i.e. they are intuitively first associated with the verb ("what we think words mean"). For instance, the most cognitively salient meaning of *go* would have to do with spatial motion. On the other hand, the research on large text corpora reveals an interesting fact that the cognitive salience of a word usage does not necessarily correspond to its *social salience* ("the actual meanings that we use"), such as the most socially salient meaning of *go* can possibly be the future tense *(to be going to)*, which we would hardly consider the "most typical" meaning of *go*.

The socially salient meanings arise regularly through metaphorical shifts and processes of semantic deployment or generalization, in which the given verb loses or generalizes some of its semantic features (cf. Bybee et al., 1994; Heine et al., 2001) to expand its collocability. When the general semantic feature (in other words actually cognitive category) is relevant for an entire class of lexemes (such as future is for verbs), the distribution of the given lexeme (here a verb or its particular morphosyntactic form) gradually ceases to be limited by the collocability of the primary meaning, and the lexeme step by step turns into a universally usable language element. At that stage, it is perceived as a part of the grammar system. This process is called *grammaticalization* (Hopper, 1987) or *grammaticization* by some (Bybee, 1985; Bybee et al., 1994). It is a gradual process that spreads from isolated words, collocations, and phrases. Grammar is to be understood as, in Hopper's terms, "a real-time, social phenomenon", which is "always in process but never arriving, and therefore emergent", and not "the only, or even the major, source of regularity, but instead grammar is what results when formulas are rearranged, or dismantled and re-assembled, in different ways."

Evidently, there is a transition area between phraseology and syntax. There are millions of idiomatic expressesions that arose as new collocations or phrases constituted by a cluster of collocates as new sematic units in their own right. When the tightness of the collocation lies mainly in the cooccurrence of two autosemantic lexemes (e.g. *be in one's shoes*), rather than in the cooccurrence of one of several lexemes with a given morphosyntactic constellation in the environment (*keep* + *-ing*), it is well in place to refer to it as to a *lexicalized* expression, which is a phraseological term. Theoretically, we could make a closed list of idiomatic expressions occurring in a given language.

On the other hand, there are a number of systematically occurring collocations of lexemes and structural elements that are not easily captured by the phraseological approach; e.g. the colloquial structure *don't go doing something*, which intensifies the negative connotation of the event in question (or indicates it when the verb itself is stylistically neutral):

(1) *…that poor man probably gets compared to that character all the time. Don't go bothering him.* [COCA]

(2) *"It's okay," she said. "I'm fine. Don't go bothering about me when you've got Georgy lying here in this state."* [COCA]

(3) *"You'll do fine," he told her confidently on one of their walks along the Danube. "Just don't go marrying an Australian. I must have my little girl back someday."* [COCA]

(4) *Is it too much to ask, Jack, honey, that just once after we make love you don't go rushing off like there's a three-alarm fire?* [COCA]

(5) *"In the defense world, you're notified of bids, you negotiate a long-term relationship and you don't go knocking on doors and say, Here's our brochure…"* [COCA]

The actual meaning of this particular structure, in which the semantically heaviest part, namely the *-ing* form of the verb governed by *go*, is freely variable, is very general. An idiomatic expression, on the contrary, typically bears a complex meaning of its own and can either be paraphrased (e.g. *push out the daisies = be dead*) or related to a particular situation (*damn it! – you say it when you are very irritated, cheers – you say it when making a toast, or when informally thanking somebody, or when parting somebody informally*). This is, however, not the case when the semantically heavy part is just the one in the construction that can be replaced.

The collocational interplay between lexical items and structural elements that we are used to perceiving as syntax, is obscuring the borderline between grammar and lexis. For this phenomenon, Hoey (Hoey, 1998) re-used the term *colligation*, originally coined by J.R. Firth for *collocation*. Hoey (Hoey, 1998, quotation taken from Hunston, 2001) defines *colligation* the following way:

- The grammatical company a word keeps (or avoids keeping) either within its own group or at a higher rank.
- The grammatical functions that the word's group prefers (or avoids).
- The place in a sequence that a word prefers (or avoids).

Such observations have been made earlier; cf. Hunston in Hunston (2001, p. 15): "If we take seriously Sinclair's assertion that there is no longer any sense in distinguishing between lexis and grammar […], then the distinction between collocation and colligation to a large extent disappears. On the other hand, the term *colligation* is helpful in drawing attention to the fact that the evidence in many instances of naturally-occurring language can be used to explain behaviour that is traditionally associated with grammar. Just as the discipline called 'lexis' has been assisted

by corpus-based approaches to collocation, so the discipline 'grammar' benefits from corpus-based approaches to colligation."

An attitude that no longer makes a distinction between grammar and phraseology can be very useful in describing semantically depleted uses of common lexical verbs that have not achieved universal collocability with a clearly defined word class (be it part of speech or even a generally known semantic criterion such as 'animate nouns', 'verbs denoting states', etc.). Such patterns of uses do not often penetrate grammar textbooks, but are on the other hand in a way too vague to be described as multiword units in lexicons. They are often ignored by native speakers as 'untypical' uses. Their cognitive salience can be extremely low, while their social salience can be high at the same time; and that is why they deserve special attention.

2. Basic Verbs

This paper aims at a lexical description of socially, but perhaps not enough cognitively salient uses of frequent lexical verbs in Swedish on the background of Czech, with some implications for the lexical description of basic verbs in general. It results in a draft of a production lexicon of Swedish basic verbs[1] for advanced Czech learners of Swedish, with focus on their uses as light verbs (see Section 3).

Verbs possess the ability to bring entities into relations and create propositions. An analysis of the most frequent lexemes in Swedish (Viberg, 1990) shows an interesting fact: there are far fewer verbs in the language than there are e.g. nouns. The Swedish frequency dictionary (Allén, 1972) contains 39 486 nouns but about 8,5 times fewer verbs (4 649).

We reflect the manifold features of different entities by a vast amount of nouns at our disposal, whereas we evidently need a significantly smaller amount of verbs to describe the relations these entities enter. Besides, there is an evident preference for just a selection of verbs. Viberg (Viberg, 1990) observed in Swedish, in full accordance with Zipf's law, that almost one half (45.5%) of the verb occurrences is represented by the 20 most frequent verbs. Almost every second verb used in the language is then one of the top-twenty.[2] This implies that some verbs have an extreme potential to fit into many different contexts.

For these verbs, Viberg coins the label *basic verbs*. According to Viberg, they are characterized by the following features:

1. They are simple stems rather than derivations or compound words.
2. They have a phonologically simple form.

[1] a term coined by Viberg in Viberg (1990)

[2] The 20 most frequent nouns cover only 8.1% of noun occurrences, the 20 most frequent adjectives cover 24.2% of adjective occurrences and the 20 most frequent adverbs have similar rate as verbs – 42.1% of adverbial occurrences.

3. Their conjugated forms are often irregular.[3]
4. They occur in the respective languages with high frequency.
5. Typologically, they have a broad distribution (their equivalents exist in many languages)
6. They have many "secondary" meanings[4].
7. They have a significant potential to become grammatical markers.
8. They act as syntactic prototypes (i.e. they allow for many valency patterns and occur in more compound words and derivations).
9. They are preferred at the early stages of first as well as of second language acquisition.

For Swedish, the top 20 verbs are the following:

1. *är/vara (to be)*
2. *ha (to have)*
3. *kunna (can)*
4. *ska (shall)*
5. *få (to get)*
6. *bli (to become)*
7. *komma (to come)*
8. *göra (to do, to make)*
9. *finnas* (existential *to be*, lit. *to be found*. Similar to the German *es gibt*.)
10. *ta (to take)*
11. *säga (to say)*
12. *gå (to go)*
13. *ge (to give)*
14. *se (to see)*
15. *måste (must)*
16. *vilja (to want)*
17. *stå (to stand)*
18. *visa (to show)*
19. *böra (ought)*
20. *gälla (to apply, to be valid)*

For the purpose of this paper, only a subset of what Viberg calls basic verbs is analyzed: verbs that are able to act as light verbs[5]. Copula verbs and modal verbs are ignored.

[3]This can indicate that the respective forms are acquired by rote learning and remain further unanalyzed by speakers (cf. Bybee, 1985).

[4]Many studies on this have been published in the Scandinavian area also by other authors. Among others Ekberg (1993), Fenyvesi-Jobbágy (2003), Hansen (1974), Jakobsson (1996), Jensen (2000), Malmgren (2002), Pihlström (1988), Reuter (1986).

[5]see Section 3

As a rule, the verbs in question rank among the top 50 most frequent verbs. It is mostly verbs of spatial motion, location, and physical control. The most typical members of this group are *stå (stand), ligga (lie), sitta (sit)* and their causative counterparts *ställa, lägga,* and *sätta*, *ge (give), ta (take), hålla (hold/keep), gå (go), komma (come), göra (do/make), falla (fall), fälla* (causative to *falla*), *bjuda (offer), visa (show/exhibit), möta (meet/face)*, and *få (get)*. A list of potential light verbs was obtained earlier by extracting verb-noun collocations from the 20-million morphosyntactically tagged Swedish corpus PAROLE (Cinková, 2004).

3. Light Verb Constructions

Light verbs and light verb constructions (henceforth LVC's) are an interesting instance of semantic shifts in lexical verbs. Their numerous definitions set by many different linguists agree that LVC's [6]) consist of a lexical verb and a noun phrase and that it is the noun that carries the semantic weight. The verb is deprived of its original meaning. It only delivers morphosyntactic categories and, possibly, some semantic features to the resulting event description. Not seldom, the valency behaviour of the verb changes when the verb acts as a light verb. For instance, *give a sigh* has nothing to do with *giving* but with *sighing*, which is supported by the fact that *give* in this case obviously opens no addressee slot.

Lexical verbs which lose their concrete meaning when combined with abstract nouns and nominalizations and which occur in such combinations very productively, appear to be very common in modern European languages, but also beyond Europe, as already noted by R. Jakobson (for reference see Jelínek, 2003, p. 50). They were even observed in South-Asian languages (Butt, 2003), which are linguistically as well as culturally very distant from the European languages.

Butt in (Butt, 2003) claims that although light verbs potentially are a universal linguistic phenomenon, they have different structural features in the respective lan-

[6]This paper uses a term coined by Jespersen in Jespersen (1954), but they are also known under many other names. In English linguistics it is e.g. *support verbs, support verb constructions, expanded predicates, verbo-nominal phrases, delexical verbs, stretched verbs*.

German linguistics has studied *Funktionsverbgefüge* and *Funktionsverben* (also under different terms) intensively since the term was coined by von Polenz (Polenz, 1963). Interest in this issue rose especially with the onset of generative and transformational grammar (among others in Rothkegel's studies on fixed syntagms Rothkegel, 1973). To be mentioned are also at least Persson's studies on causativity (Persson, 1975;Persson, 1992), as well as the research in German as a foreign language (Helbig and Buscha, 1996 and Günther and Pape, 1976).

The terms, especially the German terms as *Funktionsverben, Nominalisierungsverben, verblasste Verben, Streckformen*, etc., cannot be used interchangeably. Some authors using the respective variants were observing only the combinations of a verb and its direct object, others only the combinations of a verb and its prepositional object. For a summarizing comparison of the light-verb related terms in German and English see e.g. Hanks et al. (2006).

guages[7]. Hence all syntactic tests for defining light verbs and light verb constructions are language-specific (p. 24 in the web-released manuscript of Butt, 2003). E.g., in Germanic languages, the following criteria are commonly quoted:

- Light verb constructions with the predicate noun in the position of the direct object cannot be passivized.
- The predicate noun cannot be replaced with an anaphoric expression.
- There should be at least an option of the predicate noun to occur without a determiner (a criterion applied to Swedish, see Dura, 1997).

Few verbs are light under all circumstances: there belong those that combine only with nominalizations or event nouns, such as *perform, carry out*[8]. The syntactic behaviour of the word combination is an important clue for all verbs that can either act as lexical verbs **or** as light verbs according to their context. However, the syntactic criteria do not apply 100%.

Hanks et al. point out in (Hanks et al., 2006) that "lightness is a matter of degree", and that "some uses [of verbs that can act as light verbs, S.C.] are lighter than others" (p. 441). They emphasize the collocational and semantic criteria for deciding whether a verb use is light or not: "The problem lies in the expectation that necessary and sufficient conditions can be established for delicate grammar categories, as opposed to characterizations of typical features. Light verbs typically focus attention on an event or process, and events and processes are very often expressed in nouns that are nominalizations (i.e. cognates of verbs) – but the focus is still on the event, even when the direct object is a word that denotes a physical entity" (p. 443). They introduce the notion of *semantic lightness* in their analysis of the verb – direct object combinations, and there is no apparent reason not to relate this term also to verb – prepositional object combinations, which their paper does not address.

Butt (Butt, 2003) draws an interesting conclusion from diachronic English studies, which supports favouring semantic and collocation criteria over syntactic the syntactic – although in their function similar to auxiliary verbs, light verbs, unlike auxiliaries, do not underlie the grammaticalization process in the development of a given language: "Light verbs straddle the divide between the functional and lexical in that they are essentially lexical elements but do not predicate like main verbs" (p. 4 and 13 in the web-released manuscript of Butt, 2003).

[7] E.g. in Butt's example from Urdu, a light verb construction even requires a second lexical verb attached to the light verb in the verb-noun structure.

[8] In this context it is to be added that evaluative expressions that are neither nominalizations nor event nouns act as such in light verb constructions; e.g. *He committed something horrible.*

4. Light Verb Constructions as Collocations

LVCs can be regarded as a type of collocation. Malmgren (Malmgren, 2002, p. 12)[9] describes a number of candidate LVCs, calling them a kind of "prototypical collocations" that consist of a semantically impoverished verb and an abstract noun. The abstract noun keeps its meaning, hence it is considered to be the more stable member of the collocation – the collocational base (or *node*, see Sinclair, 1993). Its verbal collocate is generally unpredictable.

Inspired by Mel'čuk's Meaning-Text-Theory (Mel'čuk, 1996), Malmgren analyzes Swedish verbal collocates and associates them with nouns by means of the lexical function Oper. Fontenelle (Fontenelle, 1992, p. 142) also claims that "Support Verbs roughly correspond to the type of lexical relation that can be encoded through the Oper Lexical Function used by Mel'čuk".

The understanding of nouns as collocational bases in verb + abstract noun constructions is clearly shared by Čermák, (e.g. František Čermák, 1995): "Abstract nouns seem to follow a few general patterns in their behaviour, which seem to be more structured, allowing for much less freedom than concrete nouns. The patterns the abstract nouns enter are determined by their function and meaning".[10]

While Helbig and Buscha were seeking to identify a distinct class of "Funktionsverben", and Baron and Herslund (Baron and Herslund, 1998), Rothkegel (Rothkegel, 1973), and Persson (Persson, 1975, Persson, 1992) were trying to define light verb constructions by the semantic relation between the noun phrase and the verb, Fontenelle, Malmgren, and Čermák focused on the noun, in full accordance with the pregnantly formulated observation of Hanks (Hanks, forthcoming): "…it seems almost as if all the other parts of speech (verbs and function words) are little more than repetitive glue holding the names in place".

Even in the cross-linguistic perspective, it is usually the noun that is the common denominator for the equivalent light verb constructions: "The verb […], although often the only one that is correct and idiomatic, can seem totally arbitrary. In another language – mutatis mutandis – totally different verbs often occur which work as place holders; that is why prototypical collocations often cause translation problems" (Malmgren, 2002, p. 11, and cf. Schroten, 2002).[11] Malmgren further notes that "sometimes, though by far not always, one can anticipate a sort of metaphorics" in the choice of the verb. According to Malmgren, the eventual metaphors can be traced back and explained ex post facto, but they are definitely not predictable within any one given language, let alone cross-linguistically.

[9] Malmgren's starting point is the system-oriented understanding of collocations coined especially by German linguists as Hausmann and Heid (Heid, 1998, p. 302) rather than the original English contextualist approach to collocations.

[10] Though Čermák explicitly avoids the term 'collocation', using the expression 'stable combinations' instead, among which "some are undoubtedly more frequent than others".

[11] The quotations of Malmgren, Ekberg and Dura were translated from Swedish by S.C.

5. Semantic Aspects of LVCs

From the semantic point of view, the noun seems to be a part of a complex predicate rather than the object (or subject) of the verb, despite what the surface syntax suggests (cf. Schroten, 2002, p. 93, and Boje, 1995, pp. 53, 145). As already stated by many authors (e.g. Helbig and Buscha, 1996), light verbs are in fact lexical verbs that have to some extent lost their lexical meaning, in order to provide the predicate nouns with verbal morphological categories (which is the feature that makes them resemble a verb class according to Helbig and Buscha (1996) – *Funktionsverben*, and Jelínek (2003, p. 40) – *operational verbs* (*operační slovesa*.)

Many students of this topic have observed that verbs, when occurring in an LVC, start to carry more abstract semantic features. Rothkegel (1973) considers the semantic bleaching[12] of the verb to be the antipode of verbal polysemy. She shows that the meaning of a given lexical verb in LVCs neither matches any of its meanings outside LVCs, nor does it create new meanings when associated with the respective noun phrases, which implies that instead of just being deprived of a part of its original meaning, the lexical verb acquires an additional, more abstract meaning that is reserved for the verb's occurrence in LVCs.

Butt (2003, p. 18 of the web-released manuscript) proposes that light verbs are characterized precisely by the ability to express general features, as described by Rothkegel (1973). However, Butt is explicit in that she does not regard light verb uses as semantic derivations of the primary meanings of the verbs, but contrary to that, she assumes that "the lexical specification of a handful of verbs (somewhere between 5 and 20) cross-linguistically allows for a use as *either* a main verb *or* a light verb. Some common examples crosslinguistically are the verbs for *come, go, take, give, hit, throw, rise, fall*, and *do/make*. […] Their lexical semantic specifications are so general that they can be used in multitude of contexts, that is, they 'fit' many constellations."

6. LVCs and Event Structure

LVCs are often referred to as a means of modifying the event structure of a locution, especially in languages such as Swedish, which do not (regularly) indicate aspect by morphological means (i.e. by stem vowel alternations or affixes). In such languages the aspect remains underspecified, unless lexical markers (e.g. temporal adverbs) are employed in the utterance. A kind of event structure opposition is assumed between an LVC and its corresponding synthetic predicate (when there is one). Butt (2003, p. 18 of the web-released manuscript) in accordance with many other authors, emphasizes that "light verbs modulate or structure a given event predication and do so in a manner similar to that of modifiers with respect to semantic notions such as

[12]She quotes other authors' terms, such as 'das Verblassen der Merkmale bei den Verben", "Bedeutungsentleerung", "depletion of the designatum".

benefaction, suddenness, etc.[13] […] The light verbs also tend to add further information about the aktionsart of the complex predication. In particular, there is often a telic/boundedness or a causation component." In this respect they have a function similar to verbal prefixes or particles (Butt, 2003, p. 16).

LVCs are built as compositional events or constructions consisting of a 'verbal' and a 'nominal' subevent. Yet the 'verbal' event does actually never 'take place' due to the semantic depletion in light verbs (cf. Fillmore et al., 2003). The given light verb only passes some semantic features on to the 'nominal' event. Durative events are by definition atelic (e.g. *to have problems*), with the reservation that multiple telic 'nominal' events combined with a durative atelic light verb express iterativity, e.g. *to suffer from attacks*.

LVCs denoting transitions (i.e. changes of state) are generally regarded as telic (cf. Pustejovsky, 1991), no matter what telicity value the given light verb would have if used as a lexical verb outside the LVC. Bjerre (1999) puts it this way: "LVCs denoting transitions are invariably achievements[14], either inchoatives or causatives […], the SV [i.e. *support verb*, which is the term Bjerre preferres to *light verb*. S.C.] always denotes an underspecified subevent$_1$.[…] Not surprising *terminative* is the negative counterpart of *inchoative*."

Bjerre's examples make it more clear: "*Situationen kom ud af kontrol – [The situation came out of control]* denotes a situation in which the resultant state is the negative of that in *Situationen kom under kontrol [The situation came under control]*. […] This may be paraphrased: (subevent$_1$:) The situation was under control when something happened as a result of which (subevent$_2$:) the situation was out of (= not under) control". Bjerre notes that light verbs denoting transitions are either achievement verbs with inherently underspecified subevent$_1$ (*come, bring* etc.), or they are verbs of motion or location which lose their specific relation when used as light verbs.

7. Productivity vs. Lexicalization in LVCs

Whereas traditional views emphasize that it is mostly the lexicalized units that tend to show a specific syntactic behaviour and, therefore, LVCs are to be considered as more or less lexicalized phrases, Ekberg (1987) and Dura (1997), as well as Persson (1992), concentrate on the apparent productivity of LVCs and the regular production patterns they form. Ekberg notes that many lexicalized phrases "have an almost completely or at least partly predictable meaning and new ones can be formed according to productive rules within the grammar" (Ekberg, 1987, p. 32), while Dura goes even further, adding that "even the newly-formed phrases show the same syntactic restrictions as the lexicalized ones" and interpreting this phenomenon as evidence that

[13] Cf. also Schroten (2002).

[14] Transitions are further divided into two subtypes. In *achievements* the subevent$_1$ is underspecified, unlike in *accomplishments*, e.g. *Carl built a house* (accomplishment) × *The expedition reached the top of a mountain* (achievement). See Bjerre (1999).

"these restrictions indicate that something is meant as a lexicalization rather than that they are the result of lexicalization" (Dura, 1997, pp. 1–3). She considers article-less verb-noun combinations to be an evidence that there is "a kind of word combination that is not controlled by the regular syntax but aims at lexical composition" and that it is thus "possible to form new phrases which can act as lexical units. The ordinary syntax is oriented at combining lexical units with obligatory grammatical categories, but there even seems to be another syntax, a syntax which allows language users to build larger conceptual units without involving the grammatical categories". Dura and Ekberg approach the issue from the semantic side, though they seek to draw syntactic conclusions. The syntactic criteria are eventually more important for Dura and Ekberg than they are for Hanks and others.

8. Grammatical Interference in Lexicalized Collocations?

When the morphosyntactic behaviour of a multi-word cluster systematically deviates from the regular grammar rules, it is traditionally regarded as intensively lexicalized, i.e. several words are thought of as growing together into one single semantic unit. Moreover, Dura (1997, see above) suggests that the cause – consequence relation also works the other way round: collocations that **are meant** by the speakers to be perceived as a single semantic unit are deliberately taken out of the regular language system.

Many authors since the onset of corpus linguistics have observed that the regular language use to a significant extent consists of prefabricated blocks. Needless to say, this phenomenon goes far beyond idioms and terminology. For instance, Wray (2002) builds her hypotheses on formulaic sequences on the premise that "although we have tremendous capacity for grammatical processing, this is not our only, nor even our preferred, way of coping with language input and output. […] much of our entirely regular input and output is not processed analytically, even though it could be" (p. 10).

Light verb constructions appear to be such formulaic clusters. Collocations that sometimes behave according to grammar rules and sometimes do not, would normally be regarded as somewhere half-way to the ultimate lexicalization; i.e., they would be expected to exhibit only irregular behaviour in the future development of the given language.[15] However, morphosyntactic realizations of semantically transparent collocations in text may not just vary in the extent to which they comply with the rules of grammar in terms of 'right' versus 'wrong', but, on the contrary, different

[15]This is of course not the case of idiomatic expressions, whose idiomatic meaning is inseparable from their morphosyntactic realization; e.g. *abandon ship*. Example 1 implies that the ship is thought to be sinking, whereas Example 2 lacks this implicature:

(1) *Abandon ship!*

(2) *They abandoned the ship in a bay near Hong Kong.*

grammatical realizations of collocations can have different semantic/pragmatic implicatures in the particular context according to the speaker's preference. A default behaviour of lexicalized semantically transparent collocations may often be irregular (e.g. zero article, no modifiers allowed, etc.), but the corpus evidence suggests that there is not necessarily a clear ban on a step back to the regular grammar when the morphosyntactic features help reflect the communicational intentions of the speaker in a particular discourse situation.

In other words, the assumption is that regular morphosyntactic behaviour is reintroduced when the speakers explicitly want to add the semantic features triggered by regular morphosyntactic behaviour, but they are by no means obliged to do it. The presence or absence of semantic differences between two or more alternative morphosyntactic structures is very much context-dependent, and the semantic oppositions can be obscured by the fact that they happen to be irrelevant in a particular context. That implies that the alternative expression forms will not always be mutually exclusive, but that the speakers only have the option to select the non-default pattern when they feel a particular reason for doing that.

To mention a Swedish example, the light verb construction *sätta rekord (set a record)* is normally used without an article, even when *rekord* is modified by one or more adjectives (adjective modifiers usually require the use of an article in Swedish):

(6) *Mustafa Mohammed satte personligt rekord.*
Mustafa Mohammed set a personal record.

(7) *Stefan Holm klarade 2,37 i Globen och satte nytt personligt rekord.*
Stefan Holm made 2.37 in Globen and set a new personal record.

The collocation *sätta rekord (set a record)* appears to be a very lexicalized one, judging from the predominating zero article. The large Swedish corpus Konkordanser showed that the absolute majority of the occurrences of *sätta rekord* had no article preceding *rekord*. The Konkordanser subcorpora yielded 223 occurrences of the forms *sätta, sätter, satte* and *satt*, respectively, with *rekord* following within the same sentence[16]. The noun *rekord* occurred with the indefinite article only 17 times. The percentual rates were the following:

- 2 % in the infinitive
- 0 % in the present tense
- 11 % in the simple past tense
- 9 % in the perfect tense

The definite singular form *rekordet* and the definite plural form *rekorden* occurred 11 times and once in collocation with *sätta*, respectively.

[16]Unfortunately, in Konkordanser, modern Swedish texts (newspapers and fiction) are split into 14 subcorpora, and the interface does not allow multiple selection. None of the subcorpora in Konkordanser is either tagged or lemmatized, and the interface does not support CQL. Simple Boolean queries or wildcard searches can be performed, but they cannot be combined, which significantly limits the searching power.

The 29 hits with (any) article represented 12% of the total of 235 hits.

The most frequent case (indefinite article) does not seem to be affected by tense. A closer analysis of the broader contexts showed at least one situation in which the insertion of the indefinite article may be triggered by the context (approx. 1/3 of the hits with the indefinite article) – it is when the discipline in which the record was set is specified later in the text (selection):

(8) *Svensson har satt ett oslagbart svensk rekord som sportjournalist: under cirka 49 år hade han fast jobb på samma redaktion i samma tidning, Arbetet i Malmö.*
Svensson has set an unbeatable Swedish record as a sports journalist: for approximately 49 years he had had a regular job at the same publishing office, at the same newspaper, Arbetet i Malmö.

(9) *Förre RIK-aren Peter Gentzel har satt ett nytt rekord i tyska Bundsliga. Den svenske landslagsmålvakten har på 34 omgångar tagit hela 53 straffar för Nordhorn.*
Former RIK-player Peter Gentzel has set a new record in the German Bundesliga. The goalkeeper of the Swedish national team has got 53 yellow and red cards for Nordhorn in 34 rounds.

(10) *Massorna, som köade i en halvmil för att slutligen komma till Hyde Park, satte ett nytt rekord i levande opinionsbildning.*
The crowds that were queuing for a half mile in order to finally get into Hyde Park set a new record in live opinion making.

(11) *Anette var andra halvlekens gigant och satte då ett personligt rekord. – Har aldrig gjort åtta mål i en och samma halvlek i elitserien.*
Anette was the giant of the second half and it was then that she set a personal record. – I have never shot eight goals in a single half in the elite series.

In other two cases (one with an indefinite pronoun) the sentence describes an unreal or non-specific condition:

(12) *Han säger att visst, landslaget skulle väl vara kul och visst sätta ett svenskt rekord skulle väl också vara kul, men det är saker han inte går och tänker på.*
He says that yes, the national team would obviously be cool and obviously it would also be cool to set a Swedish record, but that is stuff he doesn't go thinking about.

(13) *Om jag sätter något rekord så kommer det snart någon och slår det.*
Even if I set a record, someone else will soon come and break it.

Also setting two entities in contrast normally requires an article, as can be seen in Example 14:

(14) *Hägerstenskillen […] satte ett personligt rekord och tangerade ett: Han presterade 60 kilo i stöt (tangerat pers.) och 47,5 kilo i ryck (personligt med 2,5 kilo).*
The guy from Hägersten […] set a personal record and attacked another one: He lifted 60 kg ……

In addition, the discipline in Example 14 was specified later.

Example 15 originates from a context where records were expected in several different disciplines. A certain swimming discipline was the first discipline in the entire competition where it happened: a European record was set. In this particular context, the European record, which is a unique uncountable entity in the context of one single discipline, is regarded as countable and a member of a set.

(15) *Engelsmannen Adrian Moorhouse blev den första att sätta ett Europarekord i Strasbourg.*
The Englishman Adrian Moorhouse was the first one to set a European record in Strasbourg.

In all the other 10 hits except one, the noun *rekord* with the indefinite article was modified by one or two adjectives. All of the adjectives denoted restrictive attributes. The use of a restrictive attribute implies that that particular record was one of a set, which is normally a good reason for employing an article. Nevertheless, the zero-article is strongly preferred in this context and with the modifiers *svensk (Swedish), personlig (personal), ny (new)*, even when they concatenate. No differences in the broader context were observed that would explain why the article was used. Only a sample is presented here.

(16) *Även om serien inte var perfekt satte han ett nytt prydligt personligt och svenskt rekord med 387,60 poäng.*
Even though the series was not perfect he set a new nice personal and Swedish record by 387,60 points.

(17) *Orbit Air vann både försök och final i fjol och satte ett nytt svenskt rekord.*
Orbit Air won both the trial and the final last year and set a new Swedish record.

The definite article (found 12 times) was consequently used when referring back to one particular record mentioned before – either to the same entity (the same discipline, the same year, the same person), or to a contrasting entity. Only a selection is presented.

(18) *Hennes svenska rekord på 1.500 meter på 4.09,0 är internationellt gångbart och den tiden är ingen yttersta gräns för Gunilla. Det finns mer att ge. – När jag satte det rekordet var jag inte ens trött efter loppet. Det kändes som att dansa fram.*
Her Swedish record in the 1 500 meters at 4.09,0 is internationally accepted and this time is not the ultimate limit for Gunilla. There is more to give. – When I set that record I was not tired at all after the run. It felt like dancing.

(19) *När Bartova satte det kortlivade rekordet i Prag snodde hon det från just Flosadottir som tog sig över 4,42 …*
When Bartova set the short-lived record in Prague, she had just stolen it from Flossadottir, who got over 4,42..

(20) *Det svenska skattesystemet sätter det ena otroliga rekordet efter det andra.*
The Swedish tax system sets one incredible record after another.

It is interesting to investigate to which extent the regular grammar continues to affect multi-word clusters that already have reached the stage of lexicalization, which in principle allows them to ignore grammar. This kind of research suggests the cases in which speakers may deliberately decide **to exploit** grammar in pursuit of a particular communicative goal, since they are not forced to respect grammar for its own sake. Investigating grammar in positions where the default is not to use it at all can reveal a lot about the semantic potential of our traditional grammar categories in general.

9. Polysemy

9.1. Relations among Concrete Entities

The previous sections discussed light verb constructions and the light verbs. The majority of verbs that can be used as light verbs is also polysemous in other ways. A contrastive, corpus-based comparison of the use of basic verbs reveals that in many different contexts where Swedish employs a basic verb, the Czech equivalent is stylistically marked or more specific with respect to the given context. Quite naturally, this difference lies partly in the Czech aspectual dichotomy, which can be realized morphologically – i.e. by a stem change – as well as by derivation. Even so, however, Czech employs many more verb lemmas with mutually unrelated stems than Swedish. This implies that the Swedish basic verbs have a far higher collocation potential and a more intricate polysemy than the corresponding Czech basic verbs. In other words, a Swedish learner of Czech must learn many different verbs with a relatively low collocation potential to produce idiomatic text, while a Czech learner of Swedish must acquire very elaborate cognitive maps of collocations appropriate for a few verbs, respectively.

Fig. 1 shows one instance of this equivalent discrepancy: to express that X caused Y to sit in prison, underspecifying whether condemned to or literally escorted, Swedish uses predominantly the verb *sätta*, which is stylistically neutral. Alternatively (with far lower frequency) it uses *kasta (throw)*, which is expressive. In Czech, a number of verbs is used in place of these two Swedish ones, with the frequency counts decreasing continuously, with no abrupt drops. The counts were obtained from the corpora PAROLE (SW) and SYN2005 (CZ) (Hajič, 2004 and Spoustová et al., 2007).

The discrepancy between the collocation potential of Czech and the Swedish basic verbs grows even more evident in cases like Fig. 2 when the collocate of the given Swedish basic verb is not a single noun but a set of non-synonymous nouns that all have the same semantic relation to the basic verb. Here Czech operates with a vast amount of not mutually interchangeable verbs, which are chosen in accordance with the semantic features of the respective noun collocates. There is, unlike in Swedish,

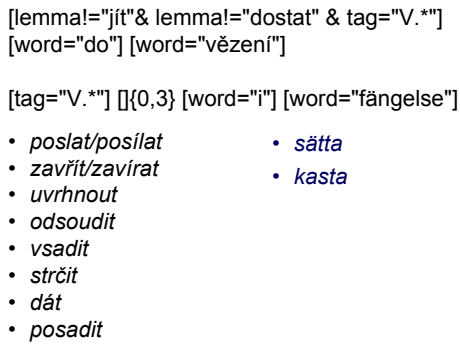


Figure 1. X puts Y into prison: verbs for put in Swedish vs. in Czech

no superior verb that would be universally used with all these nouns. The counts were obtained from the same corpora as those in Fig. 1

9.2. Swedish Spatial Conceptualization

On the one hand, Swedish seems to operate with fewer verbs than Czech. On the other hand, there is a conceptual area where Swedish systematically requires a higher degree of lexical specification than Czech. Swedish does not have any direct equivalent to the Czech dát *(give)* in the sense *put (place something somewhere)*. The speakers of Swedish must learn to choose the right verb from the set *sätta, ställa*, and *lägga*, depending on the spatial orientation of the object being moved, on the character of the target location, or even on whether the object is being attached to its target destination (e.g. with glue) or whether it keeps its new position by itself. Needless to say, a conventionalized world knowledge specific to the Swedish language community comes into the play.

To name a few examples that a Czech speaker would never resolve correctly unless he has explicitly learned them: Something that can be regarded as attached or stuck is mostly regarded as "sitting" and, accordingly, "being put into a sitting position". Thus a football can "sit" in a broken window pane, and what a post-it pad usually does on a door is also "sitting". Hence also the motivation for the example illustrated

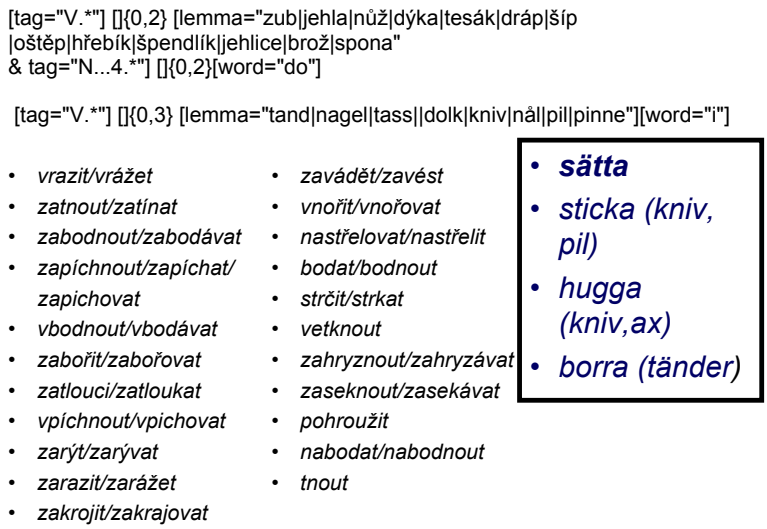


Figure 2. X inserts a sharp object into Y: verbs for insert in Swedish vs. in Czech

by Fig. 2. Besides, you can also *sätta* a plate on the table, as well as you can *ställa* it (*put* vertically or something vertical), while the plate, once placed on the table, stands there (*stå*). Jakobsson (1996) claims that a plate can also *ligga (lie)*, but only when it is positioned upside down or when it is broken. (No cooccurrence of *ligga* and *tallrik* (*lie* and *plate*) was found in PAROLE to prove it, though.) The motivation is that the functional part of the plate points up, which gives the concept of the entire object a vertical flavour, although it is actually flat and horizontal.

