NODALIDA-2007

FRAME 2007:
Building Frame Semantics Resources for Scandinavian and Baltic Languages

Workshop organizers:
Pierre Nugues and Richard Johansson
Department of Computer Science, Lund University

LUND UNIVERSITY

Department of Computer Science
http://nlp.cs.lth.se

Institute of Computer Science
http://math.ut.ee

NODALIDA-2007

FRAME 2007:
Building Frame Semantics Resources for Scandinavian and Baltic Languages

Workshop organizers:
Pierre Nugues and Richard Johansson
Department of Computer Science, Lund University

LUND UNIVERSITY

Department of Computer Science
http://nlp.cs.lth.se

Institute of Computer Science
http://math.ut.ee

Preface

Annotated data with role-semantic information are becoming an ever more important resource for many semantic systems. They form the core element to develop large coverage, high-performance, and reusable semantic parsers, classifiers as well as applications that include lexicography, term and information extraction, semantic processing of the web, text-to-scene conversion systems, etc.

Existing examples of role annotated corpora/resources include for English: FrameNet, PropBank, and VerbNet, for German: Salsa, and for Spanish: Spanish FrameNet. However, the two main initiatives outside English take FrameNet as a semantic pivot and attempt to derive or adapt frames to the target language using manual work or semiautomatic systems.

As frequently observed, the itemization of frames and lexical units and their manual annotation in a corpus is an expensive task that requires a relatively long-term and dedicated commitment. Such an effort is currently beyond the reach of most research teams in the Nordic/Baltic area, which could impair the quality, and possibly the existence, of future semantic applications in these languages. This makes the construction of a role-semantic annotated corpus and the design of automatic or semiautomatic transfer methods a challenge as well as an opportunity.

These proceedings contain the seven papers of the FRAME 2007 workshop, which was held on May 24, 2007 in Tartu, Estonia. They provide perspectives from various areas on current research in frame semantics and we hope they will foster new ideas to start the construction of role-annotated corpora in the Nordic/Baltic region and possibly share it across families of related languages.

This workshop would not have been possible without the authors and their contribution. We would like to thank them all as well as the organizers of Nodalida.

Richard Johansson, Pierre Nugues

The Workshop organizers

ISBN 978-91-976939-0-5

ISSN 1404-1200

Report 90, 2007

LU-CS-TR: 2007-240

Print: E-husets tryckeri, Lund 2007

Contents

Wordnets, Framenets and Corpus-based Contrastive Lexicology

Åke Viberg
Department of Linguistics and Philology, Uppsala University
Ake.Viberg@lingfil.uu.se

1 Introduction

In this paper, a Swedish FrameNet will be looked upon as a complement to Swedish WordNet (SWN), a first version of which was completed a few years ago. SWN is structured according to the principles of the original Princeton WordNet and in particular to its sequel EuroWordN (EWN). As we know, the basic unit in the wordnets is a synset, a set of synonyms which represent a certain meaning. The synsets are related according to a number of semantic relations such as hyponymy, meronymy and antonymy. At the end of the Swedish WordNet project, around 25 000 concepts were coded (around 5 000 concepts realized as verbs and around 20 000 concepts realized as nouns). With respect to words (literals), around 6 000 verbs and 27 000 nouns were included. Lists of the words included in SWN were run against frequency lists to check that no words with high frequency had been excluded, but needless to say, the present version only represents the core of a Swedish wordnet and needs to be extended.

At present, work is being carried out to extend the Swedish WordNet and to combine it with Swedish FrameNet, which is intended to form a Swedish counterpart to FrameNet developed for English by Charles Fillmore and his associates at Berkeley. Pilot work has been carried out on Swedish FrameNet with the coding of a selection of verbs. The work will not start in full scale until proper funding has been obtained. As in SWN, the intention is – as a first stage – to produce reliable coding of the core of Swedish vocabulary, in this case with particular focus on frequent verbs and semantically related abstract nouns and adjectives. The most frequent words (in particular verbs) tend to have meanings that form complex patterns of polysemy which in many respects are language-specific even when rather closely related languages such as English and Swedish are compared. Several examples of this can be found in studies using corpus-based contrastive analysis such as Viberg (1999, 2002, 2004, 2006). Another problem is language-specific semantic differentiation between basic words such as English *think* vs. Swedish *tänka/tycka/tro* (Viberg 2005). The semantic analysis presented in studies of this kind form a point of departure for the framenet coding. There is also a natural link to wordnets. Many frame elements are closely related to superordinate terms/top concepts in wordnets (e.g. Vehicle).