In Czech, opposite to that, *sedět/stát sitta/sätta* is out of place with *plate*, nor is it usually used to express location/placement at all. On the other hand, *ležet (lie)* and *stát (stand)*, along with the corresponding causatives, are in this respect both synonymous and roughly equally frequent, as the large Czech corpus SYN2005 reveals.

9.3. Polysemy in Relations with Abstract Entities

Basic verbs belong to lexemes that "encode major orientation points in human experience" (Bybee et al., 1994, p. 10) and as such they have a mighty potential of metaphorical shifts. This paper focuses mainly on those semantic shifts that can be regarded as grammaticalization. However, the lexical description would be incomplete if it ignored those semantic shifts that have little potential to expand their collocation potential to become a universally distributed auxiliary; the more so that the shifts are language-dependent. A contrastive view is therefore absolutely necessary here. This section is dedicated to metaphorical uses of basic verbs, in the sense of "figurative" rather than what we intuitively perceive under "grammaticalized" (although, as noted above, there is no clear boundary between these two groups, and we would better perceive them as two ends of a scale rather than two sets).

Metaphorical uses that do not expand their collocation potential often arise through what Heine et al. (2001) call *Metaphorical extension from one semantic domain to another*[17]. Metaphorical abstraction relates concepts across semantic domains.[18]
Metaphorical abstraction is the way humans conceptualize the non-concrete aspects of the world. It is the naive picture of the world, in which it does not matter what the world actually is like, but what humans believe it is like. The naive view of the world is anthropocentric. Thus the closest and most discrete objects are parts of the human body and objects that can be physically manipulated. They help to 'manipulate' the less distinct entities in discourse by acting as metaphorical vehicles (Lakoff, 1987).

Heine et al. assume that the semantic domains make a hierarchy of metaphorical abstraction, through which source structures develop into target structures:

PERSON-OBJECT-ACTIVITY-SPACE-TIME-QUALITY[19]

Just to illustrate one pair, the SPACE-to-QUALITY transfer means that structures suggesting that an object is located at a place or aims in a direction regularly express that the object finds itself in a certain state or a certain situation:

(21) *The country is **sliding into** a depression.*

(22) *Belinda fell completely in love with her daughter: 'I felt **high** for about four days, not thinking about anything but caring for her.'*

Metaphorical shifts can be explained ex-post, but they cannot be predicted. This implies that there is no general principle, according to which metaphorical uses of the source language could be universally transformed into the corresponding target language. The only solution seems to be sufficient exemplification with respect to the

[17] Nevertheless, Heine et al. also delievr many examples of established function words that have arisen through this semantic shift.

[18] though metaphorical transfers also occur within a single semantic domain

[19] Heine et al. (2001, p. 48).

learner's language background; i.e. make sure to provide cases, in which Swedish would use a verb in a way unpredictable for a Czech speaker. For instance, Czech speakers *are* in a divorce when divorcing *(být v rozvodovém řízení)*, while Swedish speakers *lie* in a divorce *(ligga i skillsmässa)*. There are hundreds of such examples, and all must be consciously learned.

10. Grammaticalization through Context-Induced Reinterpretation

10.1. Context-induced Reinterpretation

As the most frequently used terms suggest, many authorities regard generalization, which lies behind or accompanies grammaticalization, as a loss of certain semantic components, compared to the core meaning of the original lexeme: *semantic bleaching* (coined by Givón) and *weakening of semantic content* (Bybee, Perkins and Pagliuca). Yet Heine, Claudi, and Hünnemeyer (Heine et al., 2001) argue that generalization is not always a reduction of meaning (p. 40f.). They present examples of negation of the core meaning and examples of addition of further semantic components not present in the core meaning. Generalization typically occurs in the following types of semantic changes:[20]

- *Metaphorical extension from one semantic domain to another*[21]
- *Context-induced reinterpretation*[22].

Heine et al. (2001, p. 70) note that a metaphorical transfer appears rather discrete. However, they propose that the transitions from one semantic domain into another create a continuum of linguistic expressions and call this continuous grammaticalizing process *context-induced reinterpretation*. They explain it on the verb *to go* in the following sentences:

(23) *Henry is going to town.*

(24) *Are you going to the library?*

(25) *No, I am going to eat.*

(26) *I am going do to the very best to make you happy.*

(27) *The rain is going to come.*[23]

Examples 23, 24, and 25 illustrate a SPACE-TIME metaphorical transfer. In Example 23, the verb *to go* has a clearly spatial meaning, whereas in 26 and 27 it has a clearly

[20] according to Heine et al. (2001) and partly Bybee et al. (1994)

[21] see previous section and cf. Heine et al. (2001).

[22] *inference* or *conventionalization of implicature* in Bybee et al. (1994)

[23] Quoted from Heine et al. (2001, p. 70). According to an English native speaker's view, 27 sounds unidiomatic and should be rephrased as *It is going to rain*.

temporal meaning. Yet Sentences 24 and 25 are ambiguous, depending very much on the context. The sentences can be interpreted in the following way:

(28) *Henry is going to town.* SPACE

(29) *Are you going to the library?* SPACE

(30) *No, I am going to eat.* (as answer to 24) INTENTION (+ relics of spatial meaning are still present)

(31) *I am going do to the very best to make you happy.* INTENTION

(32) *The rain is going to come.* PREDICTION

Both 26 and 27 have temporal meaning, but they differ in the desire of the respective subjects to pursue the event, since *rain*, let alone the empty *it*, cannot have a will or desire, while a human can.

To explain the semantic continuum, Heine et al. (2001) introduce three idealized stages of semantic shifts:

Stage I: A lingustic form F acquires a side-meaning B in addition to its core meaning A when employed in a certain context. At this stage, the utterance can be ambiguous as long as the context (both intra- and extralinguistic) does not eliminate the ambiguity, and it can be misunderstood by the recipient. (This would apply for 25.)

Stage II: The form F can be used in contexts where only the meaning B can be employed. (This would apply for 25.)

Stage III: The meaning B becomes conventionalized and cognitively salient enough to be conceived as a second meaning of the form F, which becomes polysemous. (This applies for 26 and 27). However, the meanings A and B are conceptually linked as the transition was continuous (p. 72).

Heine et al. (2001) later revised their A-meaning-to-B-meaning model, introducing the terms *focal sense* and *non-focal sense*. In this revised model, A and B at Stage III would be focal senses. At Stage I, B would only be a non-focal sense. It would be only an exploitation of the meaning A. The meaning A is supposed to have a set of conversational implicatures in addition to its core, partial pragmatic meanings which are triggered by various contexts. When a non-focal meaning B becomes highlighted as particularly suitable for expressing a given communicational purpose, it becomes more frequent and gradually gains its own set of conversational implicatures. Then it develops into to a new focal meaning B. B, undergoing grammaticalization, then generalizes even to contexts where formerly only A was accepted.

The revised model of *context-induced reinterpretation* implies the following: when determining the meaning of a grammatical entity, not only the focal meanings have to be observed, but also the conceptually prior non-focal meanings and recurring 'later' meanings likely to develop into new focal meanings must be recorded. Sentences 26 and 27 show the completed development from a volitional to a predictional future. When the structure *be going to* is used with an agentive subject, it typically has the

meaning of INTENTION: *I am going to draw this …so that he can have a full picture.*[24] As a result of the PERSON-OBJECT metaphorical transfer[25], the volitional future construction has been exploited in order to create a new convention, which implies future in events with non-human and non-agentive subjects. The evident conversational implicature is that non-human and non-agentive subjects do not activate the *will* feature in the future since they cannot pursue any will on their own: *It is going to be hot today* (PREDICTION). However, due to the generalization of the new interpretation, the PREDICTION-meaning is extendable back to sentences with agentive and human subjects: *We are going to have a new mum.* Here the structure *to be going to* is ambiguous since without the context or knowledge of the situation it is impossible to tell whether the speakers (potentially volitional) are planning to have a new mum or whether they are rather assuming that this will happen, no matter their will.

The context-induced reinterpretation appears to be the most interesting semantic change for a lexicographer seeking out "regularities which promise interest as incipient sub-systems" (Hopper, 1987). It has also been described in other words by Hanks (*exploitations of norms* in Hanks, forthcoming) as the result of a long-termed lexicographical work with authentic language data. The next section gives an example of a context-induced reinterpretation of a Swedish construction that normally expresses the progressive tense.

10.2. A Swedish Example: *hålla på*

The verb *hålla* enriched with the particle *på* is known to have grammaticalized uses. It has three valency patterns, in which the lexical verb is represented by the hypothetical verb *verba (to verb)*:

1. *X håller på med Y (Y = noun)* (lit. *X holds on with Y*)
2. *X håller på (med) att verb-a* (lit. *X holds on (with) to verb*)
3. *X håller på och verb-ar* (lit. *X holds on and verb-s*)

The progressive use approximately corresponds to the English gerund *to be verb-ing*. It is used for backgrounding events in the discourse and to indicate ongoing processes. Unlike the English gerund it is unacceptable with verbs denoting states and with verbs denoting transitions (see above). The progressive meaning is only activated in combination with atelic verbs. The combination with telic verbs yields the tendential meaning (see below). It can be used together with verbs in passive.

The progressive meaning can be rendered by *X håller på att verb-a* as well as by the coordinated construction *X håller på och verb-ar*. Pihlström observed speakers' preference for the coordinated construction, even though it had not yet been accepted as standard in the 80's. SAG does not comment on the respective variants' stylistic val-

[24] Heine et al. (2001, p. 171ff).

[25] The trasformation of volition into prediction can be seen as the transformation of *X wants* into *X wants to happen = X will happen*.

ues but adds the same observation. According to SAG, some speakers even make a sharp semantic distinction between the two variants in that they exclusively associate *X håller på att verb-a* with tendentiality and *X håller på och verb-ar* with progressivity. However, SAG mentions another tendency that goes against this semantic distinction: the coordinated construction is strongly preferred with animate agentive subjects although it is still considered odd with inanimate non-agentive subjects:

(33) *Klimatet* ***håller på att*** *bli varmare.*
?Klimatet ***håller på och*** *blir varmare.*
The climate is becoming warmer.

PAROLE contains only 118 instances of **X håller på och verb-ar**, out of which indeed only in one the subject is inanimate (a computer) but it is agentive:

(34) *och att en dator nu* ***höll på och*** *smälte svaren.*
and that then a computer was digesting (i.e. processing) the answers.

Progressivity marking is typical in telic verbal clauses in the past tense when the context indicates that the described event was prevented from reaching the expected terminal point (Teleman et al., 1999, p. 340):

(35) *Karin* ***höll på att tvätta håret*** *men blev avbruten.*
?Karin ***tvättade håret*** *men blev avbruten.*
Karin was washing her hair but was interrupted.

Interestingly, the construction *hålla på och* as well as *hålla på att* (though less frequently) appears to acquire the meaning *constantly* (which is a sort of an opposite to progressivity).

The parallel Czech-Swedish corpus has yielded one spectacular example. It is the Swedish translation of a text originally written by B. Hrabal in very colloquial Czech:

(36) *proto se taky náš farář musel* ***v jednom tahu*** *modlit, aby nebyl tak zlej…*
därför måste också vår präst ***hålla på och*** *be* ***stup i ett****, så att han inte skulle vara så elak …*
and that's why our priest had to be praying ***all the time*** *in order not to be so evil…*

The Swedish idiom *stup i ett* is perfectly equivalent to the Czech *v jednom tahu*. However, the translator added the *hålla på* construction partly to emphasize that the priest had been praying constantly or very often, but also as an indication of colloquial register[26].

More sentences containing a combination of *hålla på* and atelic verbs were sought in PAROLE, which might be bearing the semantic component of constancy. No unambiguous declarative sentence in the past tense has been found that would be acceptable without a disambiguating adverbial. Most hits (approx. 30) were propositions

[26]This assumption was confirmed by the translator in personal communication (Larsson, 2006).

with low facticity, i.e. negative sentences, sometimes with the imperative *ska (should, ought to)*, questions, and infinitives.

All the instances from PAROLE seem to be quotations of direct speech or free indirect speech[27], which suggests that this use of *hålla på* is still confined to spoken language. Exceptions will be discussed below.

Here are a few sentences from PAROLE in which *hålla på* could be substituted with *hela tiden (all the time):*

(37) *I princip tyckte hon det verkade botten att **hålla på och** knega mellan nio och fem .*
Basically she meant that it appeared miserable to keep working from nine to five.

(38) *Ni ska inte **hålla på och** larva er sådär, för jag har ingenting att skämmas för.*
You are not supposed to keep acting like this because I have nothing to be ashamed of.

(39) *Men i längden så kan vi ju inte **hålla på att** bara försvara oss.*
But for a longer time we can't just keep defending ourselves.

(40) *Är det slut? — Det vet jag inte heller. Varför ska du **hålla på och** fråga så där?*
Is that the end? – I don't know, either. Why do you keep interrogating me like this?

(41) *Men jag tyckte det var lika bra att vara kvar och inte **hålla på att** bråka.*
But I meant the best thing to do was to stay there and not to keep fighting.

PAROLE yields just one instance of a positive declarative sentence in the present tense, and, in this particular case, *hålla på* was disambiguated by temporal adverbials in the close context (cf. Example 36):

(42) *"Det var **alltid** bara som du inbillade dig." "Du förnekar det **fortfarande**. Det är otroligt." "Det är otroligt att du **fortfarande håller på och** ältar det. Jag gillade henne aldrig."*
*"You had just been fancying it". "You **still keep denying** it. It's incredible". "It's incredible that you **still keep agonizing** over that. I **never** liked her."*

How is it that a progressive construction has acquired just the opposite meaning? The progressive *hålla på* is the default interpretation of *hålla på* with atelic verbs. It appears in positive as well as in negative declarative clauses, questions etc., in all tenses. On the other hand, the 'constancy' *hålla på* seems to almost exclusively appear in negations, questions and infinitives (this is at least what the corpus evidence says). It is negation that gives a clue for the semantic change. In a negated progressive sentence, it is not just a single moment of the given event that is negated, but it is the

[27] Wikipedia (http://en.wikipedia.org): **"Free indirect speech** (or **free indirect discourse** or **free indirect style**) is a style of third person narration which has some of the characteristics of direct speech. Passages written using free indirect speech are often ambiguous as to whether they convey the views of the narrator or of the character the narrator is describing. Free indirect speech is contrasted with direct speech and indirect speech."

entire event. For instance, the sentence *De håller på att bråka (They are fighting)* focuses just on one moment in the ongoing action. The same goes for the progressive aspect as a discourse backgrounder: *De höll på att bråka när jag kom (They were just fighting when I arrived)*. However, the negation of the sentence predicate says that **the entire event** does not take place (at the moment of reference), not that **a single moment of the event** does not take place at the moment of reference. This is best perceived in the imperative; for instance, by saying: *Don't be doing*, the speaker necessarily means: "**Stop** doing that you have been doing just long enough to annoy me". Implicitly, the event really must have been taking place.

Nevertheless, the relation between progressivity and facticity also works the other way round: when the 'constancy' *hålla på* is employed in a negative imperative with an event, it suggests that the given event is actually taking place and should be stopped [28]. In written language, the reader naturally has no way to decide whether the given event is just taking place or not. By employing the 'constancy' *hålla på* the speaker adds some kind of asserting modality:

(43) *Vi **ska inte hålla på** och keynesianskt försöka mota konjunkturer. Vi ska bygga en robust arbetsmarknad*
och en stabil privat konsumtion som bägge ska klara att anpassa sig till chocker. [29]
We are not supposed to be reaching for conjunctures in a Keynesian fashion. We are supposed to build a robust labour market….

(44) *De latinamerikanska, asiatiska och afrikanska staterna **ska inte hålla på och**blanda sig i USA´s och Europas affärer hela tiden!*[30] *The Southamerican, Asian and African states should not permanently get involved in the USA's and Europe's affairs!*

In these examples, the speakers virtually underspecify the actual event(s). What they do instead is label them with expressions that are evaluative, with clear (here negative) connotations: *reach for conjuncturalisms instead of building a solid labour market*, *get involved in someone else's affairs without being invited*. The transition of the pro-

[28] According to Larsson (2006), the authentic sentence

(1) *Jag har sett som min uppgift att övertyga mitt eget folk om att vi inte kan hålla på och förtrycka ett annat folk.*
I have considered it to be my task to convince my own nation that we cannot keep suppressing another nation.

really assumes that the suppressing is taking place. It would be unacceptable to say

(2) **…att vi inte kan hålla på och förtrycka ett annat folk genom att börja bygga järnvägar på deras mark.*
**…that we cannot keep suppressing another nation by starting to build railways in their territory.*

[29] Quoted from Google, 2006-09-19, URL<http://forum.svt.se/jive/svt/report.jspa?messageID=84154>.

[30] Quoted from Google, 2006-09-19, URL<debatt.passagen.se/show.fcgi?category=3500000000000014&conference…>.

gressive *hålla på* into a 'constant' *hålla på* is a good example of a not yet completed context-induced reinterpretation. The focal sense A is clearly progressivity. The conversational implicature associated with A is 'X is happening at the time of reference'. Constancy is the non-focal sense B. The conversational implicature associated with B is 'X has been happening to the time of reference'. The corpus evidence suggests that the sense B is still bound to contexts where ambiguity is not likely to arise (negative statements, infinitives, questions and declarative clauses with disambiguating adverbials).

11. Organizing a Lexicon of Basic Verbs

Polysemy, various stages of grammaticalization, and the morphosyntactic variability of nouns in light verb constructions – this is what any lexical description of basic verbs must be especially sensitive to. A twin-lexicon is being proposed that seeks to cover these problem areas. It consists of a valency lexicon of verbs (SweVallex) and a lexicon of predicate nouns (Predicate Noun Lexicon). Methodologically, the verb lexicon draws on the Czech valency lexicon Vallex (Lopatková et al., 2007), which is based on the valency theory of the Functional Generative Description (Panevová, 1974; Panevová, 1975), but combined with Hanks's Corpus Pattern Analysis (Hanks and Pustejovsky, 2005) and enriched with Czech equivalents in context. Unlike Vallex, whose valency frames are defined syntactically, the SweVallex frame is defined by the Corpus Pattern Analysis (CPA) criteria, which take into account both the syntactic and the semantic description of the collocates. However, each complementation has been assigned a functor (semantic label used in the Functional Generative Description) and has been classified as obligatory or optional according to the Dialogue Test used in the valency theory of FGD. Each frame (here called *pattern* in terms of Corpus Pattern Analysis) consists of a *proposition* – the given verb in conjugated form, supplied with its valency complementations. Each Swedish proposition is accompanied by one or several equivalent Czech propositions.

The Predicate Noun Lexicon captures nouns acting as predicate nouns (nominal components of light verb constructions). It captures the light verb collocates of the respective predicate nouns and sorts them according to the Mel'čukian Lexical Functions (Mel'čuk, 1996). It describes the valency of the given predicate noun with each of the light verbs, respectively, and it provides information about its morphosyntactic preferences, such as determiner/modifier insertion options.

The structure of the proposed lexicon was motivated by the needs of an advanced Czech student of Swedish. There are numerous good monolingual Swedish lexicons (in the first place Svenskt Språkbruk (Clausén et al., 2003), which do not only explain the meaning of lexemes, but also describe their behaviour in context and partly their morphosyntactic restrictions (e.g. *used only with negation*). However, even Svenskt Språkbruk pays little attention to the morphosyntactic variation and to the modifying options in phrasemes and light verb constructions.

In addition, no monolingual dictionary can anticipate all contrastive issues that arise for learners with different native-language backgrounds. A nice example is the Swedish triple *sätta-lägga-ställa* versus the English *put* (something somewhere), where the Czech equivalent *dát (give)* has the same problem as English, namely being too unspecific in comparison to Swedish. It is extremely difficult to create lexicon definitions of these three respective Swedish verbs that would teach the non-native speaker to consistently choose the proper variant: the choice is based on the Swedish native conception of items as predominantly vertical vs. predominantly horizontal, or 'axis irrelevant', in connection with other aspects (whether the item must be fixed or whether it keeps its position by itself, etc. See Section 9.2, above).

The lavish exemplification of the *put*-like reading of *sätta* makes SweVallex resemble the clue page of a textbook exercise rather than a dictionary entry. The examples are simply chosen from a number of random concordances (in case of *sätta* some 2 000 of the total 9 000 concordances). Such concordances are preferred that appear surprising to the Czech speaker (e.g. *sätta en pil i*, since Czech requires a more specific verb than the equivalent of *put* (approximately *sting*), and so for a Czech speaker *put* is absolutely unpredictable in this context).

The lexicon is bilingual, with Czech being the target language. The Czech part includes just a minimal description of the Czech equivalents. This feature makes the lexicon more or less useless to a Swedish-speaking student of Czech. Creating a Swedish-Czech lexicon as a production-focused lexicon for Czechs can also seem as missing the point; apparently, the most straightforward way for the non-native Swedish text production would be using a reliable Czech-Swedish dictionary. However, production dictionaries 'atomize' the description of the source-language units according to their equivalents in the target language, such that the picture of the uses of one single Swedish word gets lost. This is also why advanced language learners prefer using monolingual dictionaries of the source language instead of bilingual dictionaries: a good monolingual dictionary seems to help draw a cognitive map' of the given lexeme. This map is a blending of semantic features and collocation options.

What production-oriented bilingual as well as monolingual dictionaries can easily miss is a target-language-specific forewarning for collocational as well as cognitive mismatches within the given language pair. There is a need for a description system that would capture the language traps explicitly – at least those based on morphosyntax and on collocability. Such a system is tested by a Czech-related description of cognitively and collocationally difficult Swedish verbs (basic verbs), which are so frequent that nobody can avoid them, and yet they are not fully explained in the teaching materials.

SweVallex-PNL is machine-readable, and its structuring allows for an automatic extraction of a Czech-Swedish glossary. The Czech glossary obtained by the extraction of the Czech equivalents of Swedish verb uses has the advantage of being fully Swedish-centered. If the lexicon was primarily designed as a Czech-Swedish dictionary, it would be Czech-centered: the cognitive map of each word would remain

Czech, and the Swedish equivalents would be chosen in a way that would disambiguate the respective Czech-centered readings of the given Czech word *('how do I say X in Swedish?')*.

As a result, among all the potential Swedish equivalents such Swedish equivalents would be intuitively selected, whose collocational preferences are not much wider than those of the Czech source word, and the commonest verbs (which are the vaguest) would be in danger of being omitted.

On the other hand, creating an ex-post Czech glossary from a Swedish-Czech lexicon allows the learner to avoid what John Sinclair (Sinclair, 1993) noticed long ago: learning rare words instead of using the less cognitively salient uses of the commonest words. A Swedish-Czech lexicon with a Czech glossary preserves the 'cognitive maps' of the Swedish words and can be used for learning more about one particular difficult (i.e. vaguely polysemous) verb, as well as it encourages the user to use these verbs in an idiomatic, native-speaker-like way.

The issue of sense disambiguation in bilingual dictionaries is very interesting, and the approach chosen varies from dictionary to dictionary. In each described word, there is a dilemma of whether the reading split is to be based primarily on differences in the collocational preferences in the source language, or rather on differences of the equivalents in the target language. SweVallex attempts at avoiding this dilemma by defining the respective readings by corpus patterns, enhanced with functors and the information on their obligatoriness. The internal structure of the entries is described in Sections 12 and 13. As a result, the Czech equivalents of one Swedish reading are not necessarily synonymous, as Fig. 3 illustrates.

[[Human, Device--]]ACT-obl ***sätter*** [[Physical Object --]]PAT-obl [[Location, Physical Object--]]DIR3-obl
(where it is meant to come, and the entity to be placed is not perceived as primarily vertical or primarily horizontal)
[[Human, Device--]]ACT-obl ***dá umístí usadí strčí zastrčí připevní přibije přilepí přišpendlí přišije přitiskne nasadí vloží přiloží zasune*** [[Physical Object--]]PAT-obl [[Location, Physical Object--]]DIR3-obl

Figure 3. Non-synonymous Czech equivalents

In sum, Swevallex-PNL was designed with respect to the following points:

1. to describe and explain a given Swedish lexeme in detail like a monolingual dictionary,
2. to provide the morphosyntactic and collocational preferences for each reading in form of a corpus pattern,
3. to determine the underlying valency frame of each Swedish corpus pattern,
4. to provide Czech equivalents and their patterns with valency frames,
5. to list phrasemes and indicate their variability options,

6. to pay special attention to light verb constructions and their morphosyntactic preferences with respect to the definiteness of predicate nouns,
7. to inform about the options of modifier insertion in light verb constructions, and
8. to provide enough examples from the corpus.

SweVallex as well as PNL are xml files with their respective document type definitions (DTD's) and CCS templates. The data was edited in the XMLMind editor (XMLmind). The CCS templates, although they may resemble dictionary entries, have no greater ambition but to facilitate the navigation through the data during the editing, and thus, this is to be emphasized, **they are not meant as the final layout for the users**. Creating the final layout, e.g. for a CD or web release, has never been the purpose of this study, which is a purely linguistic one.

Sections 12 and 13 analyze and explain the structures of both the lexicon parts, respectively.

12. SweVallex

12.1. Macrostructure

SweVallex is the lexicon of verbs. Its structure is to the greatest extent possible derived from the structure of Vallex 2.5 (Lopatková et al., 2007), the Czech verb valency lexicon. The major deviations from the Vallex 2.5 DTD are motivated by the adaptation to Swedish and by including a second language and the Corpus Pattern Analysis.

The lexicon Swevallex consists of elements `lexeme_cluster` nested in the root element `swevallex_verbs`. Lexeme clusters bring together verbs (elements `lexeme`) that are related by word formation, e.g. *sätta, sätta sig, värdesätta, sätta på*.

```
<?xml version='1.0' encoding='UTF-8'?>
<!ELEMENT swevallex_verbs (lexeme_cluster+)>

<!ELEMENT lexeme_cluster (lexeme+)>
<!ATTLIST lexeme_cluster
              cluster_id ID #IMPLIED
>
```

Each element `lexeme` has its unique ID. Each element `lexeme` contains the elements `lexical_forms` and `patterns`. Here Swevallex starts to differ from Vallex 2.5. *Patterns* is an element of the same level as `lu_cluster` in Vallex 2.5, but its function is different. Swevallex has **patterns** (like *Corpus Patterns*, instead of LU's (lexical units) introduced in Vallex 2.5. The element `lexeme` contains the actual lexicon entry.

```
<!ATTLIST lexeme
          lexeme_id ID #IMPLIED
>
<!ELEMENT lexeme (lexical_forms, patterns)>
```

12.2. Lemma

The element `lexical_forms` consists of a lemma (element `mlemma`), or a set of lemma variants (`mlemma_variants`). If the lemma is a homograph, it gets its homograph index. The past forms are listed for each lemma separately. Reflexive pronouns as well as particles are captured in the element `admorpheme`, which is optional and can be repeated. The element `admorpheme` has an obligatory attribute, which indicates its type. The values indicate whether the morpheme is a reflexive pronoun, or a particle. This solution was adopted due to the semantically relevant variability in their order – cf.: *ställa in sig* vs. *ställa sig in*.

```
<!ELEMENT lexical_forms ((mlemma| mlemma_variants),
          admorpheme*, constraints?) >
<!ELEMENT constraints (#PCDATA)>

<!ELEMENT mlemma (#PCDATA)>
<!ATTLIST mlemma
          homograph CDATA #IMPLIED
          preteritum CDATA #REQUIRED
          supinum CDATA #REQUIRED
>
<!ELEMENT mlemma_variants (mlemma+)>
<!ELEMENT admorpheme (#PCDATA)>
<!ATTLIST admorpheme
          type  (reflex| particle) #REQUIRED
>
```

12.3. Patterns

The element `patterns` consists of at least one element `pattern`. Apart from its unique ID, each element `pattern` carries the following information in the form of attribute values: is it an idiom or not? Is the form of the verb constrained for this particular pattern in any way (e.g. does it only occur in imperative)?

```
<!ELEMENT patterns (pattern+)>
<!ATTLIST pattern
          idiom (0| 1) #IMPLIED
          verb_form_constraints CDATA #IMPLIED
          pattern_id ID #IMPLIED
>
```

Each pattern consists of the following elements: `proposition`, `czech`, and `example`.

```
<!ELEMENT pattern (proposition, czech*, example*)>
```

The proposition is the Swedish corpus pattern. It has the form of a Swedish declarative sentence in the present tense (when possible), whose predicate is the lemma

verb. Its inner participants and free modifications are rendered by slots (element `slot`), integrated in the proposition (element `pattern_text`). Each piece of `pattern_text` has an attribute value according to whether it is the lemma verb or not. The proposition can finish with a (usually English) explaining gloss, which is called implicature (element `implicature`).

```
<!ELEMENT proposition (pattern_text| slot| implicature)*>
<!ELEMENT pattern_text (#PCDATA)>
<!ATTLIST pattern_text
          verb (1| 0) "0"
>
<!ELEMENT implicature (#PCDATA)>
```

Fig. 4 shows the proposition *sätta fart på något* in the sense of starting a motor. Note that the word *fart*, which is regarded as a predicate noun, is not explicitly present in the data, but it is referred to via a reference to PNL. The CCS template (in the picture) visualizes only the ID of the given predicate noun. For more details on the description of predicate nouns see Section 13.

Idiom:0
[[Human--driver]]ACT-obl ***sätter*** [[--]]CPHR-obl ***fart*** pnl_ref:fart-saetta-5 [[Car, Motorbike, Truck, Boat, Device--]]PAT-obl **på**

Figure 4. Swedish pattern (proposition)

The Czech equivalents are also presented in form of corpus patterns with slots, pattern text, and implications. When all the equivalents presented have the same corpus pattern, they are all placed in a row of the `pattern_text` elements with the attribute value `verb=1`. When an equivalent requires a different pattern, a new Czech pattern is created. Each Czech corpus pattern is classified according to whether it is an idiom or not and whether it really is an equivalent, or just a gloss (used in case there is a lexical gap in Czech).

```
<!ELEMENT czech (pattern_text| slot| implicature)*>
<!ATTLIST czech
          match (equivalent| gloss) #REQUIRED
          idiom (1| 0) "0"
>
```

Each pattern is accompanied by examples taken from PAROLE or (extremely rarely) from Konkordanser or Google. Examples are elements with free text. Sometimes, examples are shortened, but not consequently. In light verb constructions it is often the case that the examples even include some context.

```
<!ELEMENT example (#PCDATA)>
```

12.4. Slot

A lot of linguistic information is hidden in the complex internal structure of the slots. The slots have attributes and a nested element called `occupation`, which is present at least once per slot.

```
<!ELEMENT slot (occupation+)>
```

12.5. Surface Form

The element `occupation` carries the information about the surface form of the given slot; i.e., about prepositions, lemma, number, definiteness and other restrictions (this is important with very lexicalized collocations). Occupation can also be represented by a deliberate number of references to PNL (the optional and repetitive empty element `pnl_ref` with the obligatory attribute `ref`). The elements `slot` as well as `occupation` are common for both the Swedish and the Czech patterns. Some of the internal elements of `occupation` are therefore Swedish-specific, while others are Czech-specific, and some are common.

```
<!ELEMENT occupation ((surface_form| cz_surface)*, lexical?, pnl_ref*)>
<!ELEMENT pnl_ref EMPTY>
<!ATTLIST pnl_ref ref IDREF #IMPLIED>
<!ELEMENT surface_form EMPTY>
<!ATTLIST surface_form
          form (på| om| i| till| efter| från| framför| ifrån| för| av| med
  | utan| över| genom| att| vid) #IMPLIED
          case (basic| genitive) "basic"
>

<!ELEMENT cz_surface EMPTY>
<!ATTLIST cz_surface
          cz_form (bez| do| k| kolem| na| o| od| po| pro| před
  | s| u| v| vedle| z| za) #IMPLIED
          cz_case (1| 2| 3| 4| 6| 7) #REQUIRED


>
<!ELEMENT lexical (#PCDATA)* >
<!--text: word forms. Everything else should be in the attributes-->
<!ATTLIST lexical
          lemma CDATA #IMPLIED
          number CDATA #IMPLIED
          article CDATA #IMPLIED
          other_constraint CDATA #IMPLIED
>
```

12.6. FGD-Information

The element `slot` has two obligatory attribute values: functor and its obligatoriness according to the valency theory of the Functional Generative Description.

```
<!ATTLIST slot
          functor    (ACT| PAT| ADDR| EFF| ORIG| ACMP| ADVS| AIM| APP| APPS|
ATT| BEN| CAUS| CPHR| CNCS| COMPL| COND| CONJ|
CONFR| CPR| CRIT| CSQ| CTERF| DENOM| DES| DIFF|
                   DIR1| DIR2| DIR3| DISJ| DPHR| ETHD| EXT| FPHR| GRAD|
HER| ID| INTF| INTT| LOC| MANN| MAT| MEANS| MOD|
NA| NORM| PAR| PARTL| PN| PREC| PRED| REAS|
                   REG| RESL| RESTR| RHEM| RSTR| SUBS| TFHL| TFRWH|
THL| THO| TOWH| TPAR| TSIN| TTILL| TWHEN| VOC|
VOCAT| SENT| DIR| OBST| RCMP)    #REQUIRED

          obligatoriness (obl| opt| typ) #REQUIRED
>
```

12.7. CPA-Information

The information related to the Corpus Pattern Analysis is also contained in the slot. These attribute values are implied as the CPA is less formalized at this stage of the lexicon editing than the FGD-related part.

The attribute `sem_type` contains one or more instances from the current version of the ontology used in the Corpus Pattern Dictionary, which is being built by Hanks (Hanks and Pustejovsky, 2005).

```
        sem_type CDATA #IMPLIED
```

The attribute `lex_set` contains the lexical sets.

```
lex_set CDATA #IMPLIED
```

13. Predicate Noun Lexicon

13.1. Macrostructure

The Predicate Noun Lexicon (PNL) contains entries of nouns that occur as nominal components of light verb constructions. They are typically, but not necessarily, event nouns. Besides pure predicate nouns the lexicon also contains parts of phrasemes that exhibit morphosyntactic variability. These can be nominal components of phrasemes governed by a verb, as well as dependent parts of verbless phrasemes (e.g. *pris **på någons huvud***). Dependent parts of phrasemes governed by a noun have a simplified entry.

The root element of PNL is the element predicate_noun_lexicon, which consists of at least one element pred_noun_entry or at least one phraseme_entry in deliberate order.

```
<!ELEMENT predicate_noun_lexicon (pred_noun_entry+| phraseme_entry+)* >
```

13.2. Predicate Noun Lemma

The element pred_noun_entry displays the lemma, its possible homograph index, and the basic information about its gender and declension. As with the verb entries in SweVallex, variant lemmas (e.g. orthographic variants) are allowed.

```
<!ELEMENT lemma_variants (lemma)+>

<!ELEMENT lemma (#PCDATA)>
<!ATTLIST lemma
          lemma_id ID #IMPLIED
          homonym_index CDATA #IMPLIED
          genus (utrum| neutrum| NA| neutrum_utrum) #REQUIRED
          plural CDATA #REQUIRED
>
```

The introductory part of the entry is followed by up to three lists of typical adjectival and prepositional-group collocates of the given lemma, regardless the other context (elements adjectives and pps), and the most frequent compounds that occur with the given noun as the base (element compounds). Each item of the lists of collocates is surrounded with the nested element collocate.

```
<!ELEMENT adjectives (collocate+)>
<!ELEMENT compounds (collocate+)>
<!ELEMENT pps (collocate+)>
<!ELEMENT collocate (#PCDATA)>
```

13.3. Light Verb Unit

Like the verb entries were divided into patterns, the predicate noun entries are divided according to the combinations of the given predicate noun with a particular light verb (element light_verb).

```
<!ELEMENT pred_noun_entry ((lemma| lemma_variants),

adjectives?, compounds?, pps?, light_verb+)>
```

The light-verb unit consists of the optional element czech, which can have an unlimited number of instances, along with two optional elements that cannot be repeated: definiteness and pred_noun_slots.

```
<!ELEMENT light_verb (czech*, definiteness?, pred_noun_slots?)>
```

The element light_verb contains a lot of information in form of attribute values.

The lemma of the light verb occurring in the light verb construction described is to be filled in as the first attribute value.

```
<!ATTLIST light_verb
          lemma CDATA #REQUIRED
```

Each light verb construction in PNL has its unique ID:

```
 id_for_verbslot ID #REQUIRED
```

and it is classified by means of the Lexical Functions.

```
 basic_LF (Oper1| Oper2| Copul| Func| Labor1_2| Labor2_1| NA) #REQUIRED
          phasal_LF (Incep| Cont| Fin) #IMPLIED
          causative_LF (Caus| Perm| Liqu) #IMPLIED
          anti_LF (Anti) #IMPLIED
          prox_LF (Prox) #IMPLIED
```

In addition, three properties of the verb in its light-verb use are observed: telicity, punctuality, and volitionality:

```
 telicity (telic| atelic| NA) #IMPLIED
          punctuality (punctual| durative| NA) #IMPLIED
          volitionality (volitional| non-volitional| NA) #IMPLIED
>
```

The `NA` values stand for *non-applicable*, and they are selected when they depend on the context. The attribute `volitionality` describes whether or not the event denoted by the verb normally is a volitional action (regardless of the animacy and agentivity of the agent). The simplified entry for a dependent part of a phraseme does not contain the light-verb unit:

```
<!ELEMENT phraseme_entry ((lemma| lemma_variants),slot*)>
```

When the Czech equivalent is not given in the form of a corpus pattern within the verb entry in SweVallex, it is stated here. The Czech equivalents are obtained by a combination of introspection and searches in the Czech corpus SYN2005. They are nevertheless preferably captured in SweVallex. This element is much of an auxiliary element for editing noun entries that do not have their complements in SweVallex yet. As soon as they get a corresponding entry in SweVallex, the Czech equivalent gets the form of the corpus pattern and moves there.

```
<!ELEMENT czech (#PCDATA)>
```

13.4. Noun Definiteness, Modifier Insertion

Several parameters of noun definiteness are observed in the analysis of concordances of each light verb construction:

- noun with no determiner (element `bare_noun`)
- noun with the indefinite article (element `indef_art`)
- noun with the postpositive definite article (element `def_art_post`)
- noun with both the prepositive and the postpositive definite article (element `def_art_prepost`)

- noun determined by a genitive or by a possessive pronoun (element `posgen_determiner`)
- noun determined by other non-article determiner (element `other_determiner`)

When an option is clearly predominant or, conversely, extremely rare, it is indicated by a note. When some option does not occur at all in the concordances (or there are just few concordances and they are dubious), the entire element is omitted. Each option is documented by examples. The number of the examples is not necessarily proportional to the ratio of the given option in the concordances. On the contrary, more attention is paid to the less represented options: the examples tend to be longer in context in order to make it possible for the user to find out more about its motivation (e.g. markedness in the information structure, coreferential reasons, etc.). Hypotheses about the motivation of a rare pattern, when any, are formulated in the element `note`. The examples also contain implicit information about the option of the insertion of adjectival and prepositional modifiers.

```
<!ELEMENT definiteness
(bare_noun?, indef_art?, def_art_post?, def_art_prepost?,
          posgen_determiner?, other_determiner?)>
<!ELEMENT example (#PCDATA)>
<!ELEMENT note (#PCDATA)>

<!ELEMENT bare_noun (example| note)*>
<!ELEMENT indef_art (example| note)*>
<!ELEMENT def_art_post (example| note)*>
<!ELEMENT def_art_prepost (example| note)*>
<!ELEMENT posgen_determiner (example| note)*>
<!ELEMENT other_determiner (example| note)*>
```

13.5. Slot

The last unit in the PNL entry is the slot. It has a similar structure as in SweVallex: the attributes `functor` and `obligatoriness` and the element `occupation`. Unlike in SweVallex, obligatoriness is not an obligatory attribute in PNL, as the complementations are regarded as optional by default. The attribute obligatoriness is primarily used to mark surface obligatoriness of modifiers in multi-word phrasemes; e.g. *på **rätt/fel** spår, pris på **någons** huvud*.

```
<!ELEMENT pred_noun_slots (slot*)>

<!ELEMENT slot (occupation*)>

<!ATTLIST slot
functor   (ACT| PAT| ADDR| EFF| ORIG| ACMP| ADVS| AIM| APP| APPS|
ATT| BEN| CAUS| CPHR| CNCS| COMPL| COND| CONJ|
```

```
CONFR| CPR| CRIT| CSQ| CTERF| DENOM| DES| DIFF|
            DIR1| DIR2| DIR3| DISJ| DPHR| ETHD| EXT| FPHR| GRAD|
HER| ID| INTF| INTT| LOC| MANN| MAT| MEANS| MOD|
NA| NORM| PAR| PARTL| PN| PREC| PRED| REAS|
            REG| RESL| RESTR| RHEM| RSTR| SUBS| TFHL| TFRWH|
THL| THO| TOWH| TPAR| TSIN| TTILL| TWHEN| VOC|
VOCAT| SENT| DIR| OBST| RCMP)   #REQUIRED
obligatoriness (obl| opt| typ) #IMPLIED
>


<!ELEMENT occupation (surface_form, lexical, cpa, example*, ref*)>

<!ELEMENT lexical (#PCDATA)>
<!ATTLIST lexical
          lemma CDATA #IMPLIED
          number CDATA #IMPLIED
          article CDATA #IMPLIED
          other_constraint CDATA #IMPLIED

>

<!ELEMENT ref EMPTY>
<!ATTLIST ref ref IDREF #IMPLIED>

<!ELEMENT cpa EMPTY>
<!ATTLIST cpa
          sem_type CDATA #IMPLIED
          lex_set CDATA #IMPLIED
          implicature CDATA #IMPLIED
>


<!ELEMENT surface_form EMPTY>
<!ATTLIST surface_form
          form (possgen| hos| på| om| i| till| från| för| av|
          med| utan| över| genom| att| vid) #IMPLIED>
```

14. Linking

The SweVallex-PNL lexicon comprises two parts: SweVallex, which captures verbs and their patterns, and nouns and the valency frames they have in connection with

the respective light verbs with which they combine. Apart from that, PNL captures all multi-word idioms, whose structure is too complex to be described by the SweVallex pattern system. References go currently from SweVallex to PNL (Fig. 5), or from one PNL light-verb frame to another PNL light-verb frame. Lemmas and patterns/light verb frames have their ID's in both lexicons, such that more relations among and within the entries can be displayed in the future.

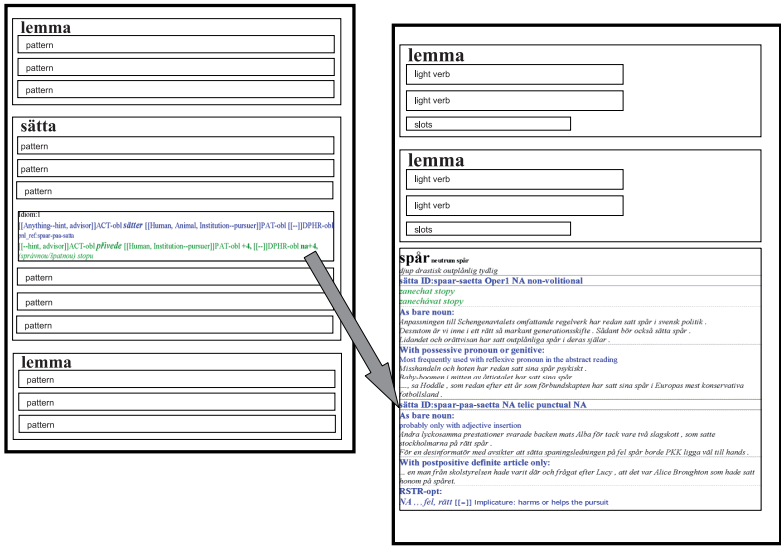


Figure 5. Reference from a pattern of sätta in SweVallex (left) to the relevant light-verb frame of spår in PNL (right)

15. Conclusion

A close corpus-based observation of basic verbs has resulted in a sketch of a learner's dictionary that would systematically and comprehensibly capture the trickiest issues of the basic verb use in Swedish. A number of linguistic theories as well as formal language description methods were critically examined, and their best features have been combined. The resulting structure is XML-based and enhanced with a simple CCS template to facilitate editing. It was tested and continuously adjusted on real corpus data.

Acknowledgements

This work was funded in part by the Companions project[31] sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434, by FP7-ICT-2007-3-231720 (EuroMatrix Plus), and by the grant of the Czech Ministry of Education No. 0021620823.

Corpora Used

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium[32].

Czech National Corpus[33] (Český národní korpus) – SYN2005. Ústav Českého národního korpusu FF UK, Praha 2005.

InterCorp[34]. Ústav Českého národního korpusu FF UK, Praha 2009.

Språkbanken vid Göteborgs universitet[35].

Corpus of Contemporary American English (COCA)[36]. Brigham Young University, Provo, Utah. Mark Davies

Bibliography

Allén, Sture. *Tiotusen i topp*. Almqvist & Wiksell, Stockholm, 1972.

Baron, Irène and Michael Herslund. Support Verb Constructions as Predicate Formation. In *The Structure of the Lexicon in Functional Grammar*. John Benjamins, Amsterdam/Philadelphia, 1998.

Bjerre, Tavs. Event Structure and Support Verb Constructions. In *Proceedings of the ESSLLI Student Session 1999*, 1999.

Boje, Frede. Hvor finder man *finde anvendelse*? In Ásta Svavarsdóttir, Guðrún Kvaran, and Jón Hilmar Jónsson, editors, *Nordiske Studier i Leksikografi Rapport fra Konferanse om leksikografi i Norden, Reykjavík 7.-10. juni 1995*, volume 3 of *Skrifter utgitt av Nordiske forening for leksikografi*, pages 51–68, Reykjavík, 1995.

Butt, Miriam. The Light Verb Jungle. *Harvard Working Papers in Linguistics*, 9(Papers from the Harvard/Dudley House Light Verb Workshop), 2003. URL `http://ling.uni-konstanz.de/pages/home/butt,quoted2007-01-19`.

Bybee, Joan. *Morphology: a study of the relation between meaning and form*, volume 9 of *Typological Studies in Language*. John Benjamins, Amsterdam/Philadelphia, 1985.

[31] `http://www.companions-project.org`

[32] `http://www.natcorp.ox.ac.uk/`

[33] `http://www.korpus.cz`

[34] `http://www.korpus.cz`

[35] `http://spraakbanken.gu.se`

[36] `http://www.americancorpus.org/`

Bybee, Joan, Revere Perkins, and William Pagliuca. *The Evolution of Grammar. Tense, aspect, and modality in the languages of the world*. The University of Chicago Press, Chicago & London, 1994.

František Čermák. Abstract Nouns Collocations: Their Nature in a Parallel English-Czech Corpus. In *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Barnbrook, Danielsson & Mahlberg, Birmingham, 1995.

Cinková, Silvie. Extraction of Swedish Verb-Noun Collocations from a Large Msd-Annotated Corpus. *The Prague Bulletin of Mathematical Linguistics 82*, pages 99–102, 2004.

Clausén, Ulla et al. *Svenskt Språkbruk – ordbok över konstruktioner och fraser*. Norstedts Ordbok and Svenska Språknämnden, 2003.

Dura, Ela. *Substantiv och stödverb*, volume 18 of *Meddelanden från Institutionen för Svenska Språket*. Göteborgs universitet, 1997.

Ekberg, Lena. *Gå till anfall och falla i sömn. En strukturell och funktionell beskrivning av abstrakta övergångsfaser*, volume A 43 of *Lundastudier i nordisk språkvetenskap*. Lund University Press, Lund, 1987.

Ekberg, Lena. Verbet *ta* i metaforisk och grammatikaliserad användning. *Språk och Stil*, 3: 105–139, 1993.

Fenyvesi-Jobbágy, Katalin. Non-literal and non-metaphorical uses of Danish komme 'come': A case study. *Jezikoslovlje*, 2003.

Fillmore, Charles J., Christopher R. Johnson, and M. L. R. Petruck. Background to FrameNet. *FrameNet and Frame Semantics. International Journal of Lexicography – Special Issue*, 16:235–250, 2003.

Fontenelle, Thierry. Co-occurence Knowledge, Support verbs and Machine Readable Dictionaries. In *Papers in Computational Lexicography, COMPLEX'92, Budapest*, 1992.

Günther, Heide and Sabine Pape. Funktionsverbgefüge als Problem der Beschreibung komplexer Verben in der Valenztheorie. In Schumacher, Helmut, editor, *Untersuchungen zur Verbvalenz: eine Dokumentation über die Arbeit an einem deutschen Valenzlexikon*, Forschungsberichte/Institut für deutsche Sprache Mannheim, pages 92–128. Narr, Tübingen, 1976.

Hajič, Jan. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, 2004. ISBN 80-246-0282-2.

Hanks, Patrick. *Norms and Exploitations: Corpus, Computing, and Cognition in Lexical Analysis*. MIT Press, forthcoming. Manuscript, obtained 2003 from the author.

Hanks, Patrick and James Pustejovsky. A Pattern Dictionary for Natural Language Processing. *Revue Francaise de linguistique appliquée*, 10(2), 2005.

Hanks, Patrick, Anne Urbschat, and Elke Gehweiler. German Light Verb Constructions in Corpora and Dictionaries. *International Journal of Lexicography – Special Issue: Corpus-Based Studies of German Idioms and Light Verbs*, 19(4):439–458, 2006.

Hansen, Erik. *Stå, sidde, ligge. Mål & Mæle*, 1(2):26–32, 1974.

Heid, Ulrich. Towards a Corpus-based Dictionary of German Noun-verb Collocations. In Fontenelle, Thierry, Philippe Hilligsmann, Archibald Michiels, AndréMoulin, and Siegfried Theissen, editors, *Actes EURALEX'98 Proceedings*, volume 1, pages 301–312, Liège, 1998. Universitéde Liège, Départements d'anglais et de néerlandai.

Heine, Bernd, Ulrike Claudi, and Friederike Hünnemeyer. *Grammaticalization. A Conceptual Framework*. University of Chicago Press, 2001.

Helbig, Gerhard and Joachim Buscha. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Verlag Enzyklopädie, Leipzig, 1996.

Hoey, Michael. Introducing applied linguistics: 25 years on. In *31st BAAL Annual Meeting: Languages and Literacies*. University of Manchester, 1998.

Hopper, Paul. Emergent Grammar. *Berkeley Linguistics Conference (BLS)*, 13:139–157, 1987. URL `http://eserver.org/home/hopper/emergence.html`.

Hunston, Susan. *Patterns of Text: In Honour of Michael Hoey*, chapter Colligation, Lexis, Pattern, and Text. John Benjamins, 2001.

Jakobsson, Ulrike. Familjelika betydelser hos STÅ, SITTA och LIGGA. En analys ur den kognitiva semantikens perspektiv. Technical report, Lunds universitet. Institutionen för nordiska språk, Lund, 1996.

Jelínek, M. O verbonominálních spojeních ve spisovné češtině. In *Přednášky a besedy z XXXVI. běhu LŠSS*, pages 37–51. MU Brno, 2003.

Jensen, Torben Juel. Kan man 'ligge' i et mentalt rum? In *Nydanske studier & almen kommunikationsteori. Artikler om partikler.* , pages 73–100. Københavns Universitet. Institut for Nordisk Filologi, København, 2000.

Jespersen, Otto. *A Modern English Grammar on Historical Principles*, volume 6. London: George Allen & Unwin & Copenhagen: Ejnar Munksgaard., 1954.

Lakoff, George. *Women, Fire and Dangerous Things. What categories reveal about the mind*. Chicago University Press, Chicago, 1987.

Larsson, Mats. translation of B. Hrabal's text "Taneční hodiny pro starší a pokročilé". personal communication, September 18 2006.

Lopatková, Markéta, Zdeněk Žabokrtský, Václava Kettnerová, Karolina Skwarska, Eduard Bejček, Klára Hrstková, Michaela Nová, and Miroslav Tichý. VALLEX 2.5 – Valency Lexicon of Czech Verbs, version 2.5. Software prototype, 2007.

Malmgren, Sven-Göran. *Begå eller ta självmord? Om svenska kollokationer och deras förändringsbenägenhet 1800-2000*. Rapporter från ORDAT. Göteborgs universitet. Institutionen för svenska språket., Göteborg, 2002.

Mel'čuk, Igor A. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In Wanner, Leo, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–105. John Benjamins, Amsterdam/Philadelphia, 1996.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description I. *Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description II. *Prague Bulletin of Mathematical Linguistics*, 23:17–52, 1975.

Persson, Ingemar. *Das System der kausativen Funktionsverbgefüge. Eine semantisch-syntaktische Analyse einiger verwandter Konstruktionen*. Liber, Malmö, 1975. PhD thesis.

Persson, Ingemar. Das kausative Funktionsverbgefüge (FVG) und dessen Darstellung in der Grammatik und im Wörterbuch. *Deutsche Sprache*, 20:153–171, 1992.

Pihlström, Sven. *Hålla på* och *hålla på och*. *Språkvård*, 2:8–10, 1988.

Polenz, Peter von. Funktionsverben im heutigen Deutsch. *Sprache in der rationalisierten Welt. Wirkendes Wort*, Beiheft 5, 1963.

Pustejovsky, James. The Syntax of Event Structure. *Cognition*, 41:47–81, 1991.

Reuter, Mikael. Lägg ribban högt. *Reuters Ruta*, Forskningscentralen för de inhemska språken 1986. URL `http://www.kotus.fi/svenska/reuter/Kotimaistenkieltentutkimuskeskus.quoted2003-04-16`.

Rothkegel, Annely. *Feste Syntagmen. Grundlagen, Strukturbeschreibung und automatische Analyse*. Linguistische Arbeiten. Niemeyer, Tübingen, 1973.

Schroten, Jan. Light Verb Constructions in Bilingual Dictionaries. In *From Lexicology to Lexicography*, pages 83–94. University Utrecht. Utrecht Institute of Linguistics OTS., Utrecht, 2002.

Sinclair, John. *Corpus, Concordance, Collocation*. Oxford University Press, 1993. 3rd edition, 1st edition 1991.

Spoustová, Drahomíra, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*, pages 67–74, Praha, Czechia, 2007. Association for Computational Linguistics. ISBN 978-1-932432-88-6.

Teleman, U., S. Hellberg, and E. Andersson. *Svenska Akademiens grammatik*. Svenska Akademien/Norstedts, Stockholm, 1999.

Viberg, Åke. Svenskans lexikala profil. *Svenskans beskrivning*, 17:391–408, 1990.

Wray, Alison. *Formulaic Language and the Lexicon*. Cambridge University Press, 2002.

XMLmind. XMLmind XML Editor Personal Edition 3.7.0. Free software version, 2000–2007. URL `www.xmlmind.com/xmleditor/`.

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 63–83

CzEng 0.9

Large Parallel Treebank with Rich Annotation

Ondřej Bojar, Zdeněk Žabokrtský

Abstract

We describe our ongoing efforts in collecting a Czech-English parallel corpus CzEng. The paper provides full details on the current version 0.9 and focuses on its new features: (1) data from new sources were added, most importantly a few hundred electronically available books, technical documentation and also some parallel web pages, (2) the full corpus has been automatically annotated up to the tectogrammatical layer (surface and deep syntactic analysis), (3) sentence segmentation has been refined, and (4) several heuristic filters to improve corpus quality were implemented. In total, we provide a sentence-aligned automatic parallel treebank of about 8.0 million sentences, 93 million English and 82 million Czech words. CzEng 0.9 is freely available for non-commercial research purposes.

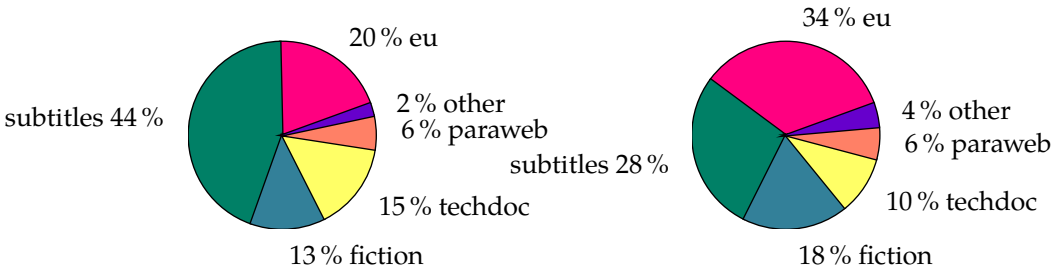
1. Introduction

Parallel corpora are essential for the training of (statistical) machine translation (MT) systems and used in other NLP tasks as well, e.g. language learning tools or terminology extraction. In the paper accompanying the previous release of CzEng (Bojar et al., 2008a), we confirmed that larger datasets usually improve the quality of MT, even if the additional data are out of the translated domain.

Some approaches to MT make use not only of large data but also of data (automatically) annotated: morphologically tagged and syntactically analyzed at a surface or a deep syntactic layer of linguistic description.

CzEng 0.9 is an extension of the previous release in both respects: we add data from several large sources like e-books and technical documentation and we use TectoMT (Žabokrtský et al., 2008) to augment the whole corpus with Czech and English automatic analyses at the morphological, analytical (surface syntactic, labelled "a-" in the sequel) and tectogrammatical (deep syntactic, labelled "t-") layers of description,

Please cite this article as: Ondřej Bojar, Zdeněk Žabokrtský, CzEng 0.9: Large Parallel Treebank with Rich Annotation. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 63–83.



(a) Sentences (b) Tokens

Figure 1. Types of parallel texts in CzEng 0.9. The depicted proportions are derived (a) from the number of included 1-1 sentence pairs, and (b) from the number of tokens (words and punctuation marks, summed for both languages).

following Functional Generative Description (Sgall, 1967; Sgall et al., 1986) and the Prague Dependency Treebank (Hajič et al., 2006).

Section 2 gives a detailed description of individual data sources included into CzEng 0.9. In Section 3, we briefly mention a general technique for fast semi-manual improvements when working with large data. The technique is then illustrated at several steps of corpus preparation, as described in Sections 4 (conversion to plain text), 5 (sentence segmentation and alignment) and 6 (automatic annotation up to the t-layer). Technical details such as sentence shuffling, the corpus structure, output file formats and corpus data size are given in Section 7 followed by the conclusion in Section 8.

2. Sources of Parallel Texts

This section gives an overview of all types of parallel text resources exploited in CzEng 0.9. The corpus is not claimed to be intentionally balanced in any sense—we simply collected as much material as possible. However, the set of covered topics is quite broad, with style ranging from formal language of laws and technical documents through prose fiction and journalistic language to colloquial language as often appearing in movies.

The proportions of the individual types of texts, which are included into CzEng 0.9, are roughly illustrated in Figure 1; detailed information is given later in Table 3. Note the difference in proportions calculated based on parallel sentences and based on words in one of the two languages.

For building CzEng 0.9, we used exclusively texts that were already publicly available in an electronic form, in most cases downloadable from the Internet. We did not do any book scanning or any other digitization activity.

2.1. Movie and Series Subtitles (`subtitles`)

Thanks to the community of movie fans, there is a huge amount of movie and series subtitles easily downloadable from several Internet subtitle archives.[1] More details about the cleaning of the data from this resource can be found in Beňa (2009), here we summarize the document alignment procedure and describe some newly implemented cleaning scripts.

As the movie/series subtitles stored in the two Internet archives were created by hundreds or thousands of contributors, one can hardly expect them to follow any strict naming conventions. First, we perform a filename normalization to represent only the following in the filename:

- the original movie/series name (from which determiners, prepositions, conjunctions and special characters were removed),
- the production year,
- the language of the subtitles (automatically detected from the file content),
- and also the series and episode numbers in the case of series.

Such normalization was reasonably reliable for de-duplicating and document-level alignment of movie subtitles, but it led to a large loss of the data in the case of series because there were too many irregularities in their original naming (or the information about episode/series number was completely missing). As mentioned in Bojar et al. (2009) an additional document-matching technique was used. Within each series, all beginning and end segments of all unpaired English subtitle files were compared with those of unpaired Czech files. The adequacy of such pairings was evaluated using a simple scoring function making use of a probabilistic translation dictionary. Then the pairs whose score was above an empirically found threshold were added into CzEng.

A number of filtering and cleaning scripts were implemented for the subtitle data, as their quality was very unstable: some authors systematically write "I'II" instead of "I'll", some others leave long passages untranslated, disregard punctuation, or disregard Czech diacritics, etc. Unless the errors were fixable with a very high reliability, we generally tend to throw out files with such a suspicious content.

Even if the subtitle data contains the highest amount of noise compared to the other sources of parallel texts, we still believe it is a valuable source because a lot of conversational phrases and colloquial language appear in them which would be difficult to find elsewhere. Moreover, the vocabulary distribution in the subtitle data probably better fit the real everyday language than e.g. European law does.

[1] CzEng 0.9 used `http://www.opensubtitles.com/` and http://www.titulky.com/.

2.2. Parallel Web Sites (`paraweb`)

Web sites with multilingual content can be an excellent source of parallel texts. For the most promising sites, it is worth implementing specialized crawlers and cleaners (and we do this for Project Syndicate, Section 2.6.1 and CzechNews, Section 2.6.3). However, we also wish to exploit the vast numbers of smaller sites.

Klempová et al. (2009) implement and evaluate a pipeline of tools that start with a few queries to search engines such as "lang:en česky" to obtain pages in English containing the Czech word for Czech from Google. Klempová et al. then crawl the whole web sites and use a combination of page structure and lexical content similarity to find parallel documents. In our current implementation, we apply a considerably simpler approach of aligning documents based on their URLs only.

Klempová et al. (2009) mention that it is surprisingly difficult to get large lists of candidate sites due to built-in limits on number of results available from search engines. We are grateful to Seznam[2], the largest Czech search engine, for an older version of all URLs of Czech Internet they index. We selected all domain names where the URLs contained a pattern indicating Czech or English language tag (e.g. "?lang=cs") and re-crawled the domains using our own crawler that specifically downloads only pages whose URL contains the language tag.

In addition to the selection of pages, we use the language tag also to find the document alignment. We have a short list of typical language tags for Czech and their variants for English, e.g. the above mentioned "?lang={cs,en}", implemented as regular expression substitution patterns. Given a website, we search the list of URLs of all documents of the website and apply the substitution. If the substitution can be applied and the resulting URL also exists, we promote the substitution pattern by a point. The highest scoring substitution pattern and the alignment of document it implies is then chosen for the given site.

Admittedly, we do not exploit the full potential of available parallel web pages: we require the pages to contain a language tag and the parallel version to differ only in the tag itself (and not e.g. the translation of the words in the URL). The advantage of aligning URLs only is the little computational cost and a relatively high accuracy.

2.3. Fiction (`fiction`)

2.3.1. E-books from Web (`ebooks`)

In the Internet, one can find a number of e-book archives such as Project Gutenberg[3] for English and Palmknihy[4] for Czech. We exploited such sources by downloading either e-book catalogs or directly the e-books files. Similarly to the case of subtitles

[2] `http://www.seznam.cz/`

[3] `http://www.gutenberg.org/`

[4] `http://www.palmknihy.cz/`

(Section 2.1), different e-book resources provided us with different metadata, so some metainformation normalization was necessary. We converted the information about the roughly 38,000 available e-books into a uniformly formatted catalog, whose entries contained

- normalized name of the book author: lowercased surname and the first letter from the first name; special rules for unifying transliteration variants (Tolstoj/Tolstoy) were applied,
- normalized book title,
- language (Czech or English),
- list of sources the book is available from.

Then the document-level alignment phase came. For each author, for whom the catalog contained at least one Czech book and at least one English book, all possible Czech-English book pairs were automatically scored. The heuristic scoring function took Czech and English titles as its input and produced a real number (weighted sum of several features) as the output. The features were based on the length similarity of the title strings, string similarity of the individual word pairs, translation probability of the individual word pairs, prefix similarity of the individual word pairs, etc. For each author, a list of Czech-English book pairs (whose score was above a certain threshold) as well as lists of remaining unpaired Czech and unpaired English books were generated. The weight and threshold values were optimized semi-automatically in several iterations, using a sample of roughly 20 authors with known book pairing.

The alignment algorithm identifies around 449 possible book pairs for 271 authors. This list was checked manually. Wrong pairs, duplicated pairs, and pairs containing poetry or dramas were excluded. 157 book pairs were confirmed as correct, and additional 102 new pairs were manually found among the unpaired books. No surprise that the simple title alignment approach did not reveal many book pairs such as in the case of Jules Verne's "Michel Strogoff" whose Czech title is "Carův kurýr" (Tsar's messenger).[5]

The e-book data, as acquired from the various archives, were stored in a highly diverse set of file formats. The need for format conversion leads to another data loss, as discussed in Section 4.

2.3.2. Kačenka corpus (`kacenka`)

Kačenka (Rambousek et al., 1997) is a Czech-English parallel corpus created by the Department of English, Faculty of Arts, Masaryk University in Brno in 1997.[6] It con-

[5] Of course, even such books could have been automatically paired supposing we already had their full texts in hand, but that was not always the case, as from some web archives it is not possible to download all books at once. That is why we performed the title-based alignment first and only then selectively downloaded the paired books.

[6] `http://www.phil.muni.cz/angl/kacenka/kachna.html`

tains texts of 12 English books and their Czech translations. The texts were manually aligned at the sentence level; this alignment has been preserved in CzEng 0.9.

All books contained in Rambousek et al. (1997) have been used when compiling CzEng 0.9. If a book pair appeared both in Rambousek et al. (1997) and in other e-book resources (Section 2.3.1), only the Rambousek et al. (1997) version was used.

2.3.3. Reader's Digest (`rd`)

Prague Czech-English Dependency Treebank (Cuřín et al., 2004) contains a parallel corpus composed of raw texts of 450 articles from the Reader's Digest, years 1993-1996, and their Czech translations.

2.4. European Union Law (`eu`)

2.4.1. JRC-Acquis (`celex`)

JRC-Acquis is a freely available parallel corpus containing European Union documents mostly of legal nature (Ralf et al., 2006).[7] It is available in 20 official EU languages. The corpus is encoded in XML, and contains roughly 8,000 millions documents per language.

We included into CzEng 0.9 all Czech-English documents pairs available in JRC-Acquis v.3.0 whose length ratio measured in characters was not too far from 1—within the interval $[1.4^{-1}; 1.4]$. If their length ratio was outside the interval, an attempt at extracting at least some parts of the documents was made: both documents were decomposed into head, body, signature and annex parts, and at least some corresponding parts were extracted if their length was inside the given interval. The motivation for this step was the following: in some cases the Czech version of the documents does only refer to the annex of the English version instead of containing the proper translation of the annex. If such document pairs are automatically sentence aligned, they might be rejected by the aligner (see Section 5.2) as a whole as they seem to be too much different, while if only their reasonably similar subparts are extracted, the chance for a successful sentence alignment grows.

2.4.2. The European Constitution proposal (`euconst`)

The European Constitution proposal from the OPUS corpus (Tiedemann and Nygaard, 2004).

[7] `http://wt.jrc.it/lt/acquis/`

2.4.3. Samples from the Official Journal of the European Union (`eujournal`)

Samples from the Official Journal of the European Union, which is a tiny collection of some rather randomly chosen issues of the the Official Journal of the European Union.

2.5. Technical documentation (`techdoc`)

2.5.1. KDE and GNOME documentation (`kde`, `gnome`)

KDE and GNOME are two most popular graphical user interface for running Linux. Both of them are open-source software projects and for both of them their Czech localizations (product translations) are available on the Internet.[8,9]

2.5.2. Microsoft glossaries (`microsoft`)

Microsoft glossaries are lists of technical terms and longer expressions and messages used e.g. in Microsoft software products. The glossaries are available for a number of languages. They are intensively used by technical translators as they constitute a rich resource of technical vocabulary. The glossaries are publicly available from the Microsoft Corporation FTP Server and its mirrors.

2.6. News texts (`news`)

2.6.1. Project Syndicate (`syndicate`)

Project Syndicate is a not-for-profit institution which currently consists of 432 newspapers in 150 countries.[10] There is a large number of newspaper articles available on its web pages, many of them existing in more language versions. Those articles that were available in English and Czech in August 2009 were used for the creation of CzEng 0.9.

2.6.2. Wall Street Journal (`wsj`)

Prague Czech-English Dependency Treebank (PCEDT, Cuřín et al. (2004)) contains English texts of Wall Street Journal articles adopted from the Penn Treebank (Marcus et al., 1993), and their Czech translations created (by human translators) specifically for the PCEDT needs.

[8] `http://www.gnome.org/projects/`

[9] `http://l10n.kde.org/`

[10] `http://www.project-syndicate.org/`

2.6.3. Czech News (`czechnews`)

The Czech news portal Aktualne.cz provides a limited selection of the news in English[11]. We implemented a custom crawler and we align the documents on the basis of links back to the Czech version available in the translated page.

2.7. User-Contributed Translations from Navajo (`navajo`)

Navajo[12] is a machine-translated Czech version of (the English content of) Wikipedia, which is a highly popular, multilingual, web-based, free-content encyclopedia. Similarly to Wikipedia, which is written and improved collaboratively by volunteers, also the content of Navajo is gradually improved by a community of volunteers who submit human-corrected translations of the individual entries. Such user-contributed Czech translations paired with their original English counterparts can be treated as a relatively reliable source of parallel texts, whose main advantage is a wide range of topics.[13] Therefore we include them into CzEng 0.9 too.

3. General Approach to Fixing Errors

Throughout the processing pipeline, we feel that the most successful correction steps are implemented using the following generic approach:[14]

1. We extend the tool in question or one of the subsequent tools to include a simple detector of suspicious positions in the corpus. We also try to automatically propose one or more possible corrections or solutions of the assumed problem.
2. We manually scan and quickly confirm or deny individual proposed solutions, e.g. by adding a prefix to each line in a text file. We carefully preserve old annotations to avoid duplicating manual effort.
3. The tool in question is extended to use the file of confirmed annotations and apply the corrections. For input with no confirmed or denied annotations, suspicious occurrences are still collected.

The main advantage of the setup is the excellent trade-off between manual labor and overall output quality. If new data are added, we can quickly add decisions for new suspicious cases. When rebuilding the whole corpus, old decisions are simply reused.

Another great advantage is the possibility to sort automatic suggestions by various criteria, such as the expected reliability (and thus little effort needed to confirm or deny a rule) or overall frequency. With time constraints on manual annotation, we can thus focus on some most important subset of the errors and leave others unsolved.

[11] `http://aktualne.centrum.cz/czechnews/`

[12] `http://www.navajo.cz/`

[13] As evaluated in Bojar et al. (2008b), about 70% of segment pairs are of reasonable quality.

[14] A similar approach proved fruitful in the pre-release corrections of PDT 2.0 (Štěpánek, 2006).

In our complex pipeline, we often take advantage of more elaborate information available in subsequent processing steps. One of the best examples is automatic suggestion of missed and superfluous sentence boundaries based on sentence alignment between Czech and English.

We used the approach in the following tasks:

- language guess based on book title, confirmed later, after the book is converted to plain text using the vocabulary of the book
- book alignment based on book titles, confirmed later by the quality of sentence-level alignment
- automatic detection and removal (upon confirmation) of page breaks and page numbers
- sentence segmentation, corrected later by sentence-level alignment

4. Handling Various Input Formats

4.1. Format Convertors

We implemented a generic wrapper of several tools to convert many file formats (pdf, doc, rtf, pdb, html, txt and also archive-like formats lit, zip and rar) to plain text encoded in UTF-8 and attempting to identify documents with malformed encoding.

Our handling of archive-like formats is rather simplistic at the moment. We make use of the archive only if the largest file in the archive clearly dominates other files and can be converted to plain text. We don't attempt to e.g. concatenate separate chapter files.