2 Language-specific differentiation

2. 1 The Swedish verbs of Thinking

The distinction between the three basic verbs of thinking *tänka*, *tro* and *tycka* is a well-known example of language-specific differentiation in Swedish. As shown in Viberg (2005), these three verbs are the major translations of English *think* in the English Swedish Parallel Corpus/ESPC (Altenberg & Aijmer 2000) in translations from English to Swedish, whereas *think* is the most frequent translation of each one of these verbs in the other direction. In the following, the discussion will be restricted to cases where the three verbs take a sentential complement.

The verb *think* appears in several frames in the FN database but the only lexical entry that is completed is related to the

Awareness frame. Two of the English examples are (with my Swedish translations):

(1) You don't *think* people ought to enjoy things	Du *tycker* inte att folk ska ha det bra
(2) He *thought* he was going to die	Han *trodde* han skulle dö

The definition of the Awareness frame is quoted in full in (D1).

(D1) "A Cognizer has a piece of Content in their model of the world. The Content is not necessarily present due to immediate perception, but usually, rather, due to deduction from perceivables. In some cases, the deduction of the Content is implicitly based on confidence in sources of information (*believe*), in some cases based on logic (*think*), and in other cases the source of the deduction is deprofiled (*know*). Note that this frame is undergoing some degree of reconsideration. Many of the targets will be moved to the Opinion frame. That frame indicates that the Cognizer considers something as true, but the Opinion (compare to Content) is not presupposed to be true; rather it is something that is considered a potential point of difference. In the uses that will remain in the Awareness frame, however, the Content is presupposed."

According to the old analysis, the sentential complements in (1) and (2) represent the FE Content. According to the newer analysis, they should rather be moved to the Opinion frame, which is defined as follows: "A Cognizer holds a particular Opinion, which may be portrayed as being about a particular Topic." This can be complemented with the definition of the FE Opinion: "The Cognizer's way of thinking, which is not necessarily generally accepted, and which is generally dependent on the Cognizer's point of view." Since the frame "indicates that the Cognizer considers something as true" (see D1), Opinion is a suitable FE for the complement of *tro* in (2). Simultaneously, this means that the FE Opinion would be different from the word *opinion* which covers also cases where evaluation rather than truth is involved. The most suitable alternative for the verb *tycka* is the frame Judgment which is defined as in (D2).

(D2) "A Cognizer makes a judgment about an Evaluee. The judgment may be positive (e.g. *respect*) or negative (e.g. *condemn*), and this information is recorded in the semantic types Positive and Negative on the Lexical Units of this frame. There may be a specific Reason for the Cognizer's judgment, or there may be a capacity or Role in which the Evaluee is judged. This frame is distinct from the Judgment_communication frame in that this frame does not involve the Cognizer communicating his or her judgment to an Addressee." An example of Judgment is: *She **admired** Einstein for his character*. Judgment_communication is illustrated with the following example: *She **accused** Einstein of collusion*.

The FE Judgment which is not mentioned in (D2) is defined as: "A description (from the point of view of the Cognizer) of the position of the Evaluee on a scale of approval." If *admire* is paraphrased 'think that someone is high on the scale of approval', this FE could be said to be incorporated into *admire* (and its Swedish counterpart *beundra*), whereas the Judgment is realized as a complement in a Swedish example such as *Hon tyckte att Einstein hade en beundransvärd karaktär* 'She thought that E had an admirable character.'

Having found suitable candidate frames for *tro* and *tycka*, the problem remains of finding a suitable frame for *tänka*, the most general of the Swedish verbs of thinking. One frequent use is to report direct and indirect thought as in (3) and (4). (cf the use of 'say' to report direct and indirect speech).

Direct thought

(3) Men oj! ***tänker*** flickan. MR	Oh, help, the girl ***thinks***.