The most problematic file format in our experience is PDF. In PDFs, the content can be internally stored in various ways (including bitmap images of book pages) and e.g. Czech accented letters are prone to lose the accent or get mis-encoded. Different implementations of PDF-to-text conversions including Acrobat Reader can run in different problems on a given file. Moreover, hyphenated words and page headers are frequent and we have also found obscure cases of HTML print-outs in PDF where the printed header changed throughout the document as the timestamp in the header was changing. We attempted to solve most of the issues manually (by converting individual PDF files to txt prior to our generic convertor) but not everything has been handled due to time constraints.

4.2. Removing Page Breaks

Some of the texts include page numbers and other repetitive sequences such as page headers throughout the document. In worst cases, such header or footer appears even in the middle of a sentence. While the magnitude of the problem is not too severe (a book has a few hundred pages, so only a few hundred sentences per book can be malformed), we attempt to fix many of the cases.

We implemented a simple heuristic to identify candidates of page breaks and manually confirm them. A page break candidate, once constructed, is essentially a simple regular expression describing the prefix, the page number placeholder and the suffix that should be removed from anywhere in the document.

Our heuristic searches for all numbers in the document. Each occurrence of a number contributes to one or more candidates depending on the actual number observed and a very short character context of the number. In essence, we require the number to be not far away from the number of the last observed occurrence attributed to the candidate. For each candidate, we store all the numbers attributed to it, the prefix and suffix seen in the first occurrence and the length of a subsequence of the prefix and suffix seen in most other occurrences.

After the whole text has been processed, we sort the candidates based on the span of numbers covered by the candidate decreased by the number of gaps in the sequence and the number of duplicated entries. The most promising candidates represent the longest sequences of numbered items with the fewest errors in numbering. In most cases, these are indeed page numbers but sometimes we find footnotes or the table of contents instead. Due to the variance in book styles, we cannot assume some average number of pages so we prefer manual inspection of the list of candidates. This also allows to make sure that the suggested prefix and suffix are correct. After the manual confirmation, all occurrences of the confirmed candidate are removed from the document.

4.3. Unwrapping

Depending on the original file format and individual typesetting rules, some of the documents are hard wrapped, some indicate paragraphs by a blank line and some indicate them by indentation.

For the purposes of sentence segmentation (Section 5.1), we need somewhat normalized format to match the training data of our segmenter.

If there are more than 30% of lines longer than 90 characters, we assume the document is not hard-wrapped. For hard-wrapped documents, we check the number of blank lines in the document, and if there are more than e.g. 500, we assume they represent the paragraph boundary. With not enough blank lines, we assume the paragraphs are indicated by indentation and we insert a blank line before every indented line. Some documents do not even use indentation, so we additionally assume there is paragraph break whenever the line is shorter than 65 characters. When unwrapping individual paragraphs, we also join hyphenated words.

Some HTML documents we got are hard-wrapped using ⟨BR⟩ tags and we generally treat the ⟨BR⟩ tag as a paragraph boundary in our simple HTML stripper, so a specific rule for this case was needed.

	1-1	2-1	1-2	1-0	0-1	3-1	Others
Overall	9,860,595	688,946	495,372	331,576	316,282	87,691	167,801
subtitles	3,721,423	189,985	145,787	158,592	76,410	14,014	27,401
eu	2,382,721	312,656	155,694	64,147	99,901	55,541	90,403
techdoc	1,350,803	21,713	18,003	18,628	3,856	1,883	2,868
paraweb	1,146,999	104,264	51,046	52,441	95,343	11,434	23,657
fiction	1,070,639	55,218	119,206	34,804	37,293	4,585	22,336
news	145,763	3,733	3,778	1,891	2,902	165	831
navajo	42,247	1,377	1,858	1,073	577	69	305

Table 1. Types of aligned text segments as detected by Hunalign in the individual sources. X-Y stands for segment pairs containing X sentences in the English segment and Y sentences the Czech segment.

5. Sentence Segmentation, Alignment and De-Duplication

5.1. Sentence Segmentation

We use the trainable tokenizer introduced in Klyueva and Bojar (2008) with a few new extensions to perform sentence segmentation. The tokenizer internally performs a deterministic "rough" tokenization and deterministically inserts markers of positions where a sentence break or token join (e.g. space-delimited thousands) may happen. A maximum entropy classifier then decides where breaks or joins indeed happen based on features of surrounding tokens. We use only the sentence break information (and occasional token joins) but use the original non-tokenized format otherwise. The reason is that we wish to use TectoMT internal tokenization (Section 6 below) which should be compatible with the whole processing pipeline.

The training set for the maximum entropy classifier was further extended to contain more examples of the document types we deal with, e.g. book texts with lots of direct speech. Usually, the training set is created manually by complementing a sample plain text with the intended tokenization and segmentation. The trainable tokenizer creates training instances for the classifier by comparing the original and tokenized text. In our case, we were able to extend the training set of texts for both English and Czech semi-automatically by finding segments aligned 1-to-2 and containing a full stop somewhere around the middle of the single segment. Most of these cases were indeed errors where either the single segment should have been split, or the two corresponding segments in the other language should have been joined (e.g. at an unrecognized abbreviation). Simply adding these sentences with the correct segmentation projected from the other language improved the accuracy on our dataset.

5.2. Sentence Alignment

We use Hunalign (Varga et al., 2005) to automatically align sentences. To reduce data sparseness, we perform a rough tokenization (at this stage, the texts are only segmented and preserve original tokenization) and lowercase and restrict each token to at most first four letters. Additionally, we use a probabilistic dictionary based on GIZA++ word alignment of the previous version of CzEng, with the identical reduction of word types.

Table 1 lists alignment types seen in various source data. On average, about 82% of segments are aligned one-to-one but e.g. for the European Law texts, the percentage falls to 75%.

5.3. De-duplication

It is a common practice in corpus preparation to remove duplicated portions of data. For some types of texts, the common simple "sort | uniq" de-duplication procedure may skew the distribution of phrases unnecessarily, making e. g. a very common phrase "Yes. = Ano." occur only once in the whole corpus.

For most sections of our corpus, we completely avoid de-duplication at the level of segments and prefer de-duplication at the level of documents (e.g. e-books). For some sources, e.g. the web collection, de-duplication is inevitable because web pages from a single site usually contain large amounts of repetitive text (that is actually seldom read by humans, unlike repetitive phrases in books).

To avoid the above-mentioned distortion, we remove duplicated aligned segments of web pages using a more sensitive context-based technique: we use a sliding window of 3 consecutive lines and print the lines in the window if no such window was printed before. For instance, for the lines "a b c a b c b d b" we get "a b c b d b". The second occurrence of "a b c" got removed but the overall distribution of "b" is influenced less.

5.4. Plaintext Checks

Sentence-aligned plaintext format is suitable for performing many simple checks to filter out either mis-aligned or simply bad segments. At this stage of corpus collection, we search and remove all suspicious sentence pairs, i.a.:

- the Czech and English sentences are identical strings (usually untranslated text from a website),
- the lengths of the sentences are too different (usually due to a wrong alignment or a wrong sentence segmentation),
- there is no Czech word on the Czech side or English word on the English side[15],

[15]We use the word lists from the British National Corpus the Czech National Corpus disregarding letter case. We prefer longer words for the test: if there are some words longer than three letters, at least one of

Bad 1-1 Segments	[%]	Most Frequent Errors
subtitles	4.6	Mismatching lengths (42.0%), Identical (27.3%), No English word (10.9%),
eu	33.3	Identical (39.9%), No English word (19.2%), Not enough letters (17.2%),
techdoc	10.2	Identical (37.9%), No English word (28.4%), Not enough letters (10.0%),
paraweb	59.5	Identical (61.7%), No English word (25.1%), Mismatching lengths (3.3%),
fiction	3.1	Mismatching lengths (54.9%), Suspicious char. (14.6%), Repeated character (6.1%),
news	3.8	Identical (54.1%), Suspicious char. (17.7%), No English word (9.3%),
navajo	11.9	Identical (40.9%), No English word (19.0%), Not enough letters (11.7%),

Table 2. Percentage of 1-1 sentence pairs rejected by various error-detection filters.

- there is a suspicious character (either non-printable one or an unlikely symbol) or a repeating sequence of a character.

Table 2 displays the percentage of 1-1 aligned sentences with one or more errors. The second column in the table lists the most frequent error in each of the sections.

Many of the errors can be corrected in earlier stages of corpus cleaning and we will continue to refine the cleaning process but for the time being, we prefer to remove all suspicious segments.

The overall most frequent error is "Identical", and we see that e.g. more than 36% of web data (61.7% out of 59.5% of erroneous segments) are removed due to this error. Unfortunately, many of the seemingly parallel web pages contain non-translated sections. The cleanest source is probably the ebooks section with some errors in segmentation or alignment (Mismatching lengths).

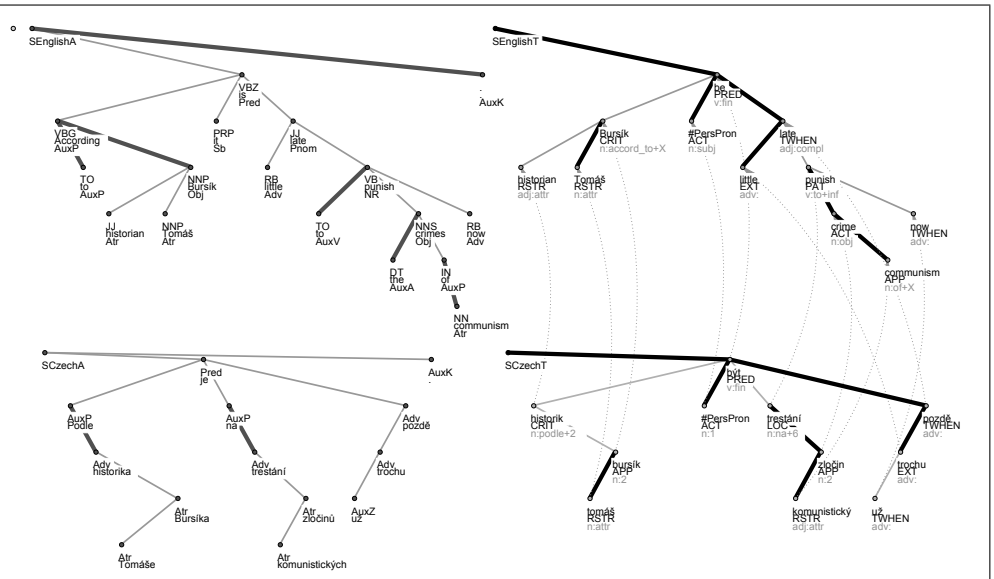
6. Sentence Annotations

The pairs of Czech and English 1-1 aligned sentences are enriched with rich morphological and syntactic annotations. The annotation scheme is adopted (with certain modifications) from the Prague Dependency Treebank 2.0 (Hajič et al., 2006). Each sentences is provided with three layers of annotation:

- morphological layer: each token (word or punctuation mark) is labeled with its lemma and morphological tag,
- analytical layer: each sentence is represented as a surface-syntax dependency tree called analytical tree (a-tree), with nodes corresponding to tokens and edges corresponding to surface-syntax dependency relations,
- tectogrammatical layer: each sentence is represented as a deep-syntactic dependency tree called tectogrammatical tree (t-tree), in which nodes have complex structure and correspond only to autosemantic words.

In addition to the PDT 2.0 scheme, a new layer containing annotation of named entities is added.

them has to be confirmed in the word list. If all words contain at most three letters, we accept also shorter words for the word list check.



English: *According to historian Tomáš Bursík it is little late to punish the crimes of communism now.*
Czech: *Podle historika Tomáše Bursíka je na trestání komunistických zločinů už trochu pozdě.*

Figure 2. Simplified visualization of parallel analytical and tectogrammatical tree representations of a sample CzEng 0.9 sentence pair.

The fully automatic annotation procedure is implemented within the TectoMT framework (Žabokrtský et al., 2008). The procedure is highly similar for both languages:[16]

1. each sentence is tokenized using a simple regular expression pattern,
2. the sentence is tagged by the Morce tagger (Spoustová et al., 2007),
3. the tokens are lemmatized; this is done already in the tagging step in the case of Czech sentences, while for English a new lemmatizer was implemented in TectoMT (Popel, 2009),
4. named entities are recognized and classified; a recognizer based on Support Vector Machines described in Kravalová and Žabokrtský (2009) is used for Czech sentences, while for English sentences we use Stanford Named Entity Recognizer introduced in Finkel et al. (2005),

[16]The procedure description is highly simplified here, in fact the procedure composes of roughly sixty subsequent blocks (basic processing units in TectoMT).

5. analytical dependency tree is created by the maximum spanning-tree parser introduced in McDonald et al. (2005) (using feature pruning described in Novák and Žabokrtský, 2007),
6. a-tree nodes are labeled with analytical functions; the values are provided already by the parser on the Czech side, while on the English side the values have to be assigned subsequently (a rule-based analytical function assigner developed in Popel, 2009 is used),
7. a t-tree is created from the a-tree by merging autosemantic a-nodes with their associated auxiliary a-nodes (e.g. a noun with a preposition and a determiner node),
8. the t-tree is labeled with grammatemes,
9. grammatical coreference links are identified in the t-tree,
10. the t-tree nodes are labeled with functors by a tool developed in Klimeš (2006),
11. finally, the resulting t-trees are aligned using the tectogrammatical aligner developed in Mareček (2008).

A sample pair of resulting sentence representations (and their alignment) is shown in Figure 2.

6.1. Line-Oriented Operations

TectoMT uses a complex XML-based file format, an instance of Prague Markup Language (Pajas and Štěpánek, 2006). While the format is excellent for the rich annotation and the interoperation of TectoMT processing blocks, it brings an additional overhead for tasks performed on large sets of sentences. Quick and simple selection of sentences matching regular expressions, counting sentences or line shuffling cannot be performed with standard utilities like grep, wc or shuf, because sentences represented in XML span over multiple lines.

To facilitate the use of line-based tools on TectoMT data, we introduce a simple modification to the file format. The new file format is called "lot" (line-oriented-tmt) and stores each sentence using XML on a single line. In other words, line breaks and indentation whitespace within the XML representation of sentences are removed. To match the line-oriented approach even closer, we omit any XML header and footer sections in "lot", so every line of a "lot" file holds a sentence. Fortunately, it is not common to store any valuable information in a header section once the text has been segmented.

Both conversion to and back from "lot" is fast and can operate on an infinite stream of sentences. When converting to "lot", we use a SAX parser to read sentence after sentence, strip any line breaks and emit the sentence. To convert back from "lot", one has to simply add a proper XML header and footer and optionally reindent the file, e.g. using "xmllint -format".

7. Corpus Structure and Size

This section provides technical details on the final shape of CzEng 0.9 data.

7.1. Dividing Data into Files, Shuffling

The Czech author law[17] permits to use short citations of published works for non-commercial educational or research purposes. To avoid the possibility of reconstructing the original texts included in CzEng, we break all documents into short fragments, shuffle them and discard any explicit information that would allow to reconstruct the original ordering of the fragments.

Let us recall that CzEng contains only sentences automatically aligned 1-1. In reality, most documents are not translated sentence by sentence, and even if this were the case, the exact sentence alignment is seldom found by the automatic procedure. So the original documents are unreconstructable from what is contained in CzEng 0.9 not only because of fragmentization and shuffling, but also because of the data losses imposed by the 1-1 requirement (and also because of other losses during pair filtering).

In order to preserve the utility of CzEng for advanced NLP techniques that extend beyond sentence boundary, such as anaphora resolution, we preserve at least short sequences of sentences, if possible. Given our processing pipeline, some breaks of the continuous flow of sentences naturally happen at sentences not aligned 1-1 or filtered by one of our plaintext checks in either of the languages. We use all these breaks and add further breaks after at most 13 consecutive sentence pairs. Due to the natural breaks, there are only 4.0 sentences per block on average. We shuffle the obtained set of blocks and assign a unique identifier to each of the blocks. Finally, the blocks are concatenated to files of about 50 to 60 sentence pairs depending on the exact sizes of the blocks in the file. We use the above mentioned line-oriented approach for these operations.

For domain-specific training or domain adaptation, the block identifiers preserve the coarse data source type (subtitles, eu, paraweb, techdoc, fiction, news, navajo) but no other meta-information is available.

7.2. Dividing Data into Sections

In order to reduce the load on the file system and to simplify selection of smaller random samples of the data for e.g. debugging, we organize the final TMT files into 100 subdirectories, each containing approximately 1500 files.

We expect many researchers to use the full size of CzEng for training their systems but some may wish to reserve a portion of the data for evaluation purposes. In order to synchronize the selection of the test set, we label about 10% of the data `dtest`

[17] The law 121/2000 Sb. including amendments up to 168/2008 Sb., see §31.

(development test set) and another 10% of the data `etest` (evaluation test set). The development test set should be used for tuning of parameters and the evaluation test should be used for final evaluation only.

The directories are thus called `train00`, ..., `train79`, `dtest80`, ..., `dtest89`, and `etest90`, ..., `etest99`.

In any case, researchers should clearly indicate which sections they used for the training and for the evaluation.

7.3. File Formats

7.3.1. CzEng in TMT Format

The main file format of CzEng 0.9 is the TectoMT file format called TMT, an instance of Prague Markup Language (Pajas and Štěpánek, 2006) based on XML. Unlike the PDT 2.0 file format, TMT allows to keep all layers of language representation in a single file. In CzEng 0.9, each TMT file is a sequence of around 50 bundles, each of them comprising morphological, analytical and tectogrammatical representations of an English sentence and of its Czech counterpart sentence, as well as their original surface string forms and their tectogrammatical alignment.

7.3.2. CzEng in Plain Text

For some applications, the rich annotation stored in TMT files is not needed or causes an unwanted bias due to our tokenization rules. Therefore, we also provide CzEng in plain text format, one sentence pair per line. The English and Czech versions of the sentence are delimited by a single tab.

We preserve the same corpus division into training and test sections. Instead of a directory `train`··, the section is stored in a single file `train`··`.gz`.

7.3.3. CzEng Export Format

The TMT format described above, is the only authoritative format of CzEng 0.9 rich annotation. However, to allow access to the rich annotation for researchers who do not wish to use the TectoMT framework with its API for TMT files, we provide CzEng 0.9 in a simple export format as well. Note that not all information from the original TMT files is preserved[18].

The export format represents each sentence pair on a single line consisting of the following tab-delimited columns: Sentence ID (including coarse CzEng source type), English a-layer, English t-layer, English lex.rf (i.e. links from English t-nodes to the

[18] We do not export all attributes of the nodes. We also remove any spaces delimiting thousands in numbers whereas the original TectoMT annotation pipeline represents space-delimited numbers in a single node with spaces in the attribute form and the a- and t-lemmas.

corresponding a-node bearing the lexical value), English auf.rf (i.e. links from English t-nodes to their auxiliary a-nodes), Czech a-layer, Czech t-layer, Czech lex.rf, Czech aux.rf, English-Czech t-layer alignment.

All the columns representing the dependency tree at a layer use so-called "factored" notation: each space-delimited word on the line represents one node of the tree. Individual attribute values of the node are delimiter by vertical bar "|". The order of the attributes is fixed for a given language and layer and usually can be guessed from attribute values.

The dependency structure of the tree is represented using two attributes: the "ord" stores the global linear order of the node in the tree starting from 1 and the "gov" contains the ord value of the governor of the node. The root of the tree has the gov value set to zero. The nodes of the tree are always listed in ascending order of ord and there are no gaps in the numbering.

In CzEng 0.9, we export these attributes:

- Czech and English a-layers: word form, lemma, morphological tag, ord, gov, analytical function.
- English t-layer: t-lemma, functor, ord (deepord), gov, nodetype, formeme, the grammatemes: sempos, number, negation, tense, verbmod, deontmod, indeftype, aspect, numertype, degcmp, dispmod, gender, iterativeness, person, politeness, resultative, and the attributes: is_passive, is_member, is_clause_head, is_relclause_head, val_frame.rf.
- Czech t-layer: t-lemma, functor, ord (deepord), gov, nodetype, formeme, the grammatemes: sempos, number, negation, tense, verbmod, deontmod, indeftype, aspect, numertype, degcmp, dispmod, gender, iterativeness, person, politeness, resultative, and the attributes: is_passive, is_member, is_clause_head, is_relclause_head, val_frame.rf.

All the columns representing some kind of links between two layers or languages are simple space-delimited pairs of indices. Unlike the ord and gov attributes, here the nodes are indexed starting from zero. In other words, e.g. the pair "0-1" of the Czech lex.rf indicates that the first Czech t-node (index 0) obtained its lexical value from the second (index 1) a-node, a typical situation of a noun with a preposition at the beginning of the sentence.

Some types of alignment allow 1-to-many links or possibly even many-to-many links. In these cases, some nodes are simply mentioned in the listing more than once.

Again, the same corpus division into training and test sections is preserved. Instead of a directory `train··`, the section is stored in a single file `train··.gz`.

7.4. CzEng 0.9 Size

Table 3 lists total number of sentences and Czech and English nodes at both layers of the annotation per section. The number of a-nodes can be interpreted as the number of "words" including punctuation.

		English		Czech	
Source	Sentences	a-nodes	t-nodes	a-nodes	t-nodes
eu	1,589,036	31,725,089	19,458,544	28,484,512	19,310,396
subtitles	3,549,367	26,550,305	16,615,991	22,175,284	16,675,187
fiction	1,036,952	17,045,233	10,861,341	15,031,926	11,102,760
techdoc	1,212,494	9,099,748	6,339,129	8,460,491	6,512,247
paraweb	464,522	4,946,552	3,666,149	4,750,757	3,667,297
news	140,191	3,196,303	2,019,758	2,945,777	2,220,789
navajo	37,239	612,826	385,292	539,659	405,484
Total	8,029,801	93,176,056	59,346,204	82,388,406	59,894,160

Table 3. Number of sentence pairs in CzEng 0.9 and number of nodes in their analytical and tectogrammatical tree representations. Artificial tree roots are not counted here, therefore the numbers of a-nodes given in the third and fifth column are equal to the number of tokens (words and punctuation marks) contained in the corpus.

7.5. Obtaining CzEng 0.9

CzEng 0.9 is available for non-commercial research purposes at:
`http://ufal.mff.cuni.cz/czeng/`

8. Conclusion

We have presented CzEng 0.9, a new release of our Czech-English parallel corpus, extended both in the data size and the depth of automatic annotation. Compared to previous versions, the corpus should be cleaner thanks to several automatic error detection techniques we implemented. Inevitably, many errors remain in the released corpus and we plan to further refine our filtering techniques and base them on the deep syntactic analyses and their alignment as well in future versions.

We believe that CzEng 0.9 is a unique resource for MT developers (definitely for the given pair of languages), and hope that that its availability will further boost the research in the field.

9. Acknowledgement

The work on this project was supported by the grants MSM0021620838, MŠMT ČR LC536, 1ET101120503 and FP7-ICT-2007-3-231720 (EuroMatrix Plus).

Bibliography

Beňa, Peter. Filmové titulky jako zdroj paralelních textů (movie subtitles as a source of parallel texts). Bachelor's Thesis, Faculty of Mathematics and Physics, Charles University in Prague, 2009.

Bojar, Ondřej, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of LREC'08*, Marrakech, 2008a.

Bojar, Ondřej, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of LREC'08*, Marrakech, 2008b.

Bojar, Ondřej, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March 2009. Association for Computational Linguistics.

Cuřín, Jan, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. Prague Czech-English Dependency Treebank, Version 1.0. Linguistics Data Consortium, Catalog No.: LDC2004T25, 2004.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics ACL 2005*, pages 363–370, 2005.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, and Marie Mikulová. Prague Dependency Treebank 2.0. LDC, Philadelphia, 2006.

Klempová, Hana, Michal Novák, Peter Fabian, Jan Ehrenberger, and Ondřej Bojar. Získávání paralelních textů z webu. In *ITAT 2009 Information Technologies – Applications and Theory*, Sept. 2009.

Klimeš, Václav. *Analytical and Tectogrammatical Analysis of a Natural Language*. PhD thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, 2006.

Klyueva, Natalia and Ondřej Bojar. UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proc. of International Conference Corpus Linguistics*, pages 188–195, Oct. 2008.

Kravalová, Jana and Zdeněk Žabokrtský. Czech named entity corpus and SVM-based recognizer. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 194–201, Suntec, Singapore, 2009. Association for Computational Linguistics. ISBN 978-1-932432-57-2.

Marcus, M. P., B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Mareček, David. Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus. Master's thesis, Charles University, MFF UK, 2008.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HTL/EMNLP*, pages 523–530, Vancouver, BC, Canada, 2005.

Novák, Václav and Zdeněk Žabokrtský. Feature engineering in maximum spanning tree dependency parser. In Matoušek, Václav and Pavel Mautner, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue*, number XVII in Lecture Notes in Computer Science, pages 92–98, Berlin / Heidelberg, 2007. Springer Science+Business Media Deutschland GmbH. ISBN 978-3-540-74627-0.

Pajas, Petr and Jan Štěpánek. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In Hinrichs, Richard Erhard, Nancy Ide, Martha Palmer, and James Pustejovsky, editors, *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pages 40–47, Genova, Italy, 2006. ISBN 2-9517408-2-4.

Popel, Martin. Ways to Improve the Quality of English-Czech Machine Translation. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, 2009.

Ralf, Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2142–2147. ELRA, 2006.

Rambousek, Jiří, Jana Chamonikolasová, Daniel Mikšík, Dana Šlancarová, and Martin Kalivoda. KAČENKA (Korpus anglicko-český - elektronický nástroj Katedry anglistiky), 1997.

Sgall, Petr. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic, 1967.

Sgall, Petr, 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.

Spoustová, Drahomíra, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Prague, 2007.

Štěpánek, Jan. Post-annotation Checking of Prague Dependency Treebank 2.0 Data. In *Proceedings of the 9th International Conference, TSD 2006*, number 4188 in Lecture Notes in Computer Science, pages 277–284, Berlin / Heidelberg, 2006. Springer-Verlag Berlin Heidelberg. ISBN 3-540-39090-1.

Tiedemann, Jörg and Lars Nygaard. The OPUS corpus - parallel & free. In *Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004*, Lisbon, May 26–28 2004. URL `http://stp.ling.uu.se/~joerg/paper/opus_lrec04.pdf`.

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing RANLP 2005*, pages 590–596, Borovets, Bulgaria, 2005.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, 2008. Association for Computational Linguistics.

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 85–104

Tectogrammatical Annotation of the Wall Street Journal

Silvie Cinková, Josef Toman, Jan Hajič, Kristýna Čermáková,
Václav Klimeš, Lucie Mladová, Jana Šindlerová, Kristýna Tomšů,
Zdeněk Žabokrtský

Abstract

This paper gives an overview of the current state of the Prague English Dependency Treebank project. It is an updated version of a draft text that was released along with a CD presenting the first 25% of the PDT-like version of the Penn Treebank – WSJ section (PEDT 1.0).

Before the January 2009 release, the conversion from the original phrase structure trees into dependency trees as well as the consistency checks were substantially enhanced to save manual work. The conversion is partly performed by scripted rules and partly by a statistical parser. To make the rules more powerful, the phrase-based Penn Treebank – WSJ was enriched with other publicly available language resources – the manual annotation of flat noun phrases and the named-entity and coreference tagging.

At the moment, 50% of the 1 million corpus have been manually annotated and consistency-checked on the tectogrammatical layer.

1. Introduction

We are presenting the first results of a manual tectogrammatical annotation of the Wall Street Journal - Penn Treebank III. We call the WSJ-PTB texts and the annotation of them the **Prague English Dependency Treebank (PEDT)**. About 50% of the WSJ-PTB have been manually annotated at the moment[1].

The Wall Street Journal section of the Penn Treebank is one of the first large manually annotated treebanks. It has become established as a standard reference corpus for statistical machine learning experiments. The PTB bracketing style has been adopted

[1]It was 25% in the draft version of this paper, which we attached to the CD with the PEDT 1.0 released in January 2009. The contents of the CD can also be accessed at http://ufal.mff.cuni.cz/pedt

Please cite this article as: Silvie Cinková, Josef Toman, Jan Hajič, Kristýna Čermáková, Václav Klimeš, Lucie Mladová, Jana Šindlerová, Kristýna Tomšů, Zdeněk Žabokrtský, Tectogrammatical Annotation of the Wall Street Journal. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 85–104.

by corpora of other languages, which strengthened the prominence of the original WSJ-PTB corpus. Although WSJ in practice is a restricted-domain corpus, which may affect its usability for general NLP tasks [2] (cf. e.g. Oepen, 2007 and Gildea, 2001), we believe that building an additional syntactico-semantic annotation on WSJ is sensible. After having built and refined the Prague Dependency Treebank, a one-million corpus of Czech 1990s newspaper texts with manual syntactico-semantic annotation (Hajič et al., 2006), we have adapted the PDT annotation scheme to English. We decided to draw on a corpus manually annotated in a widely known format, since the option of comparing both annotation schemes can be particularly useful for some users. In addition, familiar text examples facilitate the understanding of the new annotation scheme by users, and, in turn, we benefit from the constant confrontation with the PTB bracketing style while creating the annotation guidelines (Cinková et al., 2006). Most importantly, the original manual annotation has provided an excellent input for the conversion.

While creating the annotation guidelines, we made a tentative annotation of English spontaneous (but slightly edited) spoken dialogs (Hajič et al., 2008; Bradley et al., 2008) in order to compensate for the style bias of WSJ-PTB and to make sure that the current annotation scheme would fit a broader range of styles than business press can offer.

2. Background

2.1. Functional Generative Description and Tectogrammatical Representation

The **Functional Generative Description** (FGD) is a stratified formal language description based on the structuralist tradition, developed since the 1960s (Sgall et al., 1986). Unique contribution of FGD is the so-called **tectogrammatical representation (TR)**. It is implemented in a family of syntactico-semantically annotated treebanks. The treebanks are typically annotated at three layers:

- morphological layer (m-layer)
- analytical layer (a-layer)
- tectogrammatical layer (t-layer).

At the m-layer the text is still a sequence of strings with added tokenization, POS tagging, and lemmatization. Each token has its unique ID. The a-layer displays the sentences as dependency trees in which each token is represented by a node. The nodes are labeled with coarse syntactic labels. The topmost layer so far is the tectogrammatical layer (t-layer), which is based on the tectogrammatical representation (TR) proposed by FGD. Conceived as an underlying syntactic representation, the TR captures the **linguistic meaning of the sentence**. By *linguistic meaning* we understand

[2] From the linguistic point of view the corpus domain restriction is not necessarily a drawback, given the linguistic research is consciously focused on local discourse patterns and local meanings (cf. e.g. Römer, 2008).

"what has been said and can be perceived without any special knowledge of the situation" but with the common understanding of basic conversational implicatures, as well as with tolerance for redundance and vagueness. E.g. unlike a strictly logical representation, the tectogrammatical representation would not deal with the question whether in the sentence *John heard a cry* there must have been a cry for John to hear, or whether John might have mistakenly interpreted a sound he had heard as a cry. On the other hand, the tectogrammatical representation would indicate that something unexpressed on the surface is likely to be understood from the context or from the situation, or that something has been deliberately left underspecified; e.g., in the sentence *I told you last night* the tectogrammatical representation of the verb *to tell* would indicate that *something* (EFF), possibly about a mentioned matter (PAT) was told to somebody, and it would indicate whether these entities could be retrieved from the verbal context or not. (While the missing argument of *tell* is in this case likely to be retrievable from the context, some ellipses systematically express generalizations; e.g., *Peter can eat* [something, anything] *alone*.)

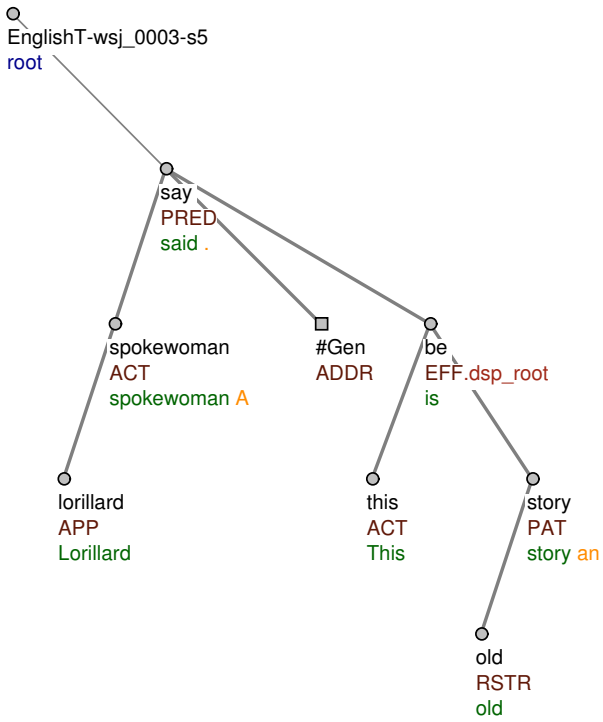
2.2. Tectogrammatical Annotation

Tectogrammatical annotation is to be held apart from the theoretical construct of tectogrammatical representation, as many annotation resolutions have been introduced for technical and consistency reasons rather than being conditioned by the theory. The dependency treebanks of the PDT family are however being continuously refined, with the ambition of adequately reflecting the FGD as a linguistic description. That is done by a step-by-step uncovering and consistent tectogrammatical representation of lexical and structural patterns.

The basic description unit of the tectogrammatical annotation is the **sentence**. Each sentence is represented as a projective dependency tree with nodes and edges (henceforth **tectogrammatical tree structure** or **TGTS**). Only **content words** are represented by nodes. Each node has a semantic label ("functor"), which renders the underlying (deep) syntactic relation of the given node to its parent node. Function words are mostly represented as attribute values in the internal structure of the respective nodes. The attribute values contain references to the analytical (surface-syntax) annotation layer instead of the forms of the function words themselves.[3] Tectogrammatical annotation, which draws on TR, captures the following aspects of text:

- syntactic dependencies
- argument structure (data interlinked with a lexicon)
- information structure (topic-focus articulation)
- grammatical and partly also textual coreference
- deletion restoration

[3] A more detailed specification of the annotation conventions is given by (Cinková et al., 2006).



A Lorillard spokewoman said, `` This is an old story. Tisková mluvčí Lorillardu řekla, "Toto je stará věc.

Figure 1.

- information on lexical derivation[4]
- semantically determined grammatical categories **(grammatemes)**[5]

Figure 1 presents the tectogrammatical tree structure (TGTS) of the sentence *A Lorillard spokewoman said: "This is an old story."*.

Each sentence is identified with a unique identifier in the **technical root** of the tree (the topmost node). This node does not reflect any part of the sentence. The topmost linguistically relevant tectogrammatical node **(t-node)** is the predicate *said*, whose tectogrammatical lemma is *say*. The internal structure of this node contains references to the analytical (dependency surface-syntax) tree of the same sentence, in

[4]so far Czech only

[5]just a tentative automatic insertion in English at the moment, not in this text

which each token is represented by a node. The references point to all analytical nodes **(a-nodes)** that affect the meaning unit rendered by the given t-node. We distinguish two types of references pointing to the analytical layer:

- reference to a content word
- reference to an auxiliary word.

The strings in darker gray in Figure 1 represent the targets of the content-word references. The lighter strings represent the targets of the auxiliary-word references.

Figure 1 also displays a few common semantic labels (functors) used in TGTS. The functors indicate the underlying syntactic relation of a given node to its parent node. A node that modifies another node is governed by that node. About 70 functors in total are used in the annotation. It is partly functors for kinds of dependences, partly functors for semantic relations between conjuncts in coordinations, and a few functors which help organize cognitively specific syntactic structures such as comparisons. Most dependent nodes can be divided into two groups: **inner participants** vs. **free modifications**. They differ in whether a valency complementation with the given functor can occur more than once as dependent on the same parent node (except for a coordination). The inner participants cannot repeat, while the free modifications can. This distinction has nothing to do with whether they are obligatory or optional. Despite their name, even free modifications can be obligatory in the valency frames of certain words (verbs, nouns, or adjectives), while inner participants also can be optional. Cf. the following example sentences:

(1) *Peter.*ACT *eats vegetables*.PAT

(2) *Peter*.ACT *eats vegetables*.PAT *and pasta*.PAT

(3) **Peter.*ACT *eats vegetables*.PAT *pasta*.PAT

versus

(4) *Peter went to Prague.*DIR3

(5) *Peter went to Prague*.DIR3 *to John's office*.DIR3

The obligatoriness vs. optionality of a valency complementation can be determined by an introspective **dialogue test** (Panevová, 1974 and Panevová, 1975). There are five inner participants: ACT (Actor), PAT (Patient), ADDR (Addressee), ORIG (Origin), and EFF (Effect). There is a sixth inner participant exclusively used with nouns: APP ("appurtenance"; i.e. association in a broader sense than ownership). Few very common free modifications can be obligatory: e.g. DIR3 (direction towards a destination), DIR1 (direction from a source location), DIR2 (direction across or through an area), TWHEN (timepoint), and MANN (manner). A complete list of functors can be found in (Cinková et al., 2006).

In Figure 1, *Lorillard* modifies *spokewoman*, and the syntactic relation between *Lorillard* and *spokewoman* is labelled as APP. The effective root (i.e. the topmost node

under the technical root, disregarding coordination nodes) of a direct speech subtree is marked with the note `dsp_root`. The predicate *say* has three obligatory participants according to the valency lexicon: Actor, Addressee, and Effect (what is being said). The Addressee is underspecified, which is why a generated node with the t-lemma substitute `#Gen (generalized)` was inserted. In general, each occurrence of a word with an argument structure (so far only verbs and verbal nouns in the English annotation) is interlinked with an instance (a **valency frame**) in the **valency lexicon**. When assigned to a lexicon frame, the occurrence of the given word must have a complete pattern of obligatory arguments **(inner participants)** determined by the valency lexicon. Generated nodes with t-lemma substitutes are inserted to complete the valency frame. A complete list of t-lemma substitutes can be found in (Cinková et al., 2006).

3. The Original Penn Treebank

The Wall Street Journal section of the Penn Treebank (Marcus et al., 1999) comprises approx. 1.25 million POS-tagged words in 49 208 sentences, which are manually annotated with constituency **bracketing** and **labels**. PTB-WSJ III keeps the PTB II (Marcus et al., 1995) bracketing style (Bies et al., 1995). Each bracket is labeled with one of the standard syntactic labels (NP, ADVP, PP, S, etc.). Since PTB II, the brackets are enriched with more detailed labeling. On the clausal level, the labels distinguish 5 types of clauses (subordinate clause, inverted question, inverted declarative sentence, direct wh-question and simple declarative clause). The phrase labels separate structural anomalies (lists, fragments, parentheses, reduced relative clauses, unlike coordinated phrases), heads of certain parts of speech (adjective, adverb, etc.) , recurrent semantic units (e.g. quantifier phrases used within noun phrases) and transition phenomena (e.g, multi-word conjunctions like *as well as, not to mention*, etc., which have coordinative as well as subordinative features). On top of phrase and clause labels, non-terminal nodes can get **function tags**. The function tags mark specific linguistic phenomena, such as the nominal function of a gerundial clause *(Baking pies is fun., I do not mind about your leaving early.)*, "dative" alternation in certain verbs (*to give*), predicate complements *(I consider Kris a fool.)*, topicalization of a phrase by the left shift in the word order *(Of the 500 barbers in Philadelphia only 10 know what they are doing.)*, and several semantic labels of adjuncts (temporal, spatial, extent, etc.). The bracketing manual gives detailed information on linguistic phenomena which were captured systematically, along with several financial-speak-specific annotation templates.

4. Complementary Annotations

Several important annotations have been built above the PTB-WSJ texts since the release of the treebank. Two lexical sources were created and interlinked with the data:

- PropBank (Palmer et al., 2004), the valency lexicon of verbs

- NomBank (Meyers et al., 2008), the valency lexicon of nouns, which in fact also comprises lexicons of predicate nouns (the nominal components of light verb constructions), adjectives and adverbs.

Both lexicons are referenced by data annotations of argument structure.

- Annotation of flat noun phrases (Vadas and Curran, 2007; Vadas, 2007)
- BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005)

4.1. Flat Noun Phrases Annotation

Complex noun phrases like *an Air Force Contract* are left flat by the original Penn Treebank annotation. Vadas (Vadas, 2007; Vadas and Curran, 2007) has created a manual annotation of the almost 61,000 complex noun phrases in WSJ-PTB, making use of the entity annotation known from (Weischedel and Brunstein, 2005). By adopting the basic priciples of the annotation of biomedical texts (Kulick et al., 2004), Vadas et al. have inserted labelled brackets around left-branching structures. The newly created constituents with noun heads have been assigned the label NML, whereas those with adjectival heads are marked as JJP.

Hence, the phrase *Air Force contract*, in the original PTB bracketing represented as

```
(NP (NNP Air) (NNP Force) (NN contract))
```

is supplemented with an NML constituent that indicates that *Air Force* is a sub-NP structure within the entire phrase:

```
(NP
(NML (NNP Air) (NNP Force))
(NN contract))
```

4.2. BBN Corpus

Weischedel and Brunstein (Weischedel and Brunstein, 2005) created a stand-off annotation of pronoun coreference along with an annotation of a variety of entity and numeric types above WSJ-PTB. The entity annotation has been designed for question-answering tasks. It distinguishes 29 categories with subtypes. The most relevant for our annotation (see Section 6) are the following categories:

- Person Name
- Person Descriptor
- Facility Name
- Facility Descriptor
- Organization Name
- Organization Descriptor
- GPE: country, city, state/province
- Work of Art.

5. Conversion

Since we launched the routine tectogrammatical annotation of PEDT, we have worked with automatically pre-generated tectogrammatical trees, which were obtained by a conversion of the original constituency trees into the FGD-based analytical trees and subsequently from the analytical trees into tectogrammatical trees. The conversion tools were recently refined and integrated into a complex English-to-Czech machine-translation system called **TectoMT** (Žabokrtský et al., 2008). The system consists of a long sequence of processing modules **(blocks)**, which perform small partial tasks. First, English tectogrammatical trees are generated from the English text input. Then the English tectogrammatical trees are transferred to Czech tectogrammatical trees. Czech analytical trees are created from the Czech tectogrammatical trees. Finally, the Czech text is created from the analytical trees.

For the automatic pre-generation of English tectogrammatical trees we have used the manually created constituency trees of WSJ-PTB converted into a PML format as input for the first sequence of blocks, by which we have obtained automatically generated analytical trees.[6] These blocks:

- lemmatize the word forms
- mark the head node (using a set of heuristic rules)
- build temporary m-trees containing morphological information (to be merged with a-trees later)
- convert constituency trees into a-trees
- apply some heuristic rules to fix apposition constructions
- apply other heuristic rules for reattaching incorrectly positioned nodes
- unify the way in which multiword prepositions (such as *because of*) and subordinating conjunctions (such as *provided that*) are treated.
- assign analytical functions (labels) if necessary for a correct treatment of paratactic constructions.

The next (much bigger) chain of blocks builds tectogrammatical trees upon the analytical trees. The procedure is the following:

- Mark a-nodes which represent auxiliary words.
- Build t-trees. Each a-node cluster formed by an autosemantic node and possibly several associated auxiliary nodes is 'collapsed' into a single t-node. T-tree dependency edges are derived from a-tree edges connecting the a-node clusters.
- Distinguish coordination members from shared modifiers (modifiers that modify all coordination members at the same time, e.g. *the kind* [*girls and boys*]).
- Modify t-lemmas when necessary, insert t-lemma substitutes for selected nodes.
- Assign functors necessary for proper treatment of coordination and apposition constructions and fix the coordination-member attributes.
- Distribute shared auxiliary words in coordination constructions.

[6] Some of the blocks used in the MT tasks have been left out when building tectogrammatical trees for manual annotation.

- Mark t-nodes which are roots of t-subtrees corresponding to finite verb clauses.
- Mark passive verb clauses.
- Assign functors in selected cases (rule based).
- Assign functors by a statistically based procedure consisting of several blocks.
- Mark t-nodes corresponding to infinitive verbs.
- Mark t-nodes which are roots of t-subtrees corresponding to relative clauses or direct speech.
- Mark t-nodes which are roots of parenthetic t-subtrees.
- Fill in or correct several internal attributes of the nodes (e.g. `nodetype`).
- Insert a reference Czech (manual) translation of the sentence.
- Assign valency frames.
- Recompute deep ordering of the nodes.
- Strip some attributes which are no longer useful when the procedure is finished.

Apart from the original TectoMT blocks, a statistical functor assigner (a recent component of a tectogrammatical parser - Klimeš, 2007) has been employed to increase the accuracy of the automatic functor pre-assignment (it is already mentioned in the above list of blocks). A preliminary measurement (the trees pre-generated with and without the assigner compared respectively with the same trees which had been manually annotated before) has proved a significant improvement on the WSJ-PTB data. The trees generated without the assigner have achieved a 57.6 % functor agreement with the reference manual annotation. The introduction of the assigner has raised the agreement to 77.3 %. That is quite good because the best interannotator agreement ever achieved was 85.7 %.

6. Rule-based pre-annotation

A significant improvement of the pregenerated tectogrammatical trees has been brought by the flat NP annotation (Vadas, 2007), which we have integrated into the WSJ-PTB data fed to TectoMT. To increase the consistency and to speed up the annotation even more, we have decided to improve the trees obtained from TectoMT by hand-written rules. These rules have been designed to apply to selected recurrent structures, which were often impossible to detect by morpho-syntactic criteria, being conditioned rather lexically or even stylistically. When creating the rules for automatical pre-annotation, the constituency trees of WSJ-PTB were first browsed with Netgraph (Mírovský, 2008) and informally described along with the tectogrammatical subtrees desired as output. These informal descriptions have been rewritten into perl scripts.

All our hand-written rules for automatic pre-annotation of WSJ-PTB are designed as "Find a specified constituency structure, locate the corresponding tectogrammatical structure and correct it". To create these rules, we have used the following features:

- WSJ-PTB terminal, nonterminal and function tags
- WSJ-PTB structure

- lemmatization
- text strings (lists of words)
- BBN entity tags

We are including a few examples of the rules here.

Phrases of the type "$600 a share"

We are looking for an NP phrase (node A) with the function tag ADV and an NP or QP phrase (node B) to the left. Node A has exactly two childnodes (both terminal), the left one having the wordform "a" and the tag "DT". In case of a match we identify the t-subtrees created from the constituency structures rooted at the nodes A and B (let us call them TSA and TSB). Then we hang TSA under TSB and assign the functor REG to the root node of TSA.

This rule has 1701 hits in the corpus. See figures 2 and 3 for the constituency and for the resulting tectogrammatical structures.

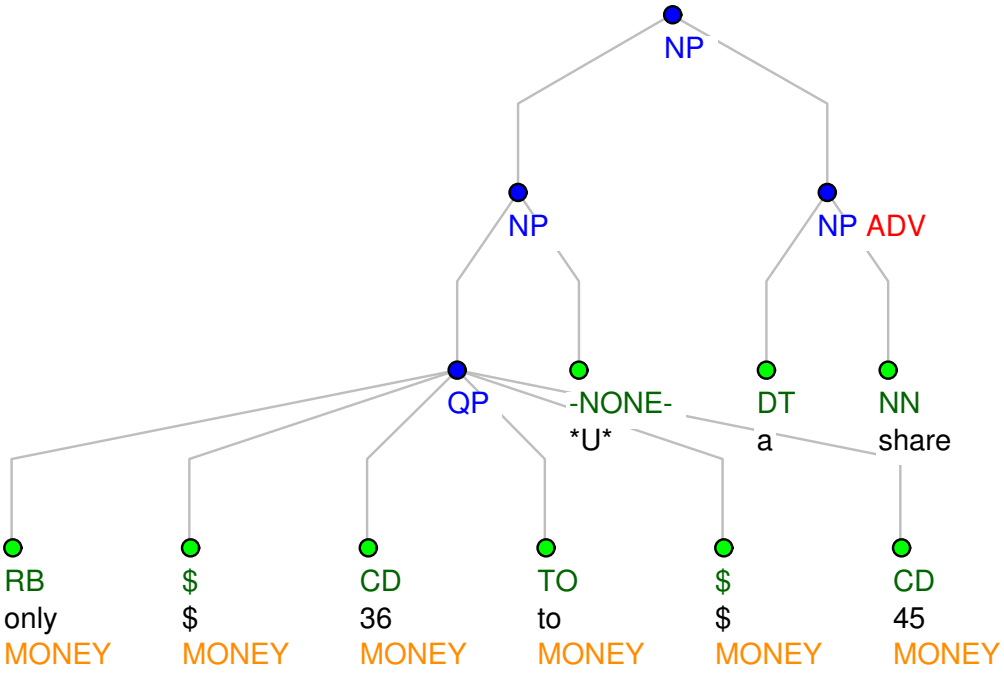


Figure 2. Example of a constituency structure of a phrase of the type "$600 a share"

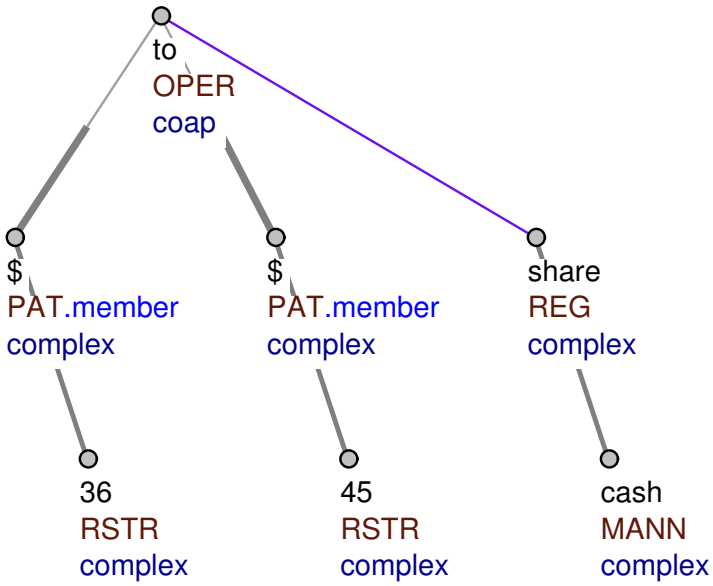


Figure 3. Example of a tectogrammatical structure of a phrase of the type "$600 a share"

Mixed Numbers

Whenever we found a mixed number (something like *3 2/7*) in the form of two terminal nodes with the tag CD, we transformed it into a tectogrammatical structure shown in Figure 4. There are 1351 mixed numbers in the corpus.

Phrases of the type "Boston, Massachusetts"

We are looking for an NP or an NML nonterminal with the phrase attribute value NAC and with the function LOC as its child (let us call it Node A). There has to be either an NP or an NML nonterminal or a noun (a terminal with a tag whose first two letters are NN) among the right siblings of the Node A – let us call it Node B. Node A has three or four childnodes. The second one is comma or left round bracket (a terminal node). If there is the fourth childnode, it has to be a comma or a right round bracket (again a terminal node). If the fourth childnode is not present and the

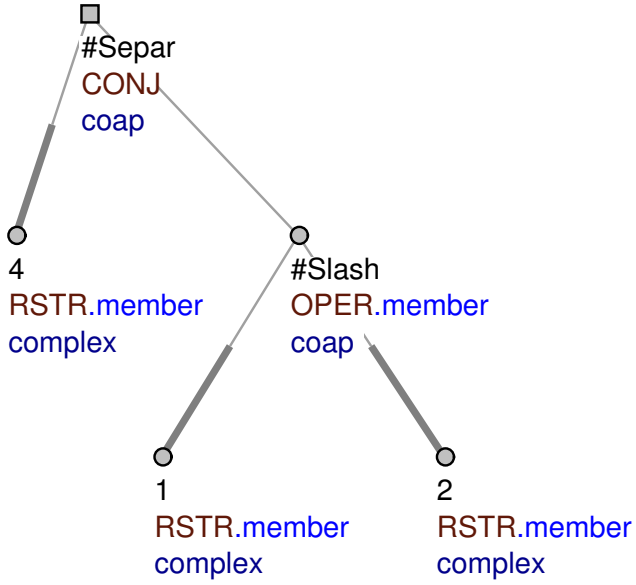


Figure 4. Example of a tectogrammatical structure of a mixed number

leftmost node of the Node B subtree satisfies the requirements, we can consider it to be the fourth child. The third childnode has to satisfy one of these three demands:

- It is an NP or an NML nonterminal and all the terminals in its subtree have the BBN-tag GPE:STATE_PROVINCE.
- It is a noun with the BBN-tag GPE:STATE_PROVINCE.
- It is a roman number (terminal node) with no BBN-tag.

The tectogrammatical counterpart of this structure is as follows. At first we identify the t-nodes which are roots of structures created from the subtrees rooted in the first and the third childnode of Node A (let's call them TR1 and TR3). Now we hang TR3 under TR1 and assign functors. TR1 should be LOC and TR3 gets the functor PAR. We also set the attribute is_parenthesis to 1 for each descendant of TR3 including the node TR3 itself. The second (and possibly the fourth) child of Node A is auxilliary and the corresponding a-node has to be properly referenced from the TR3 node. We also have to ensure that those auxilliaries do not exist as independent t-nodes and that they are not referenced from any other t-node.

There are 239 occurences of the described constituency structure in the corpus. See figures 5 and 6 for examples of the described structures. This script can with minor modifications be applied for structures consisting of person nouns and their political affiliations (e.g., *Leon Panetta, democrat*).

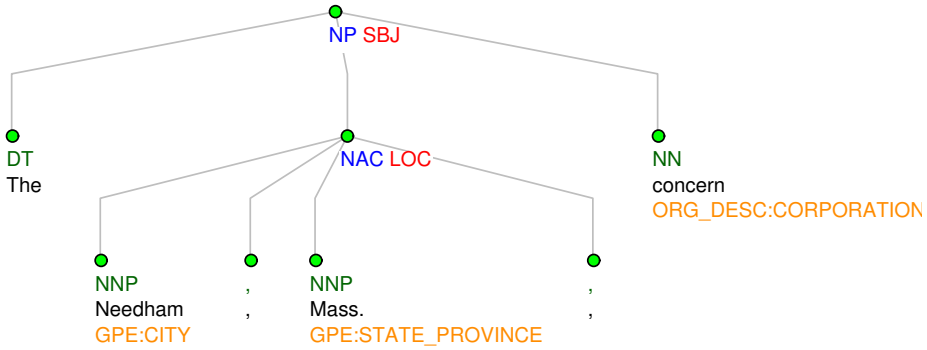


Figure 5. Example of a constituency structure of a phrase of the type "Boston, Massachusetts"

From August 2008 to November 2008 we created more than 60 rules (some of them became obsolete). The complete set of scripts was tested on one reference section (296 sentences, 7694 words). As a result we registered 1237 changes. We were measuring the agreement with manually annotated data, and we have achieved an approx. 4 % improvement in functors and 6 % in referencing auxilliaries, which is not a re-

Figure 6. Example of a tectogrammatical structure of a phrase of the type "Boston, Massachusetts"

ally substantial improvement. The agreement on other attributes has been more or less identical. However, in this case the quantity is not the only goal. Better consistency of the data is important as well. Besides applying annotation templates to structures relatively uninteresting from the linguistic point of view, such as mixed numbers, our rules annotated a number of complex and less frequent linguistically relevant phenomena throughout the corpus. Sometimes the given structures could not be processed completely, but the applied rules saved the annotators at least a part of their manual work. The overall effect of these measures on the annotation procedure would be too difficult to quantify, though. The outcomes of some rules were left for manual processing within the expert annotation (Section 10), which has positive effect on the annotation consistency as well.

7. Manual Annotation

The initial tectogrammatical annotations of English data (WSJ-PTB) date back to 2002 (Kučerová and Žabokrtský, 2002). The tectogrammatical trees have been built above analytical WSJ-PTB trees obtained by an automatic conversion from the original PTB bracketing into the format used by PDT 1.0 (Hajič et al., 2001). The automatically converted and generated data as well as this tentative manual tectogrammatical annotation were published along with parsed Czech parallel translations of WSJ-PTB as the **Prague Czech-English Dependency Treebank 1.0** (PCEDT 1.0, Cuřín et al., 2004). The PCEDT 1.0 with its 500 manually annotated tectogrammatical trees constituted the starting point for the efforts taken up 2004.

Due to substantial format changes of the "mother treebank", the Prague Dependency Treebank, before its second LDC release (Hajič et al., 2006) in 2006, the massive annotation of English data was postponed until the definite version of language-independent features of the new annotation scheme (Pajas and Štěpánek, 2006). In the meantime we concentrated on the conversion of PropBank (Palmer et al., 2004) into an FGD-compliant valency lexicon. In early 2006 we were able to convert the constituency trees into tectogrammatical trees with some of the modules which later became part of TectoMT. We also refined the initial version of the annotation manual.

Four annotators started the manual annotation in late 2006. During 2007, several more annotators were trained. At the moment we have four annotators working regularly, the rest being mostly in training, some having left the project, and some being on maternal leave. The interannotator agreement was measured approx. once a month in 2006 and early 2007. It has not been measured since March 2008, mainly because of the slow annotation pace in 2007, annotator fluctuation, and, since mid-2008, due to the intensive work on consistency checks, which all skilled annotators have been kept busy with.

The annotators work mostly off-line but send and retrieve the data via an SVN system. The data index as well as the work-progress stats are provided with a user-friendly web interface. The annotators currently correct the data produced in 2006 and 2007 by running the consistency-checking scripts upon each file and correcting the detected errors. The annotators are also asked to run the checks and correct the errors before submitting new files. A log of changes in the data is generated every month. It calculates uncorrected detected errors and the ratio of the amount of data vs. the amount of changes. Deviations from the average are examined and random samples are manually re-checked.

8. Consistency Checks

After the annotated data exceeded 12,000 trees (almost 25% of WSJ-PTB), we introduced consistency checks. Most of the scripts we use have been adopted from the Czech PDT-team (Štěpánek, 2006) and modified whenever necessary. We have added a few new, English-specific checking scripts, and we reuse some of our pre-annotation scripts. The checking scripts check mainly:

- **Paratactic structures**
  - Only a node of the appropriate type and with an acceptable functor is the root of a paratactic construction.
  - Each root of a paratactic construction has at least two descendants which are coordination members.
  - Only permitted combinations of functors occur in coordinated nodes.
- **References from t-nodes to a-nodes (content-word and auxiliary-word references)**

  - All a-nodes which represent alphanumerical tokens are referred to from the t-layer (except punctuation).
  - No a-node is referred to as a content-word from two non-generated t-nodes.
  - All t-nodes except nodes with t-lemma substitutes refer to a content word node at the a-layer.
  - A t-node, whose corresponding content-word reference at the a-layer is a noun in plural, may not refer to an a-node that represents the indefinite article.
  - T-nodes representing punctuation regarded as a content word (e.g., punctuation in asyndetic paratactic constructions) must not be represented as generated nodes.
- Tree structure
  - The effective root of the tree is either the main predicate (which might be an artificially inserted one) or the governing node of a noun group.
  - Nodes representing foreign words comply with all rules.
  - Nodes representing phrasemes comply with all rules.
  - T-nodes with t-lemma substitutes which are used for specific syntactic constructions (e.g. `#AsMuch`| `#Equal`| `#Total`) are never terminal nodes (leaves).
  - The technical root has only one descendant.
  - Each t-node has been assigned a functor.
- Valency
  - Each occurrence of a verb except *to be* and *to have* is assigned a valency frame from the lexicon.
  - The valency frame is complete according to the valency lexicon.
  - The valency frame assigned to a verb occurrence must exist in the lexicon (frames can be altered during the lexicon edits).
  - A copied verb has the same valency frame as the original.
  - All checks are dismissed when the verb node contains an annotator's comment regarding the lexicon.

This list presents only selected checks. There are approx. 80 checking scripts at the moment. Their amount is slowly but constantly growing. The annotators' comments serve as issues for new pre-annotation scripts, TectoMT improvements, or checking scripts. The comments regarding the valency lexicon are collected monthly in form of a log file with the examples and sentence identification, and they are e-mailed to the editor-in-chief of the lexicon. Besides, we are experimenting with a string-based consistency check of the tree structure and functor assignment. The data is searched for subtrees consisting of matching textual strings. Differences in the respective annotation resolutions for textual sequences are reported. This is a sample of the first tentative inconsistency survey:

```
previous month
  [month]([previous,RSTR]) 3
```

```
  [month]([previous,TWHEN]) 1
rate increase
  [increase]([rate,PAT]) 1
  [increase]([rate,ACT]) 1
size of the increase
  [size]([increase,ACT|of,the]) 1
  [size]([increase,APP|of,the]) 1
so far
  [far]([so,EXT]) 5
  [far]([so,MANN]) 1
```

Some of these reports help us uncover inconsistencies systematically made by the automatic pre-annotation and can be fixed. Many of them have to be manually checked by the annotators (see Section 10).

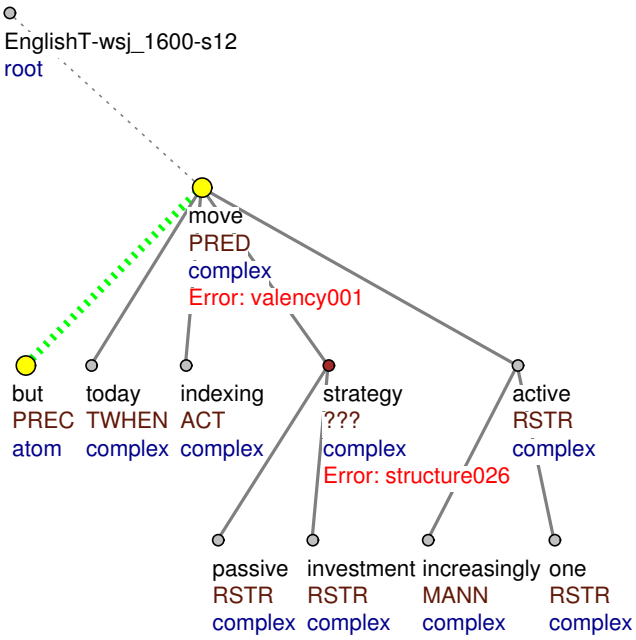
9. PEDT 1.0

The first 10 000 manually annotated and checked trees were released under the title PEDT 1.0. The CD contains the documentation along with relevant publications (including a draft version of this paper), the current version of the valency lexicon Engvallex (which is yet still being subject to revisions), and the ready-to-install package of TREd, the tree editor.

10. Discussion

The current annotation practice yields trees quite consistent in tree structure, some financial-speak specific fixed phrases, structured text like addresses and lists, and verbal valency. However, the annotation still remains inconsistent in functor assignment in adjectival and nominal phrases. We decreased this inconsistency by resigning on semantic labeling within named entities (all nodes in the subtree get the new functor NE - *Named Entity*), but we do not find this solution satisfactory, and we are going to introduce a systematic solution of noun valency in later versions of PEDT. We have tentatively merged the NomBank (Meyers et al., 2008) annotation with the PEDT data and are going to explore its benefits for an FGD-based annotation. While PropBank was driven by theoretical approaches quite similar to FGD, the NomBank approach might prove difficult to adopt. No conclusions can be drawn yet as we are just at the very start of the process.

In the next future we are going to continue improving the automatic pre-annotation by detecting problematic phrases and linguistic phenomena. As soon as the data has been annotated with the complete annotation, we will focus on the so-called **expert annotation**. This is annotation of selected structures across all corpus sections by one or a few 'expert' annotators. This procedure is meant for the annotation of particularly difficult or interesting phenomena. It is mainly supposed to further increase the



But today, indexing is moving from a passive investment strategy to an increasingly active one. Dnes se ale indexace posouvá od strategie pasivního investování ke stále aktivnějším.

Figure 7.

consistency of the annotation. Besides, it is meant to provide material for linguistic research. Figure 7 shows a TREd window with a highlighted expert-annotation task.

11. Conclusion

PEDT has been built to present the Prague Dependency Treebank-like annotation scheme to the global expert audience. The documents were chosen because of their original manual annotation and due to being a sort of a reference corpus in the NLP community, despite all linguistic objections that could be raised on how much the English used in American business press reflects the patterns of English in general. The annotation procedure has been improved, and so have the control mechanisms. Approximately 1/2 of WSJ-PTB has been annotated at the moment.

12. Acknowledgements

This work was funded in part by the Companions project[7] sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434, by FP7-ICT-2007-3-231720 (EuroMatrix+), and by the Czech Science Foundation (GA CR) project Nr. GA CR 405/06/0589.

Bibliography

Bies, Ann, Mark Ferguson, Karen Katz, and Robert MacIntyre. Bracketing Guidelines for Treebank II Style Penn Treebank Project, 1995. URL `http://www.cis.upenn.edu/~treebank/`.

Bradley, J., O. Mival, and D. Benyon. A novel architecture for designing by wizard of oz. In *Proceedings of CREATE08*, 2008.

Cinková, Silvie, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Annotation of english on the tectogrammatical level. Technical Report 35, UFAL MFF UK, 2006.

Cuřín, Jan, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. *Prague Czech-English Dependency Treebank Version 1.0*. Number LDC2004T25. Linguistic Data Consortium (LDC), University of Pennsylvania, 2004. ISBN 1-58563-321-6.

Gildea, Daniel. Corpus variation and parser performance. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2001.

Hajič, Jan, Eva Hajičová, Petr Pajas, Jarmila Panevová, and Petr Sgall. *Prague Dependency Treebank 1.0*. Linguistic Data Consortium, 2001.

Hajič, Jan, 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*. Linguistic Data Consortium, 2006.

Hajič, Jan, Silvie Cinková, Marie Mikulová, Petr Pajas, Jan Ptáček, Josef Toman, and Zdeňka Urešová. PDTSL: An annotated resource for speech reconstruction. In *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology*, 2008.

Klimeš, Václav. Transformation-based tectogrammatical dependency analysis of english. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, 2007.

Kulick, Seth, Ann Bies, Mark Libeman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, and Lyle Ungar. Integrated annotation for biomedical information extraction. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2004.

Kučerová, Ivona and Zdeněk Žabokrtský. Transforming Penn Treebank Phrase Trees into (Praguian) Tectogrammatical Dependency Trees. (78):77–94, 2002.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. *Treebank II*. Linguistic Data Consortium, 1995.

[7] `www.companions-project.org`

Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. *Treebank III*. Linguistic Data Consortium, 1999.

Meyers, Adam, Ruth Reeves, and Catherine Macleod. *NomBank v 1.0*. Linguistic Data Consortium, 2008.

Mírovský, Jiří. PDT 2.0 requirements on a query language. In *ACL 2008 Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies*, pages 37–45. Association for Computational Linguistics, 2008. ISBN 978-1-932432-06-0.

Oepen, Stephan. Beyond the science of the wall street journal. presentation slides at the Unified Linguistic Annotation (ULA) Workshop, TLT 2007, Bergen, December 5-6 2007. URL `http://tlt07.uib.no/ulaslides/stephan-ula.pdf`.

Pajas, Petr and Jan Štěpánek. XML-based representation of multi-layered annotation in the PDT 2.0. In Hinrichs, Richard Erhard, Nancy Ide, Martha Palmer, and James Pustejovsky, editors, *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pages 40–47, Paris, France, 2006. ISBN 2-9517408-2-4.

Palmer, Martha, Paul Kingsbury, Olga Babko-Malaya, Scott Cotton, and Benjamin Snyder. *Proposition Bank I*. Linguistic Data Consortium, 2004.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description I. *Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974.

Panevová, Jarmila. On Verbal Frames in Functional Generative Description II. *Prague Bulletin of Mathematical Linguistics*, 23:17–52, 1975.

Römer, Ute. A neo-firthian approach to academic writing: Uncovering local patterns and local meanings in the discourse of linguistics. oral presentation at the Conference of the American Association for (Applied) Corpus Lingusitics, Brigham Young University, Provo, Utah, USA, March 13-15 2008.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht:Reidel Publishing Company and Prague:Academia, 1986.

Vadas, David. Noun phrase bracketing guidelines. Technical report, School of Information Technologies, University of Sydney, 2007.

Vadas, David and James R. Curran. Adding noun phrase structure to the penn treebank. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

Weischedel, Ralph and Ada Brunstein. *BBN Pronoun Coreference and Entity Type Corpus*. Linguistic Data Consortium, 2005.

Štěpánek, Jan. Post-annotation checking of prague dependency treebank 2.0 data. *Prague Bulletin of Mathematical Linguistics*, (85):23–33, 2006. ISSN 0032-6585.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. Tectomt: Highly modular mt system with tectogrammatics used as transfer layer. In *Proceedings of WMT'08*, 2008.

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 105–114

Some Typological Characteristics of Czech and English and Other European Languages

Milan Malinovský

Abstract

Numerous attempts have been made to compare different types of languages from the viewpoint of their quantitative characteristics. We present a piece of research in five Indo-European languages focused on length of sentences and words, not drawing any general conclusions but bringing some partial findings about these languages' general measurable features.

1. Introduction

This article deals with the differences between quantitative characteristics of individual languages[1] comparing in particular the length of sentences and words in contemporary Czech and English. Several texts in German, French and Polish have also been included for comparison in order to illustrate the case in the three major European language families, i.e. the Germanic, Romance and Slavonic languages; in other words, to highlight certain structural differences between inflectional and analytic languages.

The aim of the study is to examine and possibly prove that in analytic languages sentences contain more words.

As well as the length of sentences, i.e. the number of words that they consist of, the focus is also on words (and the number of their constituent syllables). Detailed summaries of our research are given in Part Four.

All the texts examined are contemporary; always, a pair (triad, …) of original + translation are used to illustrate the point. In every case, only published translations were used, in order to avoid the danger of tentative subjectivity of one's own translation while judging the individual phenomena.

In spite of the fact that we have used different genres (cf. Literature) and researched a great number of texts[2], we are fully aware of the fact that our findings and

Please cite this article as: Milan Malinovský, Some Typological Characteristics of Czech and English and Other European Languages. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 105–114.

conclusions are only preliminary and should be interpreted with caution. For real quantitative conclusion to be drawn about the relationships between pairs of texts in various languages (and between different types of language), a more extensive study, utilizing a greater corpus of material, would be required.

The majority of texts used are drawn from the field of philosophy, as we believe they are fairly appropriate for research of this kind because of their (neither very concise nor excessively verbose) style.[3] The other texts come from history, political science, ecology, sociology, and so forth.

A general comment on translation: if it is true that a (good) translator is in fact a co-author of a text (and therefore not only a transposer) then it is not strictly possible to make an entirely objective comparison about the lengths of equivalent sentences in two languages (however representative they may be).

2. Method of research

The research proceeded as follows: owing to the different formats of the individual language sources, we transposed the results of calculations to 'standard' pages (30 lines times 60 keystrokes) for the comparison be utterly precise. Data was acquired by comparing these standard pages for the length of the text in the first and second (third, …) language.

2.1. Disparateness of languages: sentences

As might be expected, individual texts in the languages examined were of different lengths. Because of this, it was not technically possible to compare the lengths of parallel sentences reciprocally. Nonetheless, some comparisons of individual sentences were made; this was the case where the total number of sentences in a given text was not the same for the two (or more) languages being compared. One of the main reasons for such a disparity was the fact that in some instances a compound sentence in the original text (common in French[4], for example) was broken up into two (or three) sentences in translation.

As mentioned above, the majority of the source material was taken from philosophical texts. Now, let us examine the composition of a text in its individual translations in the following compound sentence (Patočka, 1972, p. 14; its translations are given in Literature):

(**French**): *Husserl ne l'a pas d'ailleurs développée complètement; l'ouvrage est resté inachevé; la solution qui lui servait de fil conducteur se laisse définir le plus clairement à l'aide d'une conférence prononcée par Husserl à Vienne en 1935; là, il brosse l'image d'un monde naturel, résultat d'achèvements (Leistungen) en commun des subjectivités dont ce monde constitue précisément un trait d'union sans substantialité propre; c'est le retournement complet de la vue physicaliste, il est vrai, un reversement plus que copernicien - au lieu d'îlots de subjectivité contournés d'une mer infinie de structures physicalistes, schemas abstraits dépourvus de*

qualités et de vie, on a une mer d'intersubjectivité dite transcendentale qui entoure des unités objectives constituées en unités de sens par le travail en commun des consciences communiquantes.