Indirect thought:

(4) Jag ***tänker*** blixtsnabbt att jag inte vill kyssa honom. MS	In a flash I ***think*** that I don't want to kiss him,

When *tänka* is used to report indirect thought it takes a sentential complement in the same way as *tro* and *tycka*. In principle, *tänka* can be used to report any thought, even those that represent an opinion or a judgment as in (5).

(5) På vägen ***tänkte*** han att allt hade gått bra	As he drove, ***it occurred to*** him that everything had gone well,

It seems most reasonable, however, to say that distinctions such as opinion and judgment are neutralized, and several examples such as (4) do not belong to any of these categories. Furthermore, there is often another difference, as in (5). The verb *tänka* tends to refer to the actual occurrence of a thought in the consciousness of the cognizer at a specific moment in time. Opinions and judgments are more like dispositions to think in a certain way (propositional attitudes) and need not appear in consciousness at reference time. You can say even about a sleeping person *Hon tycker att Ingmar Bergman är intressant* 'She thinks that IB is interesting'. You can 'hold' an opinion (or judgment) for a long time. The frame that appears as the best candidate for *tänka* in the uses discussed here is Mental_Activity which is defined as in (D3).

(D3) Mental_Activity "In this frame, a Sentient_entity has some activity of the mind operating on a particular Content or about a particular Topic. The particular activity may be perceptual, emotional, or more generally cognitive. This non-lexical frame is intended primarily for inheritance."

The complement of *tänka* used to report indirect thought as in (4) represents the FE Content which is defined as "The situation or state-of-affairs that the Sentient_entity's attention is focussed on." Obviously, this FE cannot be used in the revised Awareness frame, if the content is to be presupposed as indicated in (D1). A way out would be to introduce an FE like Fact to refer to the complement of LUs like *know* and *be aware*. In that case, Content could be regarded as a neutral frame which is a schematic version of more specific frames such as Opinion, Judgment and Fact. Actually, English *think* with a sentential complement could probably best be represented as neutral in this way. In many cases when *think* appears with a sentential complement in an English original text, it is necessary to use pragmatically based inferences to decide which one of the Swedish verbs *tänka*, *tro* or *tycka* is the most suitable translation.

The report of direct thought as in (3) should be treated in parallel with the treatment of direct speech in the Communication frame, which basically has the structure shown in (D4)

(D4) Communication A Communicator conveys a Message to an Addressee: *[I] TOLD [her] [it was raining]*. The Message can be refined in four ways, the most important of which are Message-Content: *I SAID [that I was planning to quit]* and Message-Form: *She SAID ["I can't stand this any longer!"]*.

By analogy with Message-Form, the direct report of thought that appears in (6) should be called Thought-Form.

(6) Nu tvingar jag dej, tänker flickan. MR	I'll make you now, the girl thought.

Note that the verb *tycka* can be used also as a communication verb as in (7).

(7) - Bra idé, tyckte Franklin. ARP	'Good idea,' said Franklin.

Bra idé is an example of Message-Form. Simultaneously, the use of *tycka* in the Swedish version requires that the content is a Judgment (cf hybrid frames, below). To sum up this section, it can be concluded that it is possible to find frames that can be used to represent the contrast between *tycka*, *tro* and *tänka,* but that requires several modifications of the existing frames to accommodate the language-specific aspects of the Swedish verbs. It remains an open question what will happen when more languages are taken into consideration. Probably, it will be necessary to accept language-specific frames that inherit part of their structure from more general frames. According to current work on linguistic relativity such as Bowerman & Levinson (2001), part of conceptual structure to which frames belong is language-specific.

2.2 The verbs of Placing

The differentiation between *sätta*, *ställa* and *lägga* which all belong to the around 50 most frequent verbs in Swedish is another well-known example. In examples like (8)-(10), a choice must be made when translating *put*.

(8) She ***put*** the bowl on a windowsill in her sun porch, GN	Hon ***ställde*** skålen på en fönsterbräda på sin solveranda
(9) I took my letter out of the envelope and ***put*** it on the table, RDO	Jag tog ut mitt brev ur kuvertet och ***la*** det på bordet,
(10) She ***put*** on a pair of cheap hoop earrings FW	Hon ***satte*** ett par enkla ringar i öronen

The verb *put* and its Swedish equivalents are realizations of the Placing frame which is defined as in (D5).