(**Czech**): *Husserl ji také nikdy úplně nerozvinul, jeho dílo zůstalo nedokončeno. Řešení, které mu bylo vodítkem, lze nejzřetelněji určit z přednášky, kterou proslovil ve Vídni roku 1935. Tam načrtává obraz přirozeného světa jako výsledku společných výkonů (Leistungen) subjektivit, jimž tento svět není ničím jiným než pojítkem bez substanciality ve vlastním smyslu. Je to vskutku úplné převrácení fyzikalistického pohledu, obrat víc než koperníkovský: místo ostrůvků subjektivity obklopených nekonečným mořem fyzikalistických struktur, abstraktních schémat zbavených kvalit i života, je zde moře intersubjektivity, zvané transcendentální, jež obklopuje objektivní jednotky konstituované společnou prací komunikujících vědomí v jednotky smyslu.*

(**English**): (The first sentence of the compound sentence stayed untranslated.) *We can best glean the solution that guided him from the lecture he delivered in Vienna in 1935. There he sketches the image of a "natural" world conceived as a product of the common achievements (Leistungen) of subjectivities for whom this world is nothing but a common link devoid of any genuine substantiality. It is, in truth, a complete inversion of the physicalist view, a more than Copernican reversal. In place of islets of subjectivity surrounded by an infinite sea of physicalist structures, of abstract schemata stripped of qualities and of life, we now have a sea of allegedly transcendental intersubjectivity surrounding objective unities constituted as unities of meaning by the common efforts of communicating consciousnesses.*

(**Polish**): *Zresztą Husserl nie rozwinął jej v pełni; dzieło pozostało niedokończone. Rozwiązanie, które posłużyło mu za nić przewodnią, można najwyraźniej określić w oparciu o wykład wygłoszony przez Husserla w Wiedniu w r. 1935. Maluje on w nim obraz świata naturalnego, wynik wspólnych dokonań (Leistungen) subiektywości, w których ów świat konstytuuje właśnie linię łączącą bez właściwej istotowości; jest to całkowite odwrócenie się od fizykalnego sposobu widzenia, przewrót większy niż kopernikański: zamiast wysepek subiektywności opasanych nieskończonym morzem struktur fizykalistycznych, abstarkcyjnych schematów pozbawionych jakości i życia, mamy morze intersubiektywności nazywanej transcendentalną, które otacza obiektywne jedności ukonstytuowane w jednościach sensu wspólną pracą komunikujących się świadomości.*

(**German**): *Hinzu kommt, daß Husserl sie nicht vollständig entfaltet hat; das Vorhaben blieb unvollendet. Im Wiener Vortrag von 1935 wird sein Leitgedanke am deutlichsten. Husserl zeichnet die natürliche Welt hier als Resultat von gemeinschaftlich erbrachten Leistungen der Subjekte, für welche diese Welt ein vereinigendes Element ohne eigene Substantialität darstellt. Er vollzieht eine totale Umkehrung der physikalistischen Sichtweise, eine mehr als kopernikanische Revolution: Statt Inseln der Subjektivität, die von einem unendlich weiten Meer physikalistischer Strukturen und abstrakter, lebloser und aller Qualität beraubter Schemata umgeben sind, haben wir hier das Meer einer als transzendental bezeichneten Intersubjektivität, welches objektive Einheiten, Sinneinheiten umschließt, die in der gemeinsamen Anstrengung kommuniziernder Bewußtseine konstituiert wurden.*

If a semicolon is considered a symbol of apposition within a compound sentence, it appears that the compound sentence of the French original (one of the most developed examples in our excerpts, consisting of 121 word units) was, when translated into four languages, translated as three or four separate sentences. This phenomenon was observed more than once in this text. Remarkably, both translators into Germanic and Slavonic languages chose to do this the same way (there are three sentences in the Polish translation of similar composition to the Czech translation). Moreover, the longest compound sentence was found in this very text (from the complete material researched): - the French version comprised 135 words, while the longest English compound sentence had 129 words. Incidence of such long sentences was more scarce in the other languages examined.

Generally speaking, the greater number of words observed in the sentences of analytic languages may be explained by the fact that in these languages the grammatical values are expressed by auxiliary words (referring, for example, to tense, person, and the category of determination). Similarly, while the Slavonic languages express syntactic relations of nouns prototypically with the help of case endings, the Germanic and Romance languages apply prepositions.

2.2. Disparateness in languages: words

In analyzing words, we considered mainly the graphic form of the word. English is a language abundant in one-syllable words. Let us illustrate this by using an excerpt of the English translation of Karteziánství a fenomenologie (Patočka, 1989, p. 307): *This world by means of which is not a world of beings, rather, it is I, a being that understands beings, that is open for them; and the core of this world which I am is a …*, in which thirty words out of thirty-seven are one-syllabic.

It is of interest whether the phonetic realization in this case coincides with the written form, or, if and when only one out of two written syllables is pronounced (in Slavonic languages no differences of this kind are found). In order to note such differences, we considered both possibilities. This is restricted mainly to English; as the research has revealed, when counting the graphic syllables, English is rather consistent with Czech, but in its phonetic realization an English syllable unit is more likely to be a word (in comparison with Czech, it is multi-lettered, cf. *since, those, scratched*, etc.), see Part Four.

3. Conclusions

Contrary to our expectations, not many long words were found in German; in English, Czech, French, Polish and German, words longer than seven syllables are rare. The one and only exception was a 10-syllable German adjective *informationstheoretischer*, in Czech there was a 9-syllable genitive form of *psychofyziologického* and 11-syllabic locative form *čtyřiadvacetikilometrovém*. For German (and English) conso-

nantal clusters are fairly common: in spite of the fact that the words look long, they are not very extensive syllablewise.

The excerpts indicate that presuppositions following from the typological characteristics of individual languages would be confirmed should more extensive language material be examined. In Czech, which is inflectional, there are fewer words on average than in English, although words are longer. The total length of a text (measured in syllables) is longer in Czech which means that a Czech word form is often longer than the corresponding English analytic form, and that Czech derived words are longer than English motivated word-forms. Czech can often easily do without auxiliaries. The length of the English written text is produced by specific features of English spelling. It should be noted that in philosophical texts, sentences are a little longer on average than in texts of other language styles. As mentioned, our collection of texts was too small for us to be able to draw any general conclusions from these partial findings.

As for the other languages, our findings suggest that in the main indices, French is in recognizable accordance with English as is Polish with Czech – again, in agreement with prediction based on typology. In some respects, German is closer to the Slavonic languages (in terms of word length), while in others it is closer to English and French (the length of sentences).

If research aimed in such a direction continues, the results may be very interesting for both general and comparative linguistics.

4. Statistical indices

The following are the numbers of pages (= standard pages), sentences, words, and syllables of the excerpted texts (in Czech and in English the extent of the texts was the same; other languages sets were restricted).

	Number of pages		sentences		words***	
Cz	264.6	81.7*	2,858**	693*	63,325	19,878*
En	278.8	93.5*	2,716**	714*	76,546	26,952*
Ge	101.4	101.4*	695	695*	24,115	24,115*
Fr	26.9	26.9*	195	195*	7,088	7,088*
Po	25.4	25.4*	202	202*	5,944	5,944*
Total	697.1	328.9*	6,666	2,499*	177,018	83,977*

Table 1. Total number of pages, sentences, and words in the texts of five languages (Czech, English, German, French, Polish).

* Part appertaining to philosophical text.
** Although the same number of texts was examined in both Czech and English, the significant difference in the number of sentences is due mainly to one of the sources in which the Czech translator divided the compound sentence of the original into more than one sentence.
*** The number of words was taken from the average number of words on a page (every fifth page).

As far as comparison of the number of words in one philosophical text (Patočka, 1972, see its translations in Literature) is concerned, the sequence is as follows: English (highest number of words), French, German, Polish, Czech (lowest).

	words on page		Average number of words in sentence		syllables in word	
Cz	239.5	246.4*	23.7	28.4*	2.4	2.4*
En	276.6	287.6*	29.5	36.4*	1.9^{I}	1.9^{*I}
					1.7^{II}	-
Ge	-	241.2*	-	33.6*	-	2.1*
Fr	-	263.7*	-	36.3*	-	2.0*
Po	-	236.2*	-	29.7*	-	2.4*

Table 2. Average number of words on page, words in sentence, and syllables in each word.

* Philosophical text.
^{I} Graphic form (cf. par. 2.2.).
^{II} Phonetic realization (ibid.).

Average number of n-syllabic words (in %)				
	Cz		**En**	
1 syllable	10.17	12.55 *	33.21 II	30.26 *I
2 syllables	24.56	23.80 *	23.86 II	18.28 *I
3 syllables	27.28	25.50 *	23.46 II	19.55 *I
4 syllables	21.10	17.47 *	15.14 II	16.87 *I
5 syllables	11.39	13.15 *	3.26 II	8.64 *I
6 syllables	3.54	5.11 *	0.76 II	4.08 *I
7 syllables	1.37	2.40 *	0.17 II	2.28 *I

Table 3. Percentage of occurrence of words according to number of syllables: Czech, English.

* Philosophical text.
I Graphic form.
II Phonetic realization.

Occurance of words (in %)			
	Ge	**Fr**	**Po**
l syllable	22.31 *	24.01 *	12.23 *
2 syllables	23.38 *	21.30 *	25.16 *
3 syllables	22.15 *	22.90 *	26.44 *
4 syllables	16.32 *	16.70 *	16.69 *
5 syllables	9.68 *	5.40 *	12.61 *
6 syllables	4.71 *	4.91 *	4.85 *
7 syllables	1.42 *	4.43 *	1.95 *

Table 4. Percentage of occurrence of words according to number of syllables: German, French, Polish.

* Philosophical text.

4.1. Ratio of the number of words in a sentence of philosophical texts

Sequence of languages: English, Czech, German (Patočka, 1976). In the given text, the English sentence has an average of 28.24 % more words than a Czech one, and 8.20 % more than a German one, that is, a German sentence has an average of 15.63 % more than a Czech one.

4.1.1. English, Czech, German, French, Polish (Patočka, 1972)

In the text provided, the English sentence has an average of 19.23 % more words than a Czech one, 6.37 % more than a German one, 9.11 % fewer than a French one, and 11.36 % more than a Polish one.

And vice versa: the Czech sentence has an average of 10.97 % fewer words than a German one, 23.77 % fewer than a French one, and 6.61 % more than a Polish one.

NOTES

[1] Within the field of quantitative linguistics, L. Uhlířová (1995) applies quantitative analysis to examine word length in Czech texts. Her findings lead to reasonable generalizations, modelled in a mathematically exact way by a general type of statistical distribution. Uhlířová claims that the distribution of word length in text is not only text-specific, text-distinctive, or even "random", but that it is subject to a more general probabilistic law. Her minute research brings evidence that the probabilistic laws found out for Czech are in full accordance with analogical probabilistic laws of the same generality which hold for other languages as well.

[2] The author expresses his thanks to Eva dos Reis for collating the statistical data.

[3] Nacherová (1994) in her study points out that "a linguistic dimension of philosophical problems has its substantiation since it is the very linguistic aspect that is, as one of manifold aspects of formulation of philosophical problems, the most significant part within the process of formulation."

[4] For completeness, let us mention that Patočka's study on "Edmund Husserl's Philosophy of the Crisis of the Sciences and His Conception of a Phenomenology of the 'Life-World'" was originally written (and later presented in Warsaw in 1971) in French.

REFERENCES

NACHEROVÁ, S., 1994. Jazyk jako prostředek formulace filozofických problémů. Charles University College of Arts.

SGALL, P., 1993. Typy jazyků a jejich základní vlastnosti. Slovo a slovesnost, 54, 271-277.

SKALIČKA, V., 1941. Vývoj české deklinace. Praha.

SKALIČKA, V., 1951. Typ češtiny. Praha.

UHLÍŘOVÁ, L., 1995. On the Generality of Statistical Laws and Indiviuality of Texts. A Case of Syllables, Word Forms, their Length and Frequencies. Journal of Quantitative Linguistics 2, 238–247.

LITERATURE

DIALOGUE, 1983. (Articles: Avoiding the Entropy Trap. Current Issues: Change and Continuity; Pluralism and Fundamentalism in the United States. Feminism, Family and Community. The New Conservatism.) The United States Information Agency Journal, 61, 16–33.

GRAY, W., HOFSTADTER, R., OLSON, K. W., 1980. An Outline of American History. United States Information Agency, 1–52.

GRAY, W., HOFSTADTER, R., OLSON, K.W., 1980. Nástin amerických dějin. Embassy of the United States of America. Praha, 1–52.

PATOČKA, J., 1989. Cartesianism and Phenomenology. In: Jan Patočka, Philosophy and Selected Writings. The University of Chicago Press, Chicago & London, 285–326.

PATOČKA, J., 1991. Cartesianismus und Phänomenologie. In: Jan Patočka, Die Bewegung der menschlichen Existenz. Phänomenologische Schriften II, Ernst Klett Verlag für Wissen und Bildung GmbH, Stuttgart, 360–414.

PATOČKA, J., 1991. Die Philosophie der Krisis der Wissenchaften nach Edmund Husserl und sein Verständnis einer Phänomenologie der Lebenswelt. In: Jan Patočka, Die Bewegung der menschlichen Existenz. Phänomenologische Schriften II, Ernst Klett Verlag für Wissen und Bildung GmbH, Stuttgart, 310–329.

PATOČKA, J., 1989. Edmund Husserl's Philosophy of the Crisis of the Sciences and His Conception of a Phenomenology of the "Life-World". In: Jan Patočka, Philosophy and Selected Writings. The University of Chicago Press, Chicago & London, 223–238.

PATOČKA, J., 1980. Filosofie krize věd podle Edmunda Husserla a jeho koncepce fenomenologie "Světa našeho života". In: Přirozený svět a pohyb lidské existence. Samizdat edition, 3, Praha, 3.3.2-3.3.21.

PATOČKA, J., 1987. Filozofia kryzysu nauki wedlug Edmunda Husserla i jego koncepcja fenomenologii "Świata życia". In: Jan Patočka, Świat naturalny i fenomenologia. Papieska Akademia Teologiczna, Wydzial filozoficzny, Kraków, 140–157.

PATOČKA, J., 1976. Karteziánství a fenomenologie. In: Samizdat edition, Praha, 1–38.

PATOČKA, J., 1972. La philosophie de la crise des sciences d'après Edmond Husserl et sa conception d'une phénoménologie du "monde de la vie". In: Archiwum historii, filozofii i myśli społecznej, 18, Warszawa, 3–18.

SAVICKÝ, P., HLAVÁČOVÁ, J., 2002. Measures of Word Commonness. Journal of Quantitative Linguistics, Swets & Zeitlinger, Vol. 9, No. 3, 215–231.

SPEKTRUM, 1983. (Articles: Jak se vyhnout nebezpečí entropie. Nový konzervatismus. Současné problémy: Změna a kontinuita; pluralismus a fundamentalismus ve Spojených státech. Ženské hnutí, rodina a společnost.) Embassy of the United States of America, 43, Praha, 16–33.

WESTWOOD, J., 1988. The Pictorial History of Railroads. Gallery Books, Smith Publishers Inc., New York, 8–143.
WESTWOOD, J., 1994. Železnice - obrazové dějiny. Columbus, Praha, 8–143.

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 115–134

Improving English-Czech Tectogrammatical MT

Martin Popel, Zdeněk Žabokrtský

Abstract

The present paper summarizes our recent results concerning English-Czech Machine Translation implemented in the TectoMT framework. The system uses tectogrammatical trees as the transfer medium. A detailed analysis of errors made by the previous version of the system (considered as the baseline) is presented first. Then several improvements of the system are described that led to better translation quality in terms of BLEU and NIST scores. The biggest performance gain comes from applying Hidden Tree Markov Model in the transfer phase, which is a novel technique in the field of Machine Translation.

1. Introduction

We report on a work in progress on developing English-Czech machine translation (MT) system called TectoMT.[1] This system participated at the Workshop on Statistical Machine Translation (WMT) in 2008 and 2009 (Žabokrtský et al., 2008; Bojar et al., 2009). The translation is carried out in three phases: analysis, transfer and synthesis. Similarly to Bojar et al. (2008a), the transfer phase implemented in TectoMT uses tectogrammatical trees and exploits the annotation scheme of the Prague Dependency Treebank, but (unlike in the cited work) the transfer does not use Synchronous Tree Substitution Grammars.

In Section 2, we shortly describe our baseline system. In order to identify its most prominent errors, their types and sources, we have manually annotated a sample of 250 sentences; the resulting error analysis is presented in Section 3. Modifications of our baseline system and their evaluation are described in Section 4. One of the most important modifications – the introduction of Hidden Markov Tree Models to the transfer phase – is explained in Section 5.

[1] http://ufal.mff.cuni.cz/tectomt/

Please cite this article as: Martin Popel, Zdeněk Žabokrtský, Improving English-Czech Tectogrammatical MT. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 115–134.

2. Baseline system

The TectoMT version which participated in WMT 2009 is used here as the baseline system. In this version, the translation process consists of about 80 steps implemented in so-called *blocks* (basic TectoMT processing units). We give here only a brief overview.

2.1. Analysis

Each sentence is tokenized (roughly according to the Penn Treebank conventions), tagged by the English version of the Morce tagger (Spoustová et al., 2007), and lemmatized in order to obtain the morphological layer (m-layer). Maximum Spanning Tree dependency parser (McDonald et al., 2005) is applied to create analytical trees (a-trees). These are then converted to the tectogrammatical ones using a sequence of heuristic blocks: Functional words (such as prepositions, subordinating conjunctions, articles etc.) are removed. Only morphologically indispensable categories (called *grammatemes*) are left with the tectogrammatical nodes (t-nodes). The information about the original syntactic form is stored in attributes called *formemes*.[2] Several other attributes are filled (e.g. functors, coreference links, named entity types).

2.2. Transfer

First, the topology of target-side t-trees is copied from source-side t-trees. Probabilistic dictionaries provide n-best lists of lemmas and formemes. In the baseline scenario, formemes are translated independently for every node as the most probable variant from the n-best list. Consequently, lemmas are translated as the most probable variant that is compatible with the already chosen formeme. The compatibility is ensured by a set of rules. Additional rule-based blocks are used to translate other t-layer attributes (grammatemes) and to change topology and word order where needed.

2.3. Synthesis

In this phase Czech analytical trees are created from the tectogrammatical ones (auxiliary nodes are added), but the process of synthesis continuously goes on (morphological categories are filled, word forms are generated), so that in the last block, the sentence is generated by simply flattening the tree and concatenating the word forms.

[2]Formemes are not used in Prague Dependency Treebank, they were introduced to TectoMT with regards to the needs of MT (Žabokrtský et al., 2008). Formemes cannot be considered as a genuine component of the tectogrammatical layer of language description, but they facilitate formalizing the relation between tectogrammatics and surface syntax and morphology. Examples of formemes are: n:subj – semantic noun in subject position, n:for+X – semantic noun with preposition *for*, v:because+fin – semantic verb as a head of subordinating finite clause introduced by *because*, v:without+ger – semantic verb as a gerund after *without*, adj:attr – semantic adjective in attributive position.

3. Error annotations and analysis

Manual analysis of translation errors is expensive and time-demanding, but it can identify types and sources of errors. This knowledge is very helpful for developers of MT systems, that perform transfer on some level of abstraction that is higher than simple phrase-to-phrase. There are many papers on manual evaluation of MT errors, (e.g. Koehn and Monz, 2006), but they are mostly limited to scoring *fluency* and *adequacy*. Some papers (Hopkins and Kuhn, 2007) use manual analysis based on some form of *edit distance*, i.e. the number of editing steps (of various types) needed to transform the system output into an acceptable translation. One of the most detailed manual analysis frameworks is the Error Classification Scheme described in Vilar et al. (2006), which classifies errors into a hierarchical structure.

Our proposed error analysis framework is similar to that of Vilar et al. (2006), but instead of three hierarchical properties of errors (*type, subtype* and *sub-subtype*) we have five properties: *seriousness, type, subtype, source* and *circumstances*. Errors are marked in text by *error markers* which the annotator simply inserts in front of relevant words. If needed, one word may have more than one error marker. Every error marker describes all the five properties of an error. Details about the error analysis framework including several examples of annotated text can be found in Popel (2009).

Source		Description	#errors
Analysis	`tok`	tokenization errors	16
	`tagger`	PoS tagging errors	37
	`lem`	lemmatization errors	1
	`parser`	errors associated with parsing and related tasks (building a-layer from m-layer)	300
	`tecto`	tecto-analysis errors (building t-layer from a-layer)	68
Transfer	`noniso`	errors caused by the assumption of t-tree isomorphism (which is currently required in the TectoMT translation)	109
	`other`	other errors associated with the transfer (translation of lemmas, formemes, grammatemes, noun gender assignment,...)	845
	`syn`	synthesis errors (generation of text from the target t-layer)	42
	`?`	source unknown	45
		total	1463

Table 1. Distribution of translation errors with respect to their sources

Circumstance	**Description** – errors associated with …	**#errors**
ne	named entity	104
num	numbers (numerals)	40
coord	coordination or apposition	117

Table 2. Distribution of translation errors with respect to their circumstances

The first author of this paper annotated 250 sentences. Tables 1 – 3 show numbers of occurrences of errors for categories *source*, *circumstances*, *type* and *subtype*.[3] As expected, most errors lie in the transfer phase. Only 8% of errors are caused by the unfulfilled presumption of isomorphic t-trees, whereas 56% are other transfer errors that could be repaired within the node-to-node transfer paradigm.[4] Another notable source of errors is parsing – 21%. We have found that 39% of these parsing errors are associated with coordinations. Also other observations indicate that the parsing of coordinations is a significant problem in TectoMT: There were 89 coordinations in the test data and more than half of them is parsed incorrectly which results in 1.13 serious errors per coordination on average.

The most common type of error is a wrong choice of lemma (lex = 37%), followed by a wrong choice of formeme (form = 33%) and grammateme (gram = 10%). Several subtypes of lex were classified (compound words, errors associated with named entities or reflexivity of lemmas), but most lex errors remain unclassified. We have not carried out any subclassification of form errors except registering problems with the Czech formeme *v:že+fin*. Among subtypes of gram, the most problematic one is the choice of correct gender[5] and number.

[3]We have also distinguished between serious and minor errors, but for brevity, this last category (*seriousness*) is not shown in the tables. Errors with types punct, order and case were mostly minor, other types were mostly serious.

[4]This finding is for us – TectoMT developers – very important. Of course, we are aware of the cases that cannot be translated within the node-to-node paradigm (e.g. *take part* → *účastnit se*, *make X public* → *zveřejnit X*) and we plan to solve them in TectoMT in future. However, those 8% is a relatively small number and thus we primarily focus on more frequent types of errors.

[5]It is well known that when translating from English to Czech, gender must be sometimes guessed from context, since English does not indicate gender for verbs, but Czech does.

Type Subtype	Description	# errors
lex	wrong lemma	**544**
asp	wrong aspect of a verb	6
se	wrong reflexivity, e.g. t-lemma *stát_se* instead of *stát*	15
neT	named entity translated, but should remain unchanged	11
neU	named entity unchanged, but should be translated, because the original form is not acceptable in the target language	4
neX	assumed named entity unchanged, but should be translated, because it is not really a named entity (*Bill was approved.*)	8
com	unchanged word due to an unprocessed compound word	13
unk	unchanged (possibly missing in the dictionary) word other than neU, neX and com	6
other	default value when no subtype is specified	481
form	wrong formeme	**481**
ze	formeme v:že+fin instead of v:rc or v:fin	39
other	default value when no subtype is specified	442
gram	wrong grammateme and related errors	**151**
gender	wrong grammateme of gender (feminine, neuter, masculine animate, masculine inanimate)	41
person	wrong grammateme of person (first, second, third)	3
number	wrong grammateme of number (singular, plural) except cases classified as numberU (see below)	26
tense	wrong grammateme of tense (simultaneous, preceding, subsequent)	5
mod	wrong verbal, deontic, dispositional or sentence modality	18
deg	degree of comparison (positive, comparative, superlative)	4
neg	negation (affirmative, negative)	19
svuj	switched m-lemma *svůj* with *jeho*, *její*, …	17
numberU	number unchanged, but should be changed e.g. *Ministry of Finance*(sg) → *Ministerstvo financí*(pl)	8
other	default value when no subtype is specified	10
phrase	phrases, idioms, deep syntactic structures that cannot be translated node-to-node.	**81**
miss	missing words that are not covered by the types above	**19**
extra	superfluous words that are not covered by the types above	**36**
punct	punctuation errors	**64**
brack	missing, superfluous or displaced brackets	24
other	default value when no subtype is specified	40
order	wrong word order (except cases classified as punct)	**64**
case	switched upper/lower case	**23**

Table 3. Distribution of translation errors with respect to their types and subtypes

4. Modifications and their evaluation

We have implemented several modifications to our system in order to improve the translation quality. We present here an overview of the most important modifications.

4.1. Analysis

- We have done slight modifications of the tokenization, so for example *3rd* is not split into two tokens anymore.
- We have developed a new implementation of the lemmatization – it fixes some errors made by the original implementation and it is more than 70 time faster.
- We have improved the parsing in the following two ways without actually changing the parsing algorithm or its features:
  We have implemented rule-based blocks that fix some frequent "mistakes" made by the parser. Some of these "mistakes" are real errors, but some are caused by different parsing guidelines concerning for example auxiliary verbs or multiword prepositions.
  We noticed that in the analysed sample, there are 22 sentences with parentheses and only 2 of them are parsed correctly. Sometimes the parenthesis is incorrectly divided and each part attached to another parent. Sometimes there are parsing errors also in the rest of the sentence, but these errors disappear, when we try to parse the sentence without the parenthesis. By parsing the parenthesis and the rest of the sentence independently we ensure that the parenthesis remains in its own subtree, which is then attached to the main sentence tree.
- *Analytical function* is the key attribute of the a-layer. It specifies the type of dependency relation of a node to its governing node. The baseline system used analytical functions only to mark coordinations and prepositions. We have added a block that recognizes also other types of dependencies, e.g. subject, object, predicate, adverbial, attribute, auxiliary verb, article. As there are no guidelines for English analytical functions yet, we had to decide how to annotate phenomena without any Czech equivalent (articles, phrasal verb particles, infinitive marker *to*, negation *not*). For details see Popel (2009).
- We have implemented a new procedure that builds the t-layer from the a-layer. It exploits analytical functions, which makes the procedure more clear. It deals with special cases that were not solved properly in the baseline implementation. We have aimed at a robust implementation that can handle also some cases with inaccurate parsing. Also, we have aimed at a modular implementation – the procedure is divided into five blocks and three of them are language independent.

4.2. Transfer

Our new design of the transfer phase is more modular. We have created 10 new blocks which can be combined in various translation scenarios.

- rule-based blocks that translate some special phenomena, e.g. ordinal numerals (*1st*, *32nd*, *999th*) can be translated by a simple rule (to *1.*, *32.*, *999.*),
- blocks that save all translation variants proposed by the dictionaries[6] to the attributes of nodes,
- blocks that rerank these variants using either more detailed models (e.g. valency formeme translation dictionary) or rules (e.g. the rule that filters out verbal lemmas whose aspect is incompatible with the given context),
- a block that selects the optimal combination of lemmas and formemes for every node using Hidden Tree Markov Model (HMTM). This is discussed in detail in Section 5.

4.3. Synthesis

- Word forms are generated according to lemmas and morphological categories. In theory, the word form should be fully specified by the lemma and morphological tag and there is a deterministic Czech word form generator suited for the task (Hajič, 2004). In practice, the tags are "underspecified", because they are generated from the t-layer that was translated from English. Some categories are not known and must be guessed.
  We have created a module that includes a subroutine for generating all forms of a given lemma whose tags match a given regular expression. The word forms are sorted according to their frequency. The model was trained on the corpus SYN (with 500 million words) of Czech National Corpus.[7]
- Commas (more precisely, a-nodes corresponding to commas) are added to boundaries of finite clauses. We have refined the rules for special cases such as quotations. We have also created a new block that coindexes all nodes belonging to the same finite clause.

4.4. Evaluation

Aside from evaluating the total difference of BLEU score between the baseline and our new modified version of TectoMT (see Table 4), we want to evaluate also the effect of each modification separately. However, many of the modified blocks would not work with the baseline system, because we have meanwhile added some functionality also to TectoMT internals. Therefore, we have chosen the opposite way – we take the new modified system, substitute one or more blocks with their baseline equivalent

[6] We use a probabilistic dictionary of lemmas (Rouš, 2009) created from the parallel corpus CzEng (Bojar et al., 2008b) and other sources as a replacement for the older PCEDT dictionary (Cuřín et al., 2004). For the translation of formemes we use the so-called valency formeme translation dictionary, which models the probability of target formeme given source formeme and source parent's lemma, and simple formeme-to-formeme dictionary as a fallback.

[7] `http://www.korpus.cz`

system	BLEU	NIST
baseline	0.0659	3.9735
modified	0.0981	4.7157

Table 4. BLEU&NIST evaluation of the new system

(called "original implementation") and we measure the impairment caused by the absence of the modification in question. This value can be loosely interpreted as an improvement caused by the modification, but we must be careful, because there may be "interferences" between some blocks.

We divided the evaluation data of WMT 2009 Shared Task (`news-test2009`) into two parts:

- First 250 sentences were used for the manual annotation of errors of the baseline implementation (as presented in Section 3).
- The rest (2 777 sentences) is our test set. Tables 4 and 5 are evaluated on this test set.

Modification	diff (BLEU)	diff (NIST)
original analysis	**0.0078**	**0.1363**
—original tokenization	0.0008	0.0105
—original lemmatization	0.0006	0.0294
—original parsing	0.0072	0.3006
—original building of t-layer	0.0053	0.1024
original transfer	**0.0171**	**0.4189**
—without HMTM	0.0130	0.2483
original synthesis	**0.0031**	**0.0621**
original quotation marks	0.0085	0.1757
all above together	**0.0322**	**0.7422**

Table 5. Modifications of analysis, transfer and synthesis

Note on BLEU&NIST scores reliability

Correct opening and closing quotation marks are in Czech „ and “. These symbols are produced by TectoMT as a translation of English “ and ”. However, reference translations in WMT09 training and test data use plain ASCII quotes ("). Statistical MT systems trained on such data produce of course also ASCII quotes. For the purpose of a fair comparison with those systems, we have created a simple block `Ascii_quotes` that converts correct Czech directional quotes to incorrect ASCII ones. We were surprised how a large "improvement" can be achieved with this block on our test data –

0.0085 BLEU (0.1757 NIST). This fact only confirms that neither BLEU nor NIST can be used as the ultimate measure for comparing two MT systems of different types.

5. Hidden Markov Tree Models

5.1. Motivation

Most errors are caused by the transfer of lemmas and formemes
In the manual annotation of translation errors we have discovered that more than half of all errors are caused by the transfer phase and 92% of these errors are wrong lemmas and wrong formemes. The choice of correct lemma and formeme is of course a very difficult task and the quality of translation depends heavily on the quality of the dictionaries used. However, even with an ideal dictionary many errors will occur if we just select the most probable variant for each node without considering the context.

Two meanings of the word *speaker*
For example, word *speaker* with the sense *loudspeaker* should be translated as *reproduktor* and according to the lemma dictionary used in our scenario the translation probability is $\mathrm{P}(\textit{reproduktor}\,|\,\textit{speaker}) = 0.45$. When the sense is *spokesperson*, the correct translation is *mluvčí* and $\mathrm{P}(\textit{mluvčí}\,|\,\textit{speaker}) = 0.26$. Perhaps, there were more texts about loudspeakers than texts about spokespersons in the CzEng parallel corpus upon which the dictionary is based. The baseline system translates every word *speaker* as *reproduktor*, so we encounter errors in phrases like *speaker of the Ministry of Transport*.

Linear context and tree context
In phrase-based MT, the context used to select the best translation of a word is linear – basically, the context is a phrase, i.e. a string of surrounding words. There are some experiments with "phrases with gaps" (Simard et al., 2005), but in most systems a phrase is defined as a contiguous string of words (not necessarily forming a phrase in a linguistic sense).

We believe that it is more appropriate to use a local tree context, i.e. the children and the parent of a given node. Not only that it is appropriate according to linguistic intuition, but it should help us to face the data sparseness.

For illustration, consider the before-mentioned example with the phrase *speaker of the Ministry of Transport*. Human translators recognize from semantics that the *speaker* is a human being (not a loudspeaker) and translate it as *mluvčí*. Phrase-based MT systems can learn the whole phrase or possibly just the phrase *speaker of the Ministry*, but they must also learn phrases like *speaker of the Chinese Ministry*, *speaker for the Foreign*

Ministry, *speaker for the Indian External Affairs Ministry* etc. in order to translate them correctly.[8]

When using the local tree context, we can for example learn that *speaker* should be translated as *mluvčí* if it has a child node with the lemma ministry. This way we cover all the before-mentioned phrases including the unseen ones. Another knowledge learned from a parallel dependency treebank may be that *speaker* should be translated as *mluvčí* if its parent node has the lemma name (e.g. in phrases *speaker's name*, *name of the next speaker*) or that *speaker* should be translated as *reproduktor* if its parent node has the lemma buy (e.g. in a phrase *buy an expensive speaker*).

How to learn, represent and use tree context?

The obvious question is how can we learn, represent and use such knowledge. The preceding paragraph formulates the knowledge in a form of rules. Although this approach could be used in MT (rules can be automatically learned from the treebank), it is difficult to combine it with probabilistic methods. We have decided to represent the knowledge in a form of a model that describes the probability of a node given its parent node. More precisely, we model the probability of a lemma and formeme of the dependent node given a lemma and formeme of the governing node.

The model can be learned from a treebank using maximum likelihood estimate, but similarly to traditional (linear) language models it is necessary to smooth the probabilities and there are many possible ways how to perform the smoothing.

Tree context: bilingual or target-language?

The probabilistic model introduced in the previous paragraph is a monolingual tree model and can be learned from a target-language treebank (Czech in our case). With the availability of parallel treebanks we can develop also "bilingual tree models". An example of bilingual tree model is the valency formeme translation dictionary. It specifies the probability of formeme of the target-side node given formeme of the source node and lemma of the source node's parent.

Ideally, we would like to use more complex bilingual tree model that defines also target-side lemmas and that is conditioned also by other attributes (lemma of the source node, lemmas of its children etc.). This complex model would supersede both

[8]The example if oversimplified. First, in phrase-based MT systems, it is the target-language model that should cover such long phrases, so it would be more accurate to present Czech translations of the phrases. Second, we suppose that the hypothetical phrase-based system is trained on the same parallel corpus as our dictionary, so $P(reproduktor|speaker) > P(mluvčí|speaker)$ and similarly for backward probabilities $P(speaker|reproduktor) > P(speaker|mluvčí)$. Otherwise, there would be no need for the language model to cover the phrases, if the translation model itself would choose the correct translation. Third, since the phrases learned by phrase-based MT systems are usually not constrained to linguistically adequate constituency phrases, it is possible that the system will learn that *speaker of the* should be translated as *mluvčí*. However, there are plenty of more relevant examples of long-distance dependencies that are not covered even by 6-gram or 7-gram language models.

formeme and lemma dictionaries as well as the target-language tree model. However, we do not have enough parallel data to reliably train such a model. Since the amount of monolingual training data is much larger, we try to exploit it as much as possible.

First attempts at using tree context
In the baseline translation of lemmas and formemes, the only usage of tree context was in the valency formeme translation dictionary. Moreover, lemmas and formemes were translated almost independently – there was only a rule to check for compatibility of a lemma with a formeme, but no probabilistic model describing their joint or conditional probability. In other words, the target-language tree model was not used in the baseline implementation.