(D5) Placing. "Generally without overall (translational) motion, an Agent places a Theme at a location, the Goal, which is profiled. In this frame, the Theme is under the control of the Agent/Cause at the time of its arrival at the Goal." Example: David [Agent] placed his briefcase [Theme] on the floor [Goal]

In this case, there is no way to mark the contrasts with the existing frame elements. On the other hand, close to 70 English verbs are given in the list of verbs that evoke this frame without any systematic indication of what differentiates them. Of course, it is an open question to what extent this is desirable. For certain purposes, FN may be used to extract information of a more general kind and in that case the Placing frame provides adequate information, and a more fine-grained analysis may be regarded as a cumbersome extravagance. However, if FN is used as a model for contrastive analysis, it is essential to be able to tease apart similarities and language-specific features. The Placing frame is part of an interlingua that shows what English and Swedish have in common. One characteristic where English is special with respect to Swedish is the relatively high number of verbs sharing the meaning 'put into a container', where the Goal is incorporated in the verb as in *archive*, *bag*, *box*, *bottle*, *cage*, *crate*, *pocket* and *shelve*. Examples of the analysis of such verbs are: *The items* [Theme] *are then* ***bagged*** [Goal] *by the Scenes of Crime Officer* [Agent] and *My* [Agent] *main task was to* ***bottle*** [Goal] *wine* [Theme]. Even if a few verbs of this type exist in Swedish such as *arkivera* 'archive', such verbs are usually translated with the container specified as part of the Goal realized as a PP as in (11) where *box* is expressed as 'pack in boxes'.

(11) ***boxing*** plums was not the work to satisfy a youth like Joseph. JC	***packa*** plommon ***i lådor*** var inte den sortens sysslor som tilltalade en yngling

som Joseph.

The English verbs of the type 'put in a container' can be described as a kind of incorporation of the Goal into the verb. One way to represent the differentiation between verbs such as Swedish *lägga* and *ställa* would be to describe this as an incorporation of an FE like Result. The contrast between *ställa* and *lägga* has to do with the resulting orientation of the Theme (in Upright vs. Horizontal position), whereas *sätta* in the most typical case signals attachment of the Theme to the Goal. (A more detailed description of the semantic contrasts are given in Viberg 1998.)

3. Hybrid frames

Incorporation of frame elements is one way of extending the English framenet to account for patterns in other languages. Another characteristic of framenet which makes it possible to account for new data is the use of hybrid frames. In this section, the use of hybrid frames to account for verbs referring to sounds are presented as the major example. Actually, there are a rather large number of verbs that in various ways refer to a characteristic sound, as the verbs in (12) and (13), which are typical examples of the Make_noise frame defined in (D6).

(D6) Make_noise "A physical entity, construed as a point-Sound_source, emits a Sound. This includes animals and people making noise with their vocal tracts." Example: The wind [Sound_source] howled.

(12) Baklastarna ***råmade*** och ***tjöt***. MPC:MN	The bulldozers ***bellowed*** and ***roared***.
(13) Nora kunde höra att det ***mullrade*** till nånstans MG	Nora could hear ***a rumbling*** somewhere.

Characteristically, the verbs referring to sound are used with many different meanings. The verb *tjuta* and *mullra*, for example, can be used as motion verbs (14-15) and as communication verbs (16-17).

(14) Lukas drog i ångvisslan: som ett fasans skri ***tjöt*** ångan **ut ur** ventilen. ARP	Lukas jerked the cord of the steam whistle and like a scream of terror, steam ***screeched*** **out of** the valve.
(15) /---/ när tågen ***mullrade*** **förbi** över oss. RJ	/---/ when the trains ***roared*** **past**.
(16) - Det var inte mitt fel, ***tjöt*** pojken. MPC:LM	"It wasn't my fault!" the boy ***wailed***.
(17) — Haha! ***mullrade*** slaktaren det var inte mycket att bita i! ARP	'Ha-ha!' ***rumbled*** the butcher. 'Nothing much to bite there!

As motion verbs, *tjuta* and *mullra* in (14-15) can be described with the general Motion frame (D7).

(D7) Motion "Some entity (Theme) starts out in one place (Source) and ends up in some other place (Goal), having covered some space between the two (Path)."

However, simultaneously as the verbs in (14-15) describe a motion, they also describe various types of sound emission. To catch this, a hybrid frame like Motion_noise defined in (D8) is used in FrameNet.

(D8) Motion_noise "This frame pertains to noise verbs used to characterize motion. Motion_noise verbs take largely the same Source, Path and Goal expressions as other

types of Motion verbs." Example: The limousine purred forwards [Path] into the traffic [Goal]

In a similar way, the hybrid frame Communication_noise (defined below) is used to describe examples such as (16-17). This is an amalgamation of the Communication frame (D4) with the frame Sound_movement (D9), which is primarily used with verbs that describe the motion of a sound realized linguistically as a noun.

(D9) Sound_movement "A Sound emitted by a Sound_source, which construed as a single point, moves along a Path. Rather than the Sound_source itself, the Location_of_sound_source may be mentioned. Essentially, this frame denotes the (semi-) fictive motion of the Sound." Example: Laughter [Sound] echoed through the hall [Path]

Typical Swedish examples taken from the Bank of Swedish (RomI = Novels I) are shown in (18-19).

(18) Ugliks tjut ***ekade*** mot väggarna. RomI	Uglik's scream ***echoed*** off the walls. (My transl.)
(19) Babyn däruppe ***tjöt*** genom trossbottnen. RomI	The baby upstairs ***screamed*** through the double ceiling. (My transl.)

Together with the Communication frame, this frame forms the hybrid frame Communication_noise defined in (D10).

(D10) Communication_noise. Hybrid of Communication (D4) and Sound_movement (D9):"This frame contains words for types of noise which can be used to characterize verbal communication. It inherits from Communication (possibly more specifically Communication_manner) and the Sound_emission frame (which simply characterizes basic sounds of whatever source, including those made by animals and inanimate objects). As such, it involves a Speaker who produces noise and thus communicates a Message to an Addressee." (The Sound_emission frame cannot be found in the database. The closest correspondent I have been able to find is Sound_movement.)

In several cases there is a clear reference to the motion of the sound such as *ner* 'down' in (20).

(20) Och hur hon ***skrek*** **ner** mot Eeva-Lisa att hon skulle ut. MPC:POE	And how she ***screamed*** **down** at Eeva-Lisa that she had to go.

Actually, it is possible to find examples with most communication verbs where there is a clear reference to the motion of the sound. The (semi-fictive) motion of the sound is referred to even in some examples with Statement verbs such as *säga* 'say' as in (21).

(21) Till Fögelke ***sade*** jag **genom dörrspringan**: ta Lejbus' sax, RomII	To Fögeleke, I ***said*** **through the crack of the door**: Take Leibus's pair of scissors (My transl.)

In principle, it is possible to use a wide range of communication verbs in the same context; you can promise or threaten or tell a story through the crack of a door. In the present version of FrameNet, the FE Medium is used within the communication frame: "Medium is the physical entity or channel used by the Speaker to transmit the statement." One of the examples provided is: Kim preached to me over the phone [Medium]. In examples of this type, Medium is an appropriate analysis but examples such as (21) are more naturally interpreted with reference to a hybrid frame combining Motion and Communication. Oral

communication is often conceived as the transmission of messages via sound that travels between speaker and hearer.

A tricky case is the description of directional complements of visual perception verbs. Modern science tells us that vision is the result of light moving from a perceived entity to our retina where it gives rise to a chain of recodings at various levels. Ordinary language is based on several, partly contradictory conceptualizations, one of which seems to rest on the assumption that something moves from our eyes: Examples such as (22-24) describe a motion away from the perceiver. Consider also expressions like *cast an eye on* which have parallels in many languages.

(22) Och sju trädgårdar kunde hon ***se*** **från** sitt fönster. MG	**From** her window she could ***see*** seven gardens
(23) De ***kikade*** **in genom** de gardinlösa fönstren HM2	They ***peeked*** **in through** undraped windows
(24) Hon ***tittade*** **upp mot** husen MPC:LM	She ***looked up at*** the houses

Winer et al (2002) account for a number of psychological studies which show that the belief that vision includes emanations from the eyes is present among American college students. Actually, this belief – referred to as the extramission theory of perception – was held also by Greek philosophers and existed even in scientific circles until Kepler's work on the retinal image.