One of the first attempts at exploiting the target-language tree model performed a top-down depth-first traversal through the t-tree translated by the baseline system. Its main idea was to choose the best lemma and formeme according to a loglinear combination of three models: translation probability of lemma, translation probability of formeme and target-language tree model created by Václav Novák. The main difference from HMTM and the tree-modified Viterbi algorithm presented in this paper is that the top-down traversal allows only local optimization based on the parent node (but no children nodes), whereas the tree-modified Viterbi algorithm searches for the global maximum.

Why do we need Hidden Markov Tree Models?
The apparent weak point of the before-mentioned top-down traversal occurs when the correct lemma or formeme can be determined only from the children rather than from the parent (e.g. *He is a speaker of the ministry* versus *It is an expensive speaker*). Of course, if we use a similar algorithm with bottom-up traversal, these cases will be handled correctly, but errors will be introduced in the opposite cases – when the correct lemma or formeme can be determined only from the parent, but not from children (e.g. *according to the speaker* versus *buy a speaker*).

Not only that both the types of cases (parent/children are important for translation) are frequent, but sometimes we need to know the parent as well as the children to choose the correct translation. The child-parent dependencies are chained in the tree, so we need to find the combination of lemmas and formemes that results in the maximal global probability of the whole tree. Hidden Markov Tree Models provide a theoretical background for the tree-modified Viterbi algorithm, which can efficiently find the global maximum.

5.2. Description of HMTM

Related work
Hidden Markov Models (HMM, see Chapter 9 in Manning and Schütze (1999))[9] belong to the most successful techniques in Computational Linguistics. There are many modifications of HMM: arc-emission versus state-emission, epsilon-emission, HMM with Gaussian distribution of emission function etc. Hierarchical Hidden Markov Models, which are used for Information Extraction (Skounakis et al., 2003), make use of tree structures, but they still primarily work with linearly organized observations/states.

Hidden Markov Tree Models (HMTM) were introduced by Crouse et al. (1998), and used in applications such as image segmentation, signal classification, denoising and image document categorization. More information about HMTM can be found in Diligenti et al. (2003) and in Durand et al. (2004). The latter article contains also a detailed explanation of the tree-modified Viterbi algorithm. Parts of this Section are based on Žabokrtský and Popel (2009), where HMTM are introduced for dependency-based MT, and on Popel (2009).

Formal definition
Suppose that

- $V = \{1, \ldots, |V|\}$ is a set of tree nodes, $r \in V$ is the root node and $\rho : V \setminus \{r\} \rightarrow V$ is a function determining the parent node of each non-root node.
- $\mathbf{X} = (X_1, \ldots, X_{|V|})$ is a sequence of random variables taking values from a state space S. Random variable X_v is understood as a *hidden state* of the node v and $P(X_v|X_{\rho(v)})$ is called *transition probability*.
- $\mathbf{Y} = (Y_1, \ldots, Y_{|V|})$ is a sequence of *observable symbols* taking values from an alphabet K. $P(Y_v|X_v)$ is called *emission probability*.

We further introduce the following notation:

- $subtree : V \rightarrow 2^V$ is a function mapping a node v to a set of all nodes of the subtree rooted in v, i.e. $subtree(v) = \{w \in V : \exists w = z_1, \ldots, z_n = v, \forall i \in \{1 \ldots n-1\} \quad \rho(z_i) = z_{i+1}\}$.
- $\mathbf{X}(v)$ is a sequence of hidden states of the subtree rooted in v, i.e. $\mathbf{X}(v) = \{X_w : w \in subtree(v)\}$. Hence $\mathbf{X} = \mathbf{X}(r) = \{X_r, \mathbf{X}(w) : \rho(w) = r\}$.
- Analogously, $\mathbf{Y}(v)$ is a sequence of symbols of the subtree rooted in v.

Similarly to stationary first-order state-emitting HMM, we formulate three independence assumptions for HMTM:

[9] To avoid any terminological confusion, we should note that by HMM we mean only Hidden Markov *Chain* Models.

1. **stationary property** (analogy to time invariance property of HMM)
   $\forall \nu, w \in V \setminus \{r\} : P(X_\nu|X_{\rho(\nu)}) = P(X_w|X_{\rho(w)})$ &
   $\forall \nu, w \in V : P(Y_\nu|X_\nu) = P(Y_w|X_w)$,
   i.e. transition and emission probabilities are independent of nodes.
2. **tree-Markov property** (analogy to limited horizon property of HMM)
   $\forall \nu \in V \setminus \{r\}, \forall w \in V \setminus subtree(\nu) : P(\mathbf{X}(\nu)|X_{\rho(\nu)}, X_w) = P(\mathbf{X}(\nu)|X_{\rho(\nu)})$,
   i.e. given $X_{\rho(\nu)}$, all hidden states of the subtree rooted in ν are conditionally independent of any other nodes.[10]
3. **state-emission property**
   $\forall \nu, w \in V : P(Y_\nu|X_\nu, X_w, Y_w) = P(Y_\nu|X_\nu)$,
   i.e. given X_ν, Y_ν is conditionally independent of any other nodes.

Let $\nu_1, \ldots, \nu_n$ be children of the root r, then using the tree-Markov property and mathematical induction we get:

$$\begin{aligned} P(\mathbf{X}) &= P(X_r, \mathbf{X}(\nu_1), \ldots, \mathbf{X}(\nu_n)) \\ &= P(X_r)P(\mathbf{X}(\nu_1), \ldots, \mathbf{X}(\nu_n)|X_r) \\ &= P(X_r)P(\mathbf{X}(\nu_1)|X_r)P(\mathbf{X}(\nu_2), \ldots, \mathbf{X}(\nu_n)|X_r, \mathbf{X}(\nu_1)) \\ &= P(X_r)P(\mathbf{X}(\nu_1)|X_r)P(\mathbf{X}(\nu_2), \ldots, \mathbf{X}(\nu_n)|X_r) \\ &= P(X_r)P(\mathbf{X}(\nu_1)|X_r) \ldots P(\mathbf{X}(\nu_n)|X_r) \\ &= P(X_r) \prod_{\nu \in V \setminus \{r\}} P(X_\nu|X_{\rho(\nu)}) \end{aligned} \tag{1}$$

Using the state-emission property and mathematical induction we get:

$$\begin{aligned} P(\mathbf{Y}|\mathbf{X}) &= P(Y_r|\mathbf{X})P(\mathbf{Y}(\nu_1), \ldots, \mathbf{Y}(\nu_n)|\mathbf{X}(\nu_1), \ldots, \mathbf{X}(\nu_n), X_r, Y_r) \\ &= P(Y_r|X_r)P(\mathbf{Y}(\nu_1), \ldots, \mathbf{Y}(\nu_n)|\mathbf{X}(\nu_1), \ldots, \mathbf{X}(\nu_n)) \\ &= \prod_{\nu \in V} P(Y_\nu|X_\nu) \end{aligned} \tag{2}$$

From Equations 1 and 2 we can deduce the following factorization formula:

$$P(\mathbf{Y}, \mathbf{X}) = P(Y_r|X_r)P(X_r) \cdot \prod_{\nu \in V \setminus \{r\}} P(Y_\nu|X_\nu)P(X_\nu|X_{\rho(\nu)}) \tag{3}$$

[10] Our formulation of the tree-Markov property differs from the one used in Diligenti et al. (2003), which could be rewritten as
$\forall \nu, w, z \in V, \rho(w) = \rho(z) = \nu \implies P(\mathbf{X}(w)|X(\nu), \mathbf{X}(z)) = P(\mathbf{X}(w)|X(\nu))$,
i.e. given $X_{\rho(w)}$, the subtree of w is conditionally independent of its sibling subtrees.
Such assumption is too weak to be used in the last two lines of Equation 1, where we need $P(X_\nu|X_{\rho(\nu)}, X_{\rho(\rho(\nu))}) = P(X_\nu|X_{\rho(\nu)})$.
On the other hand, the formulation used in Žabokrtský and Popel (2009) is unnecessarily strong:
$\forall \nu \in V \setminus \{r\}, \forall w \in V : P(X_\nu|X_{\rho(\nu)}, X_w) = P(X_\nu|X_{\rho(\nu)})$,
i.e. given $X_{\rho(\nu)}$, X_ν is conditionally independent of any other nodes.

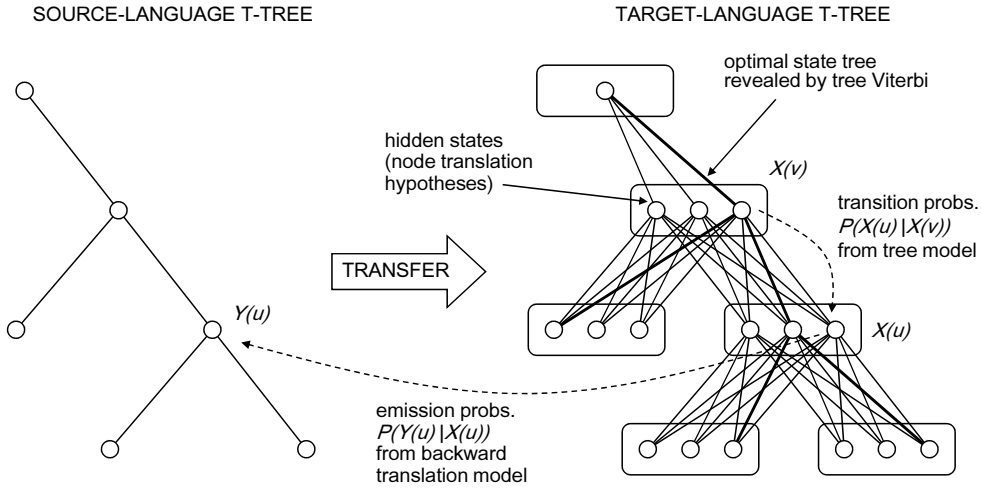


Figure 1. Scheme of the tectogrammatical transfer as a task for HMTM.

We see that HMTM (analogously to HMM, again) is defined by the following parameters:[11]

- $P(X_v|X_{\rho(v)})$ – transition probabilities between the hidden states of two tree-adjacent nodes,[12]
- $P(Y_v|X_v)$ – emission probabilities.

5.3. Application of HMTM in MT

How to estimate emission and translation probabilities?
When using HMTM in MT, labels of the source-language nodes can be interpreted as observable symbols and labels of the target-language nodes can be interpreted as hidden states (see Figure 1). In the case of TectoMT transfer, a label of a node is a pair of lemma and formeme. Therefore, the hidden states space (S) is the Cartesian product of lemmas and formemes possible for the target language and the alphabet of observable symbols (K) is the Cartesian product of lemmas and formemes possible for the source language.

HMTM emission probabilities can be estimated from the "backward" (source given target) node-to-node translation model. This node-to-node translation model can be

[11] As follows from the stationary property, the parameters are independent on the node v.

[12] The need for parametrizing also $P(X_r)$ (prior probabilities of hidden states in the root node) can be avoided by adding an artificial root whose state is fixed.

further estimated by factorization to the lemma translation dictionary and formeme translation dictionary.

HMTM transition probabilities can be estimated from the target-language tree model.

The decomposition into *translation model* and *language model* proved to be extremely useful in Statistical Machine Translation since Brown et al. (1993). It allows to compensate for the lack of parallel resources by the relative abundance of monolingual resources.

Limitations of HMTM

There are several limitations implied by the definition of HMTM, which we have to consider before applying it to MT.

The first limitation is merely a technical detail. The set of hidden states and the alphabet of observable symbols are supposed to be finite. This assumption can be easily fulfilled by introducing an artificial symbol/state for unknown tokens. However, in practice we are able to consider only a limited number of possible hidden states for each node, so the trick with an artificial symbol is not actually needed.

More serious limitations are induced by the three independence assumptions:

- **stationary property**
  We assume that the position of a node in a tree cannot influence its translation and emission probabilities. For example, this property would be violated if some words should be translated differently when being children of the main clause verb (i.e. grandchildren of the technical root).[13] According to our observations, such a dependence on the level of a node (i.e. distance from the root) is not a substantial issue.
  Another violation of the stationary property can be a dependency on word order. For example, some words should be translated differently when being at the beginning of the sentence.[13] These cases are also not a substantial problem.[14]
- **tree-Markov property**
  This assumption concerns only the target-language tree model. The conditional dependency (in the probabilistic sense) of a node on its parent corresponds well to the intuition behind dependency relations (in the linguistic sense) in dependency trees. However, there are special linguistic phenomena that violate this assumption. These phenomena are addressed in the manual for English tec-

[13] …and this difference could be determined neither from the source node nor from the target-side parent node.

[14] PDT-style tectogrammatical nodes have an attribute `deepord`, which specifies the so-called *deep word order* for the purpose of communicative dynamism. TectoMT tectogrammatical trees use this attribute for surface word order. Nevertheless, if there were a reason, the attribute could be incorporated to the source node's label to circumvent the violation of the stationary property.

togrammatical annotation (Cinková et al., 2006) in Sections: Non-dependency edges, Dual dependency and Ambiguous dependency.
Predicative complements have the so-called dual dependency – on a verb and on a semantic noun, but only the former is represented by a tree edge.[15] In the following examples[16] we mark the predicative complement with an underline; its second dependency is always the subject (*He*). *He spoke of him as of his father. He left whistling. He lives alone.*
Although not considered a dual dependency, copula constructions also violate the assumption. For example, in sentences *He is a speaker.* and *It is a speaker.* we can disambiguate the sense of the object (*speaker*) based on the subject (*He* or *It*), but these nodes are siblings, so that the probabilistic dependency cannot be directly used in HMTM.
A possible solution to circumvent these violations and hopefully improve the translation quality is to incorporate the secondary dependencies into the labels of source nodes to be handled by the translation model.

- **state-emission property**
  This property can be weakened to "arc-emission property":
  given X_ν and $X_{\rho(\nu)}$, Y_ν is conditionally independent of any other nodes, i.e.
  $\forall \nu, w \in V : P(Y_\nu|X_\nu, X_{\rho(\nu)}, X_w, Y_w) = P(Y_\nu|X_\nu, X_{\rho(\nu)})$
  A factorization formula, analogical to Equation 3, can be then proved:

$$P(\mathbf{Y}, \mathbf{X}) = P(Y_r|X_r)P(X_r) \cdot \prod_{\nu \in V \setminus \{r\}} P(Y_\nu|X_\nu, X_{\rho(\nu)})P(X_\nu|X_{\rho(\nu)}) \quad (4)$$

  With this generalization we can condition emission probabilities (i.e. translation model) on the parent node. Another (actually equivalent) method how to use a richer translation model, without the need of weakening the state-emission property, is to incorporate the needed attributes to the labels of target-side nodes.

The most limiting assumption from the MT viewpoint was not expressed explicitly yet:

- **isomorphism presumption**
  The source-language tree and the target-language tree are required to be isomorphic. In other words, only node labeling can be changed in the HMTM transfer step. This assumption concerning the tree isomorphism is problematic. As we have shown in Section 3, there are cases when it is not possible to translate a sentence correctly without violating the isomorphism presumption. On the other hand, only 8% of all translation errors in our annotation experiment were caused

[15] The latter dependency relation is indicated by the attribute `compl.rf`.

[16] We present English examples, but since the violations concern the target-language tree model, it would be more accurate to present Czech equivalents.

by such cases. Possible solutions to the problem are discussed in Popel (2009, p. 65).

5.4. Tree-modified Viterbi algorithm

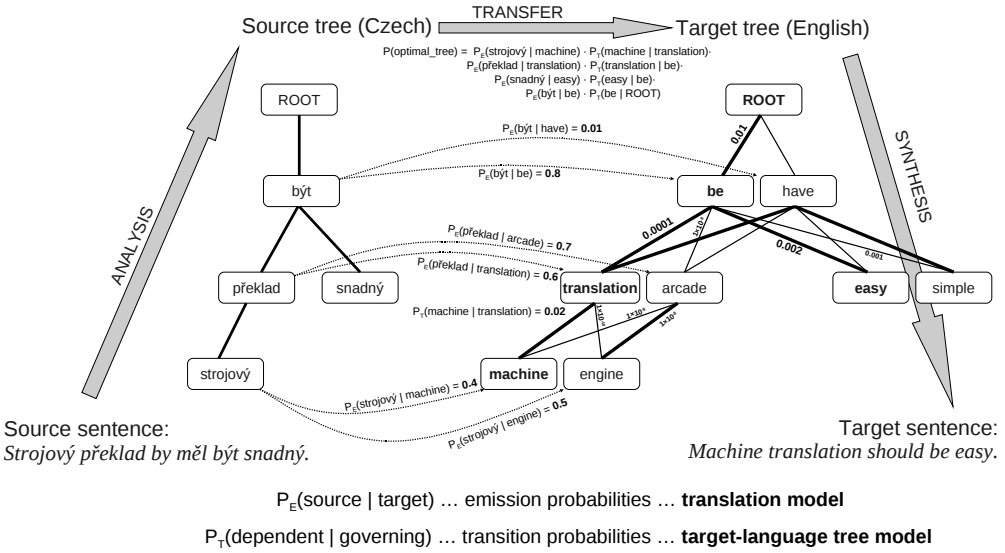


Figure 2. A simplified example of the tectogrammatical transfer as a task for HMTM. The actual translation direction is English-to-Czech, but for better illustration of the target-side t-tree, we display the Czech-to-English direction in the figure.

Naturally the question arises how to restore the most probable hidden tree labeling $\widehat{\mathbf{X}}$ given the observed tree labeling $\mathbf{Y}$ (and given the tree topology, of course). Using the factorization formula from Equation 3, we can write:

$$\begin{aligned}\widehat{\mathbf{X}} &= \arg\max_{\mathbf{X}} \mathrm{P}(\mathbf{X}|\mathbf{Y}) \\ &= \arg\max_{\mathbf{X}} \mathrm{P}(\mathbf{X},\mathbf{Y}) \\ &= \arg\max_{\mathbf{X}} \mathrm{P}(\mathrm{Y_r}|\mathrm{X_r})\mathrm{P}(\mathrm{X_r}) \cdot \prod_{\nu \in V\setminus\{\mathrm{r}\}} \mathrm{P}(\mathrm{Y_\nu}|\mathrm{X_\nu})\mathrm{P}(\mathrm{X_\nu}|\mathrm{X_{\rho(\nu)}})\end{aligned} \tag{5}$$

Similarly to the classical Viterbi algorithm, we can use dynamic programming to achieve an efficient implementation – $\mathcal{O}(|V| \cdot K^2)$ for $|V|$ nodes and K states considered for every node.

However, we cannot start at the root node and perform top-down traversal, which would be the most straightforward analogy to the classical Viterbi algorithm. Instead, the tree-modified Viterbi algorithm starts at leaf nodes and continues upwards, storing in each node for each state and each its child the optimal downward pointer to the hidden state of the child. When the root is reached, the optimal state tree is retrieved by downward recursion along the pointers from the optimal root state. Downward pointers are marked by bold edges in Figure 2.

In practice, HMTM serves us as an inspiration, though for pragmatic reasons the implementation differs in some aspects from the theory. Apart from usual practices like computing probabilities in logarithmic space and smoothing transition probabilities, we use a factorization of the translation model into two channels: lemmas and formemes. Moreover, we use a forward translation model (target given source) in addition to the backward translation model (source given target), because it proved to have a positive effect on the translation quality. The emission probability is computed as a weighted average of the models.

6. Conclusions

We have implemented several improvements of English-Czech translation system TectoMT. In order to do so, we annotated 250 sentences produced by the baseline system and identified the most prominent errors and their sources. According to the error analysis, the assumption of isomorphism between the source and target tectogrammatical trees causes only 8% of errors. This facilitates the utilization of Hidden Tree Markov Model based transfer phase, which proved to be one of the most helpful modifications we have done.

We have achieved an improvement over the baseline 0.0659 BLEU (3.9735 NIST). Our new version of TectoMT reaches 0.0981 BLEU (4.7157 NIST). Although these results are still lower than those of the state-of-the-art English-Czech MT systems, our system is rapidly evolving and we see a great potential for further improvements.

7. Acknowledgement

The work on this project was supported by the grants MSM0021620838, MŠMT ČR LC536, and 1ET101120503.

Bibliography

Bojar, Ondřej, Silvie Cinková, and Jan Ptáček. Towards English-to-Czech MT via Tectogrammatical Layer. *Prague Bulletin of Mathematical Linguistics*, 90, 2008a. ISSN 0032-6585.

Bojar, Ondřej, 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 2008b. ELRA.

Bojar, Ondřej, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 125–129, Athens, Greece, March 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W09/W09-0x22`.

Brown, Peter E., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 1993. URL `http://acl.ldc.upenn.edu/J/J93/J93-2003.pdf`.

Cinková, Silvie, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Annotation of English on the tectogrammatical level. Technical Report 35, ÚFAL MFF UK, 2006.

Crouse, Matthew, Robert Nowak, and Richard Baraniuk. Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. *IEEE Transactions on Signal Processing*, 46(4):886–902, 1998.

Cuřín, Jan, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. Prague Czech-English Dependency Treebank, Version 1.0. Linguistics Data Consortium, Catalog No.: LDC2004T25, 2004.

Diligenti, Michelangelo, Paolo Frasconi, and Marco Gori. Hidden tree Markov models for document image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25: 2003, 2003.

Durand, Jean-Baptiste, Paulo Goncalvès, and Yann Guédon. Computational methods for hidden Markov tree models - An application to wavelet trees. *IEEE Transactions on Signal Processing*, 52(9):2551–2560, 2004.

Hajič, Jan. *Disambiguation of Rich Inflection – Computational Morphology of Czech*. Charles University – The Karolinum Press, Prague, 2004.

Hopkins, Mark and Jonas Kuhn. Machine Translation as Tree Labeling. In *Proceedings of SSST, NAACL-HLT*, pages 41–48, 2007.

Koehn, Philipp and Christof Monz. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121, 2006.

Manning, Christopher D. and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing (HTL/EMNLP)*, pages 523–530, Vancouver, BC, Canada, 2005.

Popel, Martin. Ways to Improve the Quality of English-Czech Machine Translation. Master's thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2009.

Rouš, Jan. Probabilistic translation dictionary. Master's thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2009.

Simard, Michel, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. Translating with non-contiguous phrases. In *Proceedings of HLT-EMNLP*, pages 755–762, October 2005.

Skounakis, Marios, Mark Craven, and Soumya Ray. Hierarchical Hidden Markov Models for Information Extraction. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 427–433. Morgan Kaufmann, 2003.

Spoustová, Drahomíra, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha, 2007.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error Analysis of Machine Translation Output. In *Proceedings of the Fifth International Language Resources and Evaluation (LREC'06)*, pages 697–702, Genoa, Italy, May 2006.

Žabokrtský, Zdeněk and Martin Popel. Hidden Markov Tree Model in Dependency-based Machine Translation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, August 2009.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the 3rd Workshop on Statistical Machine Translation, ACL*, 2008.

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 135–147

Evaluation of Machine Translation Metrics for Czech as the Target Language

Kamil Kos, Ondřej Bojar

Abstract

In the present work we study semi-automatic evaluation techniques of machine translation (MT) systems. These techniques are based on a comparison of the MT system's output to human translations of the same text. Various metrics were proposed in the recent years, ranging from metrics using only a unigram comparison to metrics that try to take advantage of additional syntactic or semantic information. The main goal of this article is to compare these metrics with respect to their correlation with human judgments for Czech as the target language and to propose the best ones that can be used for an evaluation of MT systems translating into Czech language.

1. Introduction

In recent years a lot of research has been devoted to the field of MT evaluation. Since 2002, almost every year new MT metrics emerged that tried to establish themselves as the MT evaluation standard.

So far, the BLEU metric is considered as the golden standard in various competitions and workshops. However, some researchers have noted that BLEU is not very reliable in scoring translations on the sentence level. This can be a significant problem because MT systems usually translate source text sentence by sentence. Moreover, it is easier to collect human judgments on the sentence level because people can judge the quality of translations on the sentence level more easily than for the whole text.

In this article we examine MT metrics with respect to their correlation with human judgments on the level of the sentence and the translation system as a whole. We restrict our experiments only on Czech as target language because results for English are already available in Callison-Burch et al. (2008) and Callison-Burch et al. (2007). Because Czech belongs to a typologically different group of languages, namely the

Please cite this article as: Kamil Kos, Ondřej Bojar, Evaluation of Machine Translation Metrics for Czech as the Target Language. The Prague Bulletin of Mathematical Linguistics No. 92, 2009, 135–147.

Slavic ones with rich inflection, there can be some differences in the correlation. Some of the metrics can be more suitable for English and some of them more suitable for Czech, e.g. because of the fixed word order in English and relatively free word order in Czech.

2. Metrics

We compared the most common metrics that are used in MT systems evaluation. We used our own implementation of the metrics to compute the ratings. This was especially necessary for metrics that take advantage of syntactic or semantic information because original evaluation tools are available mostly only for English or other widespread languages like French or Spanish.

The following metrics were evaluated:

- **F-measure** is defined as the harmonic mean of *precision (p)* and *recall (r)*: $\frac{p+r}{2*p*r}$ where precision is the number of words that co-occur in the candidate and the reference sentence divided by the size of the candidate sentence, and recall is the number of words that co-occur in the candidate and the reference sentence divided by the size of the reference sentence.
- **BLEU** (Papineni et al., 2002) is based on the geometric mean of n-gram precision ($n = 1 \ldots 4$). Candidate translations that are shorter than human references are penalized by the brevity penalty which is a single value over the whole test set.
- **NIST** (Doddington, 2002) also uses n-gram precision ($n = 1 \ldots 5$), differing from BLEU in that an arithmetic mean is used, weights are used to emphasize informative word sequences and the formula for brevity penalty is different.
- **WER** (Su and Wu, 1992) is defined as the minimum number of edit operations required to transform one sentence into another normalized by the length of the reference translation

$$WER(s_i, r_i) = \frac{\min\left(I(s_i, r_i) + D(s_i, r_i) + S(s_i, r_i)\right)}{|r_i|}$$

  where $I(s_i, r_i)$, $D(s_i, r_i)$ and $S(s_i, r_i)$ are the number of insertions, deletions and substitutions, respectively, and $|r_i|$ is the length of the reference. The numerator of the equation above is also known as the Levenshtein distance.
- **TER** (Snover et al., 2006) is also based on the number of operations needed to transform the candidate sentence into the reference sentence. However, it allows one additional operation: the *block shift*. Hence, possible operations include insertion, deletion, and substitution of single words as well as shifts of word sequences.
- **PER** (Tillmann et al., 1997) is similar to WER except that word order is not taken into account. Both sentences are treated as bags of words and the set difference is judged.

- **GTM** (Turian et al., 2003) is inspired by the plain F-measure trying to eliminate (one of) its major drawbacks. Since F-measure is based only on unigram matching, two sentences containing the same words always get the same F-measure rating regardless of the correct order of the words in the sentence. GTM rewards contiguous sequences of correctly translated words. The reward is controlled by parameter e. For $e = 1$ the GTM score is the same as the plain F-measure. For $0 < e < 1$ contiguous sequences of words are rewarded and for $e > 1$ they are penalized.
- **Meteor** (Banerjee and Lavie, 2005) incrementally constructs an alignment between the candidate and the reference sentence using several modules that define which words can be matched. The modules are *exact*, *porter stem* and *WordNet (WN) synonymy*. *Exact* module matches two words if they have the same surface representation (e.g. *dog* matches *dog* but not *dogs*). *Porter stem* module matches two words if they have the same stem according to Porter stemmer (Porter, 2001) (e.g. *dogs* matches *dog*) and *WN synonymy* module matches two words if they are synonyms. Our modification of the metric replaces the *porter stem* module with *lemma* module which matches two words, if they have the same lemma. The *WN synonymy* module uses the Czech WordNet (Pala and Smrž, 2004). The alignment is then used to compute precision and recall, similarly to F-measure, only that the weight of precision is bigger than the weight of recall. Moreover, penalty is used to penalize translations with words in wrong order.
  In Lavie and Agarwal (2007), the authors optimized the parameters that are used by Meteor. We use the parameters that were obtained for English because they did not consider Czech. The new parameters put more weight on recall than before and use different coefficients in the penalty formula. We denote the original version of Meteor as *orig* and the new version without any attributes.
- **Semantic POS Overlapping** (SemPOS) metric is inspired by a set of metrics using various linguistic features on syntactic and semantic level introduced by Giménez and Márquez (2007). One of their best performing metrics was *semantic role overlapping*. Since we did not find a tool that would assign semantic roles as defined in Giménez and Márquez (2007) to words in a Czech sentence, we decided to use a slightly different metric. The *TectoMT* framework (Žabokrtský et al., 2008) can assign a semantic part of speech (semantic POS) to words. We compute overlapping for this linguistic feature as defined in Giménez and Márquez (2007). Moreover, we do not use the surface representation of the words but their `t-lemma` obtained from the *TectoMT* framework for the computation of the overlapping. As an approximation, we can say that our application of SemPOS evaluates the lexical choice of autosemantic words, taking the (semantic) part of speech into account.

Judgments per sentence	1	2	3	4	5	6	7	Total Sents.	Total Judgs.
Articles: # of sents.	119	24	8	3	5	3	3	165	267
Editorials: # of sents.	109	26	9	8	1	3	0	156	243

Table 1. Number of sentences with 1 to 7 human ratings in the test sets.

3. Test Data

The test data and human judgments were taken from the data collected at the Third Workshop on Statistical Machine Translation (Callison-Burch et al., 2007). We have chosen only systems and human judgments which had Czech as the target language. We used the human rankings of whole sentences. The judgments about syntactic constituents were not taken into account.

The output of the following systems was considered:

- BOJAR - Charles University, Bojar (Bojar and Hajič, 2008),
- TMT - Charles University, TectoMT (Žabokrtský et al., 2008),
- UEDIN - University of Edinburgh (Koehn et al., 2008),
- PCT - PC Translator (a commercial MT provider from the Czech Republic).

The test data consisted of two test sets. The first one contained a total of 90 articles which were selected from a variety of Czech, English, French, German, Hungarian and Spanish news sites. The other test set was drawn from Czech-English news editorials. The Articles test set contained 2050 sentences and the Editorials test set contained 2028 sentences. The reference translations contained only one human translation for each sentence.

The human judgments contained 243 system scores of 156 unique sentences for the Editorials test set and 267 system scores of 165 unique sentences for the Articles test set with up to 7 judgments of a single sentence. Table 1 gives the details of judgment distribution. The human judgments contained scores of the translation quality on the scale 1 to 5, one being the best. It was possible that several translations obtained the same score. The scores for the translations were only on the sentence level. We considered human scores of the same sentence as independent of each other and included all of them in the ratings.

4. Correlation with Human Judgments

To measure the correlation of the metric ratings with the human judgments we used the Pearson correlation coefficient on ranks. This coefficient captures the extent to which two different rankings correlate with each other. We used the following equation:

$$\rho = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}\sqrt{n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2}}$$

Human score	Metric score	Human rank	Metric rank
1	0.62	1.5	1
3	0.54	3	3
1	0.54	1.5	3
5	0.54	4	3

Table 2. Conversion of scores to rankings.

In the formula, n denotes the number of evaluated systems and x_i, y_i are the positions of the i^{th} system in the human and metric rank. The possible values of ρ range between 1 (all systems are ranked in the same order) and -1 (systems are ranked in the reverse order). Thus, an evaluation metric with a higher value of ρ reflects the human judgments better than a metric with a lower ρ.

4.1. Sentence-Level Correlation

To measure the sentence-level correlation we transformed the human scores to ranks for each sentence. If several systems obtained the same score, we used the average position for each of them. In the case that all systems had the same score, we did not use the human judgment. For automatic metrics, we computed the metric scores on the sentence level and converted the scores to rankings in the same manner as for human judgments. Table 2 illustrates how we created the rankings.

4.2. System-Level Correlation

Because no human judgments were available on the system level we had to synthesize them from sentence level judgments. We used the same method as in Callison-Burch et al. (2007) in order to make the results comparable. We created the system rankings based on the

- *percent of cases in which the sentences (produced by the system) were judged to be better than or equal to the translations of any other system.*

Since we had only two test sets to measure the correlation coefficients on the system level, we used bootstrapping to estimate their variance. On the system level, we obtained no ties in rankings. Then, the Pearson correlation coefficient is equivalent to the Spearman's rank correlation coefficient defined as:

$$\rho_{sp} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks for $system_i$ and n is the number of systems.

Metric	Articles	Editorials	Average
NIST	0.22±0.60 (7)	0.26±0.62 (1)	0.24
F-measure/GTM(e=1)	0.24±0.58 (1)	0.23±0.63 (4)	0.23
GTM(e=0.5)	0.24±0.58 (2)	0.23±0.63 (6)	0.23
GTM(e=2)	0.24±0.58 (3)	0.22±0.63 (10)	0.23
Meteor	0.23±0.57 (4)	0.24±0.62 (2)	0.23
GTM(e=0.1)	0.23±0.58 (5)	0.23±0.63 (5)	0.23
Meteor(orig)	0.23±0.57 (6)	0.23±0.62 (7)	0.23
PER	0.22±0.60 (8)	0.24±0.63 (3)	0.23
TER	0.21±0.60 (9)	0.23±0.62 (8)	0.22
WER	0.21±0.60 (10)	0.23±0.62 (9)	0.22
SemPOS	0.21±0.57 (11)	0.19±0.61 (11)	0.20
BLEU	0.03±0.63 (12)	0.02±0.62 (12)	0.03

Numbers in brackets indicate the relative position of the metric.

Table 3. Average sentence-level correlations for the metrics including standard deviation.

5. Results and Discussion

In the present section, we discuss various aspects of the estimated correlations to human judgments. For complete listing of results, please see Tables 7 and 8 at the end of our article.

5.1. BLEU Not Suitable for Sentence-Level Evaluation

The results of the sentence-level correlation are given in Table 3. They indicate that the correlation of the automatic metrics with human judgments is not very high (around 0.2). Perhaps more importantly, the huge variance of the correlation discards any differences between the metrics. In fact, all results lie within the error bars of the best performing metrics (NIST for the Editorials dataset and F-measure/GTM(e=1) for the Articles dataset).

The only outstanding result is the extremely low correlation for BLEU. The BLEU metric cannot predict the human judgments on the sentence level at all which makes it unsuitable for evaluation of the quality of separate sentences.

5.2. Sentence-Level Correlation Difficult for Humans

The low coefficients observed in Table 3 are, however, influenced by the quality of human judgments. The inter-human correlation coefficients are given in Table 4. They suggest that it is difficult even for human annotators to agree which sentence

	Articles	Editorials
Judgment pairs	224	156
ρ	0.56±0.48	0.56±0.50

Table 4. Number of human judgment pairs of the same sentence and the average inter-human correlations with standard deviation.

translations are good. For an illustration of two sentences see Figures 1 and 2 at the end of the paper.

The inter-human correlation coefficients were computed as follows: we took the human scores for sentences for which there were given at least two human judgments and computed the Pearson's correlation coefficient for them. If there were more than two ratings of the same sentence, we considered all possible combinations. For the Editorials test set, we obtained 156 pairs of human judgments and for the Articles test set 224 pairs.

5.3. SemPOS Best for System-Level Comparison

Table 5 presents average Pearson correlation coefficients for both test sets on the system level. We used bootstrapping to estimate the confidence intervals. We can see that the Semantic POS Overlapping metric clearly has the highest correlation, followed by the Meteor metric. The next metrics are GTM(e=0.5) and BLEU. Metrics with the lowest correlation were the distance metrics PER, WER and TER.

It is interesting that NIST, the best metric on the sentence level, finished in the second half of the chart on the system level. On the contrary, BLEU can evaluate the quality of translation much better on the system level than on the sentence level, even if it is only slightly better than the average metrics on the system level.

Note that the Semantic POS Overlapping extensively takes advantage of the automatic annotation tools. The MT output must be preprocessed first to obtain the semantic POS and `t-lemma` for the words of the translation. Hence, the performance of Semantic POS Overlapping metric can be influenced by the quality of the annotation tools.