4. Verbal particles

The frequent use of verbal particles, which is a characteristic feature of English, is not dealt with in any detail in FrameNet. Arguably, particles are even more important in Swedish. In principle, particles can often be treated as frame elements. Examples can be found in the FrameNet database, for instance in the description of the frame Self_motion, which is defined "The Self_mover, a living being, moves under its own power in a directed fashion /---/"), a typical example being: *The cat* [Self-mover] ran *out of the house* [Source]. There are also examples of FEs realized as single particles: *The cat* ran *out* [Source]. *The principal* walked *over* [Goal] *and sat down*. Examples like these are similar in English and Swedish. More problematic are cases when the direction is incorporated into the verb root as in *enter*. In this case, the Goal is realized as a direct object: *The messenger* [Theme] entered (*the room* [Goal]). The verb *enter* is related to the frame Arriving ("An object Theme moves in the direction of a Goal. The Goal may be expressed or it may be understood from context, but it is always implied by the verb itself.") Swedish does not have a direct equivalent of *enter*. Ex. (25) is taken from an English original text in the ESPC.

(25) Then he **entered** the sitting room and threw on the light. FF	Sedan ***gick*** han ***in i*** vardagsrummet och tände ljuset.

Examples like (25) may be analyzed by saying that English in this case uses the Arrival frame, whereas Swedish uses the Self_motion frame. The reference to different frames is justified by the fact that the English and Swedish versions are not equivalent out of context. The English verb *enter* is unmarked for intention and for manner of motion, whereas Swedish *gå* is intentional and always refers to walking when the subject is human. We can leave it at that or try to account for the differences by referring to a more abstract version of the motion scenario along the lines of Talmy (1985). There is a shared representation which basically looks as follows: A Theme moves [into]Path [room]Goal [by walking]Means. In

English, Path is incorporated into the verb, whereas Means which is not expressed in English must be incorporated into the Swedish verb. This difference between English and Swedish may appear relatively minor since both languages belong to the satellite-framed languages in Talmy's sense, but as is well known, there are a number of verb-framed languages, such as French, where incorporation of Path represents a basic pattern. In (26), Manner is expressed as an adverbial and in (27) it is left unexpressed, which represents the most frequent alternative. (These and the following examples from three languages are taken from the MPC corpus consisting of extracts from Swedish novels and their translations into various languages.)

(26) - Sorry, sa nattchefen när han **susade *in i*** rummet LM	"Sorry," the night editor said as he **hurtled *into*** the room,	- Désolé, lança le rédacteur en chef en ***entrant* en trombe** dans la pièce,
(27) Christina sätter nyckeln i köksdörren och öppnar, **glider *in*** och tänder ljuset. MA	Christina puts the key in the lock and opens the back door, **glides *inside*** and turns on the light.	Christina sort la clé, ouvre la porte de la cuisine, ***entre*** et allume la lumière.

A special case is represented by several Swedish particles that lack (a frequent) equivalent in English. One such particle is *ihjäl* (etymologically into Hell/Hel) as in (28).

(28) Då anmälde den andra kärringen Signe Persson för att katten hade **haft ihjäl** hennes undulat. SW	Then the other old lady made a complaint against Signe Persson, because the cat had **killed** her budgie.

In expressions such as *ha ihjäl* and arguably also *slå ihjäl*, the manner component is fairly neutralized, and it would be possible to treat them as lexical units ("phrasal verbs"). The use of the particle is, however, fully productive and can be used with many verbs expressing fine-grained manner distinctions as in (29) and (30).

(29) Den äldre albatrossungen **hackar** så ***ihjäl*** den yngre. POE	Then the older baby albatross **pecks** the younger one ***to death***.	Le bébé albatros le plus âgé ***tue*** alors le plus jeune **à coups de bec.**
(30) I stallet **törstade** hästen ***ihjäl***.	his horse ***dying* of thirst** in the stable.	Dans l'écurie, son cheval était ***mort* de soif.**

What happens in examples of this type is that the information in the main verb is degraded to a manner component whereas the particle refers to the focused event. The Killing frame is defined as follows: "A Killer or Cause causes the death of the Victim." Example: John [Killer] drowned Martha [Victim]. In this example, the manner is incorporated into the main verb. Ex. (29) can be derived from an underlying structure like: A Killer causes the death of a Victim by pecking [Means]. Ex. (30) represents an inchoative version of the Killing frame.