5.4. Effects of Lemmatization

Table 6 illustrates the effects of lemmatizing both the reference and the hypothesis of the system for selected metrics. By lemmatizing, we deliberately ignore differences in word forms. The systems are therefore not judged on the basis of morphological coherence of the output.

The column "lemma" shows correlations for texts lemmatized while preserving the number of tokens. The column "t-lemma" shows correlations for linearized tec-

Metric	Articles	Editorials	Average
SemPOS	**0.81±0.18 (1)**	**0.75±0.23 (1)**	0.78
Meteor	0.43±0.18 (2)	**0.60±0.28 (2)**	0.52
Meteor(orig)	0.43±0.18 (3)	**0.52±0.32 (3)**	0.47
GTM(e=0.1)	0.24±0.34 (9)	*0.48±0.34 (4)*	0.36
GTM(e=0.5)	0.40±0.22 (5)	0.28±0.33 (5)	0.34
BLEU	0.40±0.23 (6)	0.25±0.33 (6)	0.33
F-measure/GTM(e=1)	0.41±0.21 (4)	0.21±0.31 (7)	0.31
GTM(e=2)	0.31±0.34 (7)	0.18±0.31 (9)	0.24
NIST	0.25±0.34 (8)	0.21±0.31 (8)	0.23
PER	0.01±0.38 (10)	0.16±0.32 (12)	0.09
TER	-0.17±0.41 (11)	0.18±0.32 (10)	0.00
WER	-0.17±0.41 (12)	0.18±0.32 (11)	0.00

Results covered in the error bounds of the best result are in bold. Results covering the best result in their error bounds are in italics. Numbers in brackets indicate the relative position of the metric.

Table 5. Average system-level correlations with standard deviations for the metrics computed from bootstrapped samples (N=10000).

togrammatical trees where the number of tokens has been reduced (auxiliary words are removed, the reflexive particle becomes part of the verb t-lemma).

The results are not very pronounced, the error bars always cover the differences. In general, lemmatization tends to improve the correlation but for some metrics and some datasets, the correlation can significantly drop.

As can be seen in Table 8 at the end of the paper, SemPOS remains the best performing metric for the system-level comparison. For the sentence-level comparison, lemmatization puts the very simple PER metric higher on the scale, see Table 7.

5.5. Comparison with English

If we compare our results with the correlation coefficients on the system level that were published in Callison-Burch et al. (2008) and Callison-Burch et al. (2007), we can see that the results for Czech and English as the target language are similar. Meteor and SemPOS (which is similar to Semantic Roles Overlapping (SR) metric from Callison-Burch et al., 2007) correlate the best with human judgments, while TER (mTER in Callison-Burch et al., 2007) has one of the lowest correlation coefficients. However, almost all metrics, except for SemPOS, show correlation coefficients of only 0.3 to 0.4 for Czech compared to 0.6 to 0.8 for English. We have documented that the distance metrics PER, WER and TER are completely unsuitable for system-level evaluation for Czech. We explain this by the morphological richness of Czech—various

Metric	Dataset	word form	lemma	t-lemma
BLEU	Articles	0.40±0.23	↘0.36±0.30	↘0.14±0.46
	Editorials	0.25±0.33	↗0.43±0.35	↗0.50±0.32
F-measure/GTM(e=1)	Articles	0.41±0.21	↗0.49±0.21	↗0.56±0.24
	Editorials	0.21±0.31	↗0.29±0.34	↗0.41±0.35
GTM(e=0.1)	Articles	0.24±0.34	↘-0.19±0.35	↘0.01±0.41
	Editorials	0.48±0.34	↘0.44±0.35	↗0.66±0.23
GTM(e=0.5)	Articles	0.40±0.22	↘0.39±0.23	↗0.47±0.26
	Editorials	0.28±0.33	↗0.48±0.33	↗0.62±0.25
GTM(e=2)	Articles	0.31±0.34	↗0.64±0.26	↗0.56±0.28
	Editorials	0.18±0.31	⇝0.18±0.32	↗0.21±0.32
NIST	Articles	0.25±0.34	↗0.50±0.32	↗0.32±0.36
	Editorials	0.21±0.31	↗0.32±0.35	↗0.33±0.35
PER	Articles	0.01±0.38	↗0.21±0.42	↘-0.09±0.35
	Editorials	0.16±0.32	↗0.20±0.33	↗0.19±0.33

Table 6. Effects of lemmatization on system-level correlation.

paraphrases with the same meaning can often differ in every word form due to e.g. a different case.

6. Future Work

More accurate results about the quality of MT metrics for Czech as the target language can be obtained if the experiments we have performed on the system level would be repeated on more data. We had only 2 test sets of 156 and 165 unique sentences. Since the synthesized system-level human judgments from sentence-level scores are possible sources of errors, it would be useful to collect human judgments for whole test sets. However, people are better at scoring shorter fragments of text and they even have problems when evaluating longer sentences. We suggest to use a task-based evaluation instead of scoring the whole text. Human judgments can be collected on the basis of a set of questions after the translations are read by the annotators, like in comprehension tests in foreign language exams.

Other metrics that emerged recently can be implemented and evaluated. This concerns especially metrics that were published in Giménez and Márquez (2007). Several of them show high correlation with human judgments for English. The TectoMT framework can provide most of the required features to compute these metrics for Czech sentences.

		Human scores	
Src	Berlusconi's lack of experience in politics doomed his first government to collapse after only six months.		
Ref	Berlusconiho nedostatečné zkušenosti v politice odsoudily jeho první vládu po pouhých šesti měsících k pádu.		
BOJAR	Berlusconiho nedostatek zkušeností v politice odsoudila jeho první vláda, Berlusconi's $\text{lack}_{nom,acc}$ $\text{experience}^{pl}_{gen}$ in politics doomed his first gov._{nom}, která padla po pouhých šest měsíců. which collapsed after only six_{nom} months.	**3**	3
TMT	Nedostatek Berlusconiho zkušenosti z politik odsoudil svou první vládu Lack of Berlusconi's $\text{experience}^{sg}_{gen}$ from politics doomed his first gov._{acc} zhroutit po pouze šesti měsících. to collapse after only six months.	4	**2**
UEDIN	Berlusconi's nedostatku zkušeností v politice odsouzena k zániku jeho první Berlusconi's lack_{dat} $\text{experience}^{pl}_{gen}$ in politics doomed to downfall his first vláda, která padla po pouhých šest měsíců. gov._{nom}, which collapsed after only six_{nom} months.	4	4
PCT	Berlusconi- ovo nedostatek zkušenosti s politikami odsoudily jeho první Berlusconi- 's lack $\text{experience}^{sg}_{gen}$ with politics_{pl} doomed_{pl} his first vládu k tomu, aby zhroutila se po jen šesti měsíců. gov. to that, so that collapsed *refl* after only six months_{gen}.	**3**	3

Figure 1. Example sentence 1 with human scores

		Human scores				
Src	The former police chief has been cooperating fully with the prosecutors investigating the case, Morvai added.					
Ref	Attila Morvai se zmínil taktéž o tom, že bývalý policejní kapitán od začátku spolupracoval se státními zástupci vykonávajícími vyšetřování.					
BOJAR	Bývalý policejní šéf byl plné spolupráci s prokurátory Morvai Former police chief was full cooperation with prosecutors Morvai vyšetřování případu, dodal. investigation case_{gen}, added.	4	4	4	4	**3**
TMT	Že se bývalý policejní šéf spolupracoval plně žalobci That *refl* former police chief cooperated fully prosecutors vyšetřováním případu, Morvai přidal. $\text{investigation}_{inst}$ case_{gen}, Morvai added.	3	3	3	**2**	4
UEDIN	Bývalý náčelník policie bylo plné spolupráci s prokurátory Former chief police was full cooperation with prosecutors vyšetřování případu, morvai přidán. investigation case_{gen}, $\text{morvai added}_{pass}$.	2	3	4	4	4
PCT	Bývalý policejní šéf spolupracoval plně se žalobce Former police chief cooperated fully with prosecutor_{nom} vyšetřování případ, Morvai přidal. investigation case_{nom}, Morvai added.	**1**	**2**	**2**	**2**	4

Figure 2. Example sentence 2 with human scores

Preprocessing	Metric	Articles	Editorials	Average
lemma	PER	**0.24±0.57 (1)**	**0.28±0.61 (2)**	0.26
t-lemma	PER	**0.21±0.56 (17)**	**0.30±0.59 (1)**	0.26
lemma	F-measure/GTM(e=1)	**0.24±0.58 (2)**	**0.24±0.60 (14)**	0.24
t-lemma	NIST	**0.24±0.56 (3)**	**0.24±0.58 (15)**	0.24
–	NIST	**0.22±0.60 (11)**	**0.26±0.62 (3)**	0.24
t-lemma	F-measure/GTM(e=1)	**0.22±0.57 (12)**	**0.26±0.59 (6)**	0.24
t-lemma	GTM(e=0.1)	**0.22±0.57 (13)**	**0.26±0.59 (7)**	0.24
t-lemma	GTM(e=0.5)	**0.22±0.57 (14)**	**0.26±0.59 (8)**	0.24
–	F-measure/GTM(e=1)	**0.24±0.58 (4)**	**0.23±0.63 (16)**	0.23
–	GTM(e=0.5)	**0.24±0.58 (5)**	**0.23±0.63 (18)**	0.23
–	GTM(e=2)	**0.24±0.58 (6)**	**0.22±0.63 (24)**	0.23
–	Meteor	**0.23±0.57 (7)**	**0.24±0.62 (12)**	0.23
–	GTM(e=0.1)	**0.23±0.58 (8)**	**0.23±0.63 (17)**	0.23
–	Meteor(orig)	**0.23±0.57 (9)**	**0.23±0.62 (19)**	0.23
lemma	GTM(e=2)	**0.23±0.59 (10)**	**0.23±0.62 (23)**	0.23
–	PER	**0.22±0.60 (15)**	**0.24±0.63 (13)**	0.23
lemma	GTM(e=0.5)	**0.22±0.59 (16)**	**0.23±0.60 (22)**	0.23
t-lemma	GTM(e=2)	**0.21±0.57 (18)**	**0.26±0.59 (9)**	0.23
lemma	TER	**0.19±0.57 (24)**	**0.26±0.61 (4)**	0.23
lemma	WER	**0.19±0.57 (25)**	**0.26±0.61 (5)**	0.23
–	TER	**0.21±0.60 (19)**	**0.23±0.62 (20)**	0.22
–	WER	**0.21±0.60 (20)**	**0.23±0.62 (21)**	0.22
lemma	GTM(e=0.1)	**0.21±0.60 (21)**	**0.22±0.59 (25)**	0.21
lemma	NIST	**0.21±0.59 (22)**	**0.22±0.61 (26)**	0.21
–	SemPOS	**0.21±0.57 (23)**	**0.19±0.61 (27)**	0.20
t-lemma	TER	**0.13±0.61 (26)**	**0.25±0.62 (10)**	0.19
t-lemma	WER	**0.13±0.61 (27)**	**0.25±0.62 (11)**	0.19
lemma	BLEU	**0.09±0.60 (28)**	**0.02±0.64 (30)**	0.06
t-lemma	BLEU	**0.02±0.58 (30)**	**0.06±0.63 (28)**	0.04
–	BLEU	**0.03±0.63 (29)**	**0.02±0.62 (29)**	0.03

Results covered in the error bounds of the best result in bold.

Table 7. Sentence-level correlations with human judgments.

Preprocessing	Metric	Articles	Editorials	Average
–	SemPOS	**0.81±0.18 (1)**	**0.75±0.23 (1)**	0.78
t-lemma	GTM(e=0.5)	0.47±0.26 (7)	**0.62±0.25 (3)**	0.54
–	Meteor	0.43±0.18 (8)	**0.60±0.28 (4)**	0.52
t-lemma	F-measure/GTM(e=1)	0.56±0.24 (3)	*0.41±0.35 (11)*	0.48
–	Meteor(orig)	0.43±0.18 (9)	**0.52±0.32 (5)**	0.47
lemma	GTM(e=0.5)	0.39±0.23 (13)	*0.48±0.33 (8)*	0.43
lemma	GTM(e=2)	**0.64±0.26 (2)**	0.18±0.32 (25)	0.41
lemma	NIST	*0.50±0.32 (5)*	0.32±0.35 (13)	0.41
lemma	BLEU	0.36±0.30 (14)	*0.43±0.35 (10)*	0.40
t-lemma	GTM(e=2)	*0.56±0.28 (4)*	0.21±0.32 (19)	0.39
lemma	F-measure/GTM(e=1)	0.49±0.21 (6)	0.29±0.34 (14)	0.39
–	GTM(e=0.1)	0.24±0.34 (18)	*0.48±0.34 (7)*	0.36
–	GTM(e=0.5)	0.40±0.22 (11)	0.28±0.33 (15)	0.34
t-lemma	GTM(e=0.1)	0.01±0.41 (21)	**0.66±0.23 (2)**	0.34
–	BLEU	0.40±0.23 (12)	0.25±0.33 (16)	0.33
t-lemma	NIST	0.32±0.36 (15)	0.33±0.35 (12)	0.33
t-lemma	BLEU	0.14±0.46 (20)	*0.50±0.32 (6)*	0.32
–	F-measure/GTM(e=1)	0.41±0.21 (10)	0.21±0.31 (17)	0.31
–	GTM(e=2)	0.31±0.34 (16)	0.18±0.31 (22)	0.24
–	NIST	0.25±0.34 (17)	0.21±0.31 (18)	0.23
lemma	PER	0.21±0.42 (19)	0.20±0.33 (20)	0.21
lemma	GTM(e=0.1)	-0.19±0.35 (30)	*0.44±0.35 (9)*	0.12
–	PER	0.01±0.38 (22)	0.16±0.32 (28)	0.09
lemma	TER	-0.01±0.36 (23)	0.18±0.32 (26)	0.08
lemma	WER	-0.01±0.36 (24)	0.17±0.32 (27)	0.08
t-lemma	PER	-0.09±0.35 (25)	0.19±0.33 (21)	0.05
–	TER	-0.17±0.41 (28)	0.18±0.32 (23)	0.00
–	WER	-0.17±0.41 (29)	0.18±0.32 (24)	0.00
t-lemma	TER	-0.16±0.32 (26)	0.12±0.33 (29)	-0.02
t-lemma	WER	-0.16±0.32 (27)	0.12±0.33 (30)	-0.02

Results covered in the error bounds of the best result in bold.
Results covering the best result in their error bounds in italics.

Table 8. System-level correlations with human judgments.

7. Conclusion

This work has examined the most common MT system evaluation metrics that are currently used. The experiments have demonstrated that the most suitable metrics for evaluation of MT systems on the system level with Czech as the target language are Semantic POS Overlapping and Meteor, followed by GTM, BLEU and NIST. These results are consistent with data that were published for systems with English as the target language even though the correlation coefficients with human judgments are lower for Czech.

The evaluation of MT quality on the sentence level proved to be unsuitable because of a relatively low correlation with human judgments for all considered metrics. Due to the variance of the correlations, none of the metrics was identified as the best one. We only found out that BLEU does not correlate with human judgments on the sentence level. However, the results were influenced by the quality of human judgments which had only a moderate inter-human correlation.

8. Acknowledgment

The work on this project was supported by the grant FP6-IST-5-034291-STP (EuroMatrix), and the grants MSM0021620838 and ME838.

Bibliography

Banerjee, S. and A. Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, pages 65–72, Ann Arbor, Michigan, June 2005.

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

Callison-Burch, Chris, 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, 2007. Association for Computational Linguistics.

Callison-Burch, Chris, 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, 2008. Association for Computational Linguistics.

Doddington, George. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Giménez, Jesús and Lluís Márquez. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, June 2007. Association for Computational Linguistics.

Koehn, Philipp, Abhishek Arun, and Hieu Hoang. Towards Better Machine Translation Quality for the German-English Language Pairs. In *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Lavie, A. and A. Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, June 2007. Association for Computational Linguistics.

Pala, Karel and Pavel Smrž. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 2004(7):79–88, 2004. URL `http://www.fit.vutbr.cz/research/view_pub.php?id=7682`.

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

Porter, Martin. The Porter Stemming Algorithm, 2001. URL `http://www.tartarus.org/martin/PorterStemmer/index.html`. Last visited on July 16, 2008.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Morristown, NJ, USA, August 2006. The Association for Machine Translation in the Americas.

Su, K. and J. Wu. A New Quantitative Quality Measure for Machine Translation Systems. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 433–439, Nantes, France, July 1992.

Tillmann, Christoph, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. Accelerated DP Based Search for Statistical Translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece, September 1997.

Turian, Joseph P., Luke Shen, and I. Dan Melamed. Evaluation of Machine Translation and its Evaluation. In *Machine Translation Summit IX*, pages 386–393. International Association for Machine Translation, September 2003.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio, June 2008. Association for Computational Linguistics.

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 149–152

REVIEWS

Figures of General Linguistics

Qixiang Cen

Beijing: World Publishing Corporation (WPC), 2008, 3+2+213pp.,
ISBN 978-7-5062-8758-6/ H 1033

Reviewed by Jun Qian, Peking University

This handsome collection (Chinese title being *pu tong yu yan xue ren wu zhi*) consists of eighteen articles devoted to eighteen linguists. It can be roughly divided into four sections, i.e. (1) Western European linguists (1-90), (2) American linguists (91-112), (3) Slavic linguists (113-172), and (4) Chinese linguists (173-213).

The section on Western European linguists introduces Ferdinand de Saussure (1857––1913), Antoine Meillet (1866–1936), Vilhem Thomsen (1842–1927), Karl Verner (1846––1896), Otto Jespersen (1860–1943), Joseph Vendryès (1875–1960), Marcel Cohen (1884––1974), and André Martinet (1908–1999). As can be seen, these eight linguists are not exactly chronologically presented, since the three Danish linguists were all born earlier than Meillet. This presentation arrangement implies Cen's view of the position that Saussure and Meillet take in the history of linguistics. The article on Saussure focuses on his *Cours de linguistique générale* (1916) and Cen concludes with the statement that since its publication the revision and development of linguistic theories by all linguists have to be based on the relevant concepts in this book (16). The article on Meillet talks about his achievements in historical-comparative linguistics, historical linguistics, and general linguistics. The three articles on Thomsen, Verner, and Jespersen introduce their respective contribution, i.e. Thomsen's deciphering of the Turkic Orkhon inscriptions, Verner's Law, and Jespersen's study of English and other subjects. The article on Vendryès focuses on his *Le Langage* (1921), which Cen believes has deeply influenced linguistics in China (58). In fact, Cen and his student (the late Professor Ye Feisheng) co-translated *Le Langage* into Chinese (published in 1992). The article on Cohen is comprehensive with some account of their interpersonal relation-

ship while the article on Martinet concentrates on his book *Eléments de linguistique générale* (1960).

The section on Slavic linguists introduces Baudouin de Courtenay (1845–1929), Lev Vladimirovich Scherba (1880–1944) Nikolai Sergeyevich Trubetzkoy (1845—1929), and Roman Jakobson (1896–1982). The article on Baudouin focuses on his contribution to phonology. The discussion is based on the Russian edition of Baudouin's writings on general linguistics (1963, 1964), although probably for most of the present-day readers Edward Stankiewicz's all-English Baudouin anthology (1972; cf. Stankiewicz 1987; Birnbaum 1998) is more accessible and helpful, which Cen does not refer to. The article on Scherba makes a fairly comprehensive presentation of his contribution to phonology, grammar, lexicography, and other fields. The article on Trubetzkoy concentrates on his contribution to phonology. In Cen's view, Trubetzkoy's contribution lies not only in his coinage of certain terms, but also in his definition and illumination of their content (146). The presentation is based on Trubetzkoy's *Grundzüge der Phonologie* (1939). At the end of this article, after a brief mention of Roman Jakobson and André Martinet's further development of Trubetzkoy, Cen writes that "structural linguistics is characterized by a conspicuous feature, i.e. over-attention to the interrelationship among linguistic elements. Some even think that the sole goal of linguistics is to study these relationships, as if all the rest are irrelevant to them. And that is a big weak point of structural linguistics." (155-156) The article on Roman Jakobson summarizes his contribution to linguistics as the theory of phonemic distinctive features, clarification of the relationship between synchronic linguistics and diachronic linguistics, the study of child language and aphasia, and the theory of language functions.

The section on American linguists introduces Zellig Harris (1909–1992) and Noam Chomsky (b. 1928). The article on Harris discusses his work, the first phase of which (pre-1951) is characterized by his structural analysis of morphology and phonology whereas the second phrase of which (post-1951) is characterized by his analysis of sentences, and in between is his representative book *Methods in Structural Linguistics* (1951). Cen believes that there are many self-contradictions in Harris's approach and "it is not strange that he is scoffed at as of 'hocus-pocus group'." (98) The article on Chomsky focuses on pre-1970 Chomsky. Cen tries to explain Chomsky's phrase structure grammar, transformational grammar, and generative grammar. His illustration is mainly based on Chomsky's *Syntactic Structures* (1957). Cen believes that Chomsky's transformational generative grammar is in essence based on Descartes's rationalism, intermingled with many mathematical-logical elements, and "no matter how skillfully he applies it, it is methodologically inadvisable." (112)

The section on Chinese linguists introduces four eminent linguists, Zhao Yuan Ren (= Yuen Ren Chao, 1892–1982), Luo Chang Pei (=Lo Ch'ang-p'ei ,1899–1958), Li Fang Gui (=Fang-Kuei Li,1902–1987), and Wang Li (1900–1986). The first three are generally regarded as pioneers and great contributors in the course of modernization of Chinese linguistics (205). The article on Zhao refers to his achievements in establishing a set of Roman letters to spell Chinese characters, his field study of Chinese

dialects, and his co-translation (with Luo Chang Pei and Li Fang Gui) of the Swedish sinologist Bernhard Karlgren's (1889–1978) *Études sur la phonologie chinoise* (1915-1926; Chinese translation *zhong guo yin yun xue yan jiu* published in 1940, 781 pages). The project was far more than merely translating, as it involved addition, deletion, and rewriting, all with Kalgren's permission. The articles on Luo and Li focus on their phonological studies and historical-comparative studies respectively. Zhao and Li lived and died in the USA, and Luo died in 1958 in mainland China. The only one of them who experienced the so-called Great Cultural Revolution that swept the mainland China from 1966 to 1976 was Wang Li, a distinguished linguist, phonologist, and grammarian. The article on Wang Li gives a general description of his research and teaching. In spite of their varied experiences these Chinese linguists are characterized by two shared features. One is their international educational experience in their formative years. Zhao (Ph.D., Harvard, 1918; cf. Zhao & Huang 1998:89) and Li ((Ph.D., Chicago, 1928) studied in the USA, Wang studied in France (Ph.D., 1932), and Luo taught from 1944 to 1948 in the USA (I found out that among Leonard Bloomfield Papers at Manuscripts and Archives, Yale University Library is one letter by Luo to Bloomfield, written on June 8, 1946.)[1]. Another is their profound knowledge of their native culture (philology, literature, history, philosophy, etc.). These two features may explain why they became distinguished linguists.

Qixiang Cen (1903–1989), author of this collection, was a professor of Chinese at Peking University in mainland China. The collection was somehow intended as a supplement to his monograph *yu yan xue shi gai yao (History of Linguistics,* 1958, revised ed. 1988). Being a France-trained linguist (1928–1933), Cen understandably includes articles on Meillet Vendryès, and Cohen, from whom he took linguistic classes. Being a Chinese linguist whose research areas included Chinese dialects and languages of other ethnic groups in China, he naturally includes articles on four important Chinese linguists. Taken as a whole, the collection makes an informative and introductory reading. One distinctive flavor of the book is Cen's accounts of his interpersonal relationships with some of the linguists in question (71–72, 189–192, 198, 206, 208), which are fairly interesting.

References

Birnbaum, Henrik. 1998. *Sketches of Slavic Scholars.* Indiana University: Slavica Publishers.

Karlgren, Bernhard. 1915–1926. *Etudes sur la phonologie chinoise.* Upsala: K. W. Appelberg; Leyde: E.-J. Brill.

Stankiewicz, Edward. 1972. *A Baudouin de Courtenay anthology: the beginnings of structural linguistics.* Translated and edited with an introduction by Edward Stankiewicz. Bloomington, Indiana: Indiana University Press.

[1] http://hdl.handle.net/10079/fa/mssa.ms.0635

Stankiewicz, Edward. 1987. Baudouin de Courtenay: Pioneer in Diachronic Linguistics. In Hans Aarsleff, Louis G. Kelly and Hans-Josef Niederehe, eds. *Papers in the History of Linguistics,* 1987, 539-549. Amsterdam: John Benjamins.

Zhao, Xinna & Peiyun Huang. 1998. *Zhao Yuan Ren nian pu (The Chronicles of Zhao Yuan Ren).* Beijing: The Commercial Press.

Note from the Editors

The Editors of PBML have been pleased very much to learn that Professor Jun Qian (English Department, Peking University), who is a frequent contributor to our Bulletin, has been awarded by the Czech Ministry of Foreign Affairs the *Jan Masaryk Bronze Medal* for promoting Czech culture in China.

Qian's introduction of the Czech culture to the Chinese people began in 1990. He has published three books and many papers closely related with the Prague School or Czech scholars, the most important one of which is *Structural-Functional Linguistics: The Prague School* (1998). This book is the first one on the Prague School in Chinese and it won a book prize and a research prize awarded by China's Ministry of Education.

To help Chinese have access to Czech scholars' writings, Qian made strenuous efforts to urge Chinese publishers to reprint these writings. At present three books by Czech scholars have been reprinted in China, with Qian's detailed introduction for each of them. The three books by Czech scholars are as follows:

Mathesius, Vilém. 2008. *A Functional Analysis of Present Day English on a General Linguistic Basis.* [Qian's introduction, 13-63] Beijing: World Publishing Corporation.

Firbas, Jan. 2007. *Functional Sentence Perspective in Written and Spoken Communication.* [Qian's introduction, 13-31] Beijing: World Publishing Corporation.

Luelsdorff, Philip A., Jarmila Panevová, and Petr Sgall (eds.). 2004. *Praguiana 1945-1990.* [Qian's introduction, 1-42] Beijing: Peking University Press.

Qian's two courses *The Prague School* and *Functional Linguistics* are devoted to the understanding of Czech scholars and he has advised many theses and dissertations on their ideas. In addition, Qian has also taken an active part in translating Czech scholars' writings into Chinese and has given talks on Czech scholars.

Qian's twenty-year effort has contributed remarkably to Chinese people's better understanding of the Czech culture and to the cultivation of the younger generation's interest in this domain.

Our most sincere congratulations!

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009 153–155

NOTE

Zdeněk Kirschner died

Petr Sgall

One of the most active and most systematic researchers of the Prague group of computational linguistics, Dr. Zdeněk Kirschner, born January 15, 1924, died on the Christmas Eve 2008, three weeks before his 85th birthday.

Zdeněk joined our research group in the spring of 1970, after his return from a longer stay in Tanzania, where he, among other things, was engaged in the education and cultural orientation of the fighters for the freedom of Mosambique; some of them later visited him in Prague as esteemed representatives of their country and members of its government.

One of the first questions Zdeněk had to answer after his arrival concerned, as it was usual in the given historical situation, was his attitude towards the Soviet led invasion of Czechoslovakia. Since he did not approve this event, he was excluded from the communist party. Even so, he was allowed to apply for a position at Charles University, where twenty years earlier he obtained his PhD. degree in English studies. He then had to decide between two options at the Faculty of Arts, one of which was to join the then established institute oriented at Ibero-American studies and the other concerned our invitation to the research group represented at that time as the Laboratory of Algebraic Linguistics. We were glad that he chose the latter possibility, and luckily we were quick enough to accept him before the political difficulties became too heavy not only for him, but also for us.

Due to his exclusion from the communist party it was impossible to quote him as the author of his contributions and thus we included the first of them as an anonymous appendix to the volume *Automatische Textenbearbeitung* (Prague, Matematickofyzikální fakulta Univerzity Karlovy 1974, pp. 86-156. It constitutes an extremely systematic and detailed analysis of the functions of the English presposition in in different contexts, which even today can serve as an important source of relevant insights (in case of interest, the text may be copied and shared).

Most of Zdeněk's further contributions appeared in the internal series which followed this volume, namely *Explizite Beschreibung der Sprache und automatische Textbear-*

beitung (EBSAT), with a parallel Russian title. The first of these (again published anonymously) was his rich and many-sided dictionary of the terminology of computational linguistics (*Terminologisches Wörterbuch*, 1975), republished later in Poland in an enriched form by a group of authors under the editorship of K. Polański (1985). Already the original version was based on English terms, but it contained also their equivalents in French, German, Russian and Czech, as far as these equivalents existed.

Later, Zdeněk concentrated on issues of English-to-Czech machine translation and developed a successful system, based on an ingenious analysis of English morphemics and syntax, overcoming the difficulties connected with the absence of inflection in English, i.e. with the fact that English word forms by themselves mostly do not identify their functions without context (see Kirschner 1982;1984;1987).

Another highly important result of Zdeněk's research concerned information retrieval: His system MOSAIC was designed to account for automatic extraction of significant terms from Czech texts on the basis of the richness of the language in morphemic endings and derivational suffixes (see Kirschner 1983; Kiršner i Buraneva 1976). He worked on this approach, and also on its application to German, together with P. Pognan (Paris), before this cooperation was made impossible by some of our "coordinators" of that time, who suspected that in this way important information could escape to the imperialist western world.

Zdeněk was a wonderful personality, with a great moral and human influence on all his colleagues, on all of us. He has educated several of his followers who shared with him his involvement in natural language processing, be it on machine translation (let us mention e.g. Alexandr Rosen, a nowadays senior research worker at the Faculty of Arts of Charles University) or on information retrieval from full texts (e.g. Hana Vernerová). After having reached the age of 65, Zdeněk could only continue his collaboration with our research group as a retired specialist. Even so, his contributions played an important role in our work, and it can be only understood as our fault if we did not always exploit his efforts as they deserved. We all will miss his working spirit, enthusiasm, stimuli and good humour.

References:

EBSAT: Explizite Beschreibung der Sprache und automatische Textbearbeitung, Matematicko-fyzikální fakulta Univerzity Karlovy, Prague

(Kirschner Z.) (1975): *Terminologisches Wörterbuch. EBSAT 1.*

Kirschner Z. (1982): *A dependency-based analysis of English for the purpose of machine translation. EBSAT 9.*

Kirschner Z. (1983): *MOSAIC – A method of automatic extraction of significant terms from texts. EBSAT 10.*

Kirschner Z. (1984): On a dependency analysis of English for automatic translation. In: P. Sgall (ed.): *Contributions to functional syntax, semantics and language comprehension,* Prague:Academia, 335–358.

Kirschner Z. (1987): APAC 3-2: *An English-to-Czech machine translation system. EBSAT 13.*
Kiršner Z. i Buraneva E. (1976): Ob odnom sposobe avtomatičeskogo indeksirovanija. *Naučno-techničeskaja informacija, Serija 2, 9,* Moskva.
Polański K., ed. (1985): A terminological dictionary of algebraic linguistics. Katowice: Uniwersytet Śląski.

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009

Index to the Volumes 81–91

1. Structural linguistics

Articles

Reviews

Martin Čmejrek: *Vilém Mathesius Lecture Series 1,* 81, 2004, 55
Libuše Dušková (ed.): *Dictionary of the Prague School of Linguistics,* Rev. by Jun Qian, 81, 2004, 77–81

2. Formal description

Articles

Pavel Květoň: *Rule-based morphological disambiguation: On computational complexity of the LanGR formalism,* 85, 2006, 57–72
Jiří Havelka: *Projectivity in Totally Ordered Rooted Trees,* 84, 2005, 13–30
Radim Sova: *Genesis of Two Algebraic Theories of Language,* 83, 2005, 59–75
Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá and Daniel Zeman: *Issues of Projectivity in the Prague Dependency Treebank,* 81, 2004, 5–22

Reviews

Kenneth R. Beesley and Lauri Karttunen: *Finite State Morphology,* Rev. by Otakar Smrž, 81, 2004, 73–75

3. Semantics and discourse

Articles

Lucie Kučová, Kateřina Veselá, Eva Hajičová and Jiří Havelka: *Topic-focus articulation and anaphoric relations: corpus based probe,* 84, 2005, 5–12
Radim Sova: *The Sound-Meaning Relation in the Standard Theory of Transformational Grammar,* 84, 2005, 31–52
Jiří Semecký: *Automatic Assignment of Frame Semantics Using Syntax-Semantics Interface in LFG,* 83, 2005, 19–46
Václav Novák: *Towards Logical Representation of Language Structure,* 82, 2004, 5–86

Reviews

Agnes Celle and Ruth Huart (eds.): *Connectives as Discourse Landmarks,* Rev. by Šárka Zikánová, 88, 2007, 95–98
Sergei Nirenburg and Victor Raskin: *Ontological Semantics,* Rev. by Petr Němec, 86, 2006, 55–56
Edward Göbbel: *Syntactic and Focus-structural Aspects of Triadic Constructions,* Rev. by Eva Hajičová, 83, 2005, 77–80
Leonard Talmy: *Toward a Cognitive Semantics, Volume I, Concept Structuring Systems,* Rev. by Pavel Straňák, 83, 2005, 85–86

4. Computational linguistics and artificial intelligence

Articles

Reviews

Notes

5. Corpus annotation and parsing

Articles

Reviews

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009

INSTRUCTIONS FOR AUTHORS

Manuscripts are welcome provided that they have not yet been published elsewhere and that they bring some interesting and new insights contributing to the broad field of computational linguistics in any of its aspects, or of linguistic theory. The submitted articles may be:

- long articles with completed, wide-impact research results both theoretical and practical, and/or new formalisms for linguistic analysis and their implementation and application on linguistic data sets, or
- short or long articles that are abstracts or extracts of Master's and PhD thesis, with the most intersting and/or promising results described. Also
- short or long articles looking forward that base their views on proper and deep analysis of the current situation in various subjects within the field are invited, as well as
- short articles about current advanced research of both theoretical and applied nature, with very specific (and perhaps narrow, but well-defined) target goal in all areas of language and speech processing, to give the opportunity to junior researchers to publish as soon as possible;
- short articles that contain contraversing, polemic or otherwise unusual views, supported but some experimental evidence but not necessarily evaluated in the usual sense are also welcome.

The recommended length of long article is 12–30 pages and of short paper is 6-15 pages.

The copyright of papers accepted for publication remains with the author. The editors reserve the right to make editorial revisions but these revisions and changes have to be approved by the author(s). Book reviews and short book notices are also appreciated.

The manuscripts are reviewed by 2 independent reviewers, at least one of them being a member of the international Editorial Board.

Authors receive two copies of the relevant issue of the PBML together with 10 offprints of their article.

The guidelines for the technical shape of the contributions are found on the web site http://ufal.mff.cuni.cz/pbml.html. If there are any technical problems, please contact the editorial staff at pbml@ufal.mff.cuni.cz.

The Prague Bulletin of Mathematical Linguistics
NUMBER 92 DECEMBER 2009

LIST OF AUTHORS

Ondřej Bojar
Institute of Formal and
Applied Linguistics
Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
bojar@ufal.mff.cuni.cz

Silvie Cinková
Institute of Formal and
Applied Linguistics
Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
cinkova@ufal.mff.cuni.cz

Kamil Kos
Institute of Formal and
Applied Linguistics
Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
kamilkos@email.cz

Milan Malinovský
College of Civil Engineering, Department
of Languages
Czech Technical University of Prague
Thakurova 7
166 29 Praha 6, Czech Republic
malinov@fsv.cvut.cz

Elena Paducheva
Russian Academy of Science, Institute of
Scientific and Technical Information
125080 Moscow, Russia
elena708@gmail.com

Martin Popel
Institute of Formal and
Applied Linguistics
Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
popel@ufal.mff.cuni.cz

Jun Qian
English Department
Peking University
Beijing 100871, P.R. China
junqian@pku.edu.cn

Petr Sgall
Institute of Formal and
Applied Linguistics
Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
sgall@ufal.mff.cuni.cz

Zdeněk Žabokrtský
Institute of Formal and
Applied Linguistics
Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
zabokrtsky@ufal.mff.cuni.cz