Another example from Swedish is the particle *sönder* which is the closest correspondent to *break* (in its basic sense). Intransitive *break* is a realization of the Fragmentation_scenario ("A Whole fragments or breaks into Parts"), whereas transitive break is related to the frame Cause_to_fragment ("An Agent suddenly and often violently separates the Whole_patient into two or more smaller Pieces, resulting in the Whole_patient no longer existing as such.") Ex: *I* [Agent] smashed *the toy boat* [Whole_patient] *to flinders* [Pieces]. *Break* is also related to the frame Render_nonfunctional ("An Agent

affects an Artifact so that it is no longer capable of performing its inherent function.") In Swedish, the most frequent translation of *break* is *gå sönder* 'go apart' as in (31), when *break* is intransitive, and *slå sönder* 'strike apart' as in (32) when it is transitive (*ha* 'have' and *göra* 'do/make' *sönder* are also used within a formal and a spoken register, respectively).

(31) The glass didn't **break** in the frame. BO	Glaset i ramen ***gick*** inte ***sönder***.
(32) Jane going round plates matters; FW	Att Jane går omkring och ***slår sönder*** tallrikar, det har betydelse,

As argued in Viberg (1985), written within a different theoretical framework, Swedish *sönder* in its prototypical use combines two core components which roughly could be paraphrased as '(separate) into pieces' and 'not possible to use (in the conventional way)'. The FE Means, which is defined as "The action that the Agent performs which results in the Artifact being inoperable", can be incorporated into the verb in Swedish. Literally, Swedish uses a phrase meaning 'scream apart' in (33) to realize a meaning such as 'to cause to become nonfunctional by screaming'.

(33) Han hade **skrikit** ***sönder*** nånting. KE	He had ***damaged*** something **by screaming**.
Quelque chose s'était ***cassé*** quand il avait **crié**.	

To sum up, incorporation of frame elements appears to be a promising way to describe differences between languages related to the use or not of verbal particles.

5. Conclusion

In my view, FrameNet represents a fascinating further development of lexical databases after WordNet that today is available in some version in a large number of languages. This paper has been concerned with the use of framenets and frame semantics for corpus-based contrastive analysis. For this purpose, it is important to work out a fine-grained analysis to account for the contrasts between words (lexical units) that evoke the same frame, for example Placing, as discussed above. One way of extending framenet for contrastive purposes is the further development of the existing model by adding subframes, hybrid frames or by referring to various kinds of incorporation of frame elements. It is still an open question, however, how far this approach should be followed. For certain purposes, it may be more advantageous to combine framenet with some variety of componential analysis to differentiate between words evoking the same frame.

As for practical applications, contrastive analysis is important for work on translation and for language learning. In particular with a view to language learning with which I am most familiar, a major problem is patterns of polysemy that have a tendency to give rise to various transfer phenomena. Like Wordnet, FrameNet assigns different representations to each sense of a polysemous word. However, the relationships between various senses of a word are not accounted for in a systematic way to any greater extent. One device that appears to be useful for this purpose is found in the frame-to-frame relations such as inheritance, subframe, Causative_of and Inchoative_of. In spite of this, this is an area where much remains to be done.

References

Bengt Altenberg and Karin Aijmer. 2000. The English-Swedish Parallel Corpus: A Resource for Contrastive Research and Translation Studies. In *Corpus Linguistics and Linguistic Theory*, C. Mair and M. Hundt (eds), 15–33. Rodopi, Amsterdam and Atlanta.

Melissa Bowerman & Stephen Levinson (eds.) 2001. *Language acquisition and conceptual development*. Cambridge University Press, Cambridge.

Leonard Talmy. 1985. Lexicalization patterns: semantic structure in lexical forms. In T. Shopen (ed.), *Language typology and syntactic description. III. Grammatical categories and the lexicon*. Cambridge University Press, Cambridge.

Åke Viberg. 1985. Hel och trasig. [In Swedish. 'Whole and damaged'] *Svenskans beskrivning 15*. Göteborgs universitet, Göteborg: 529-554.

---1998. Contrasts in polysemy and differentiation: Running and putting in English and Swedish. In: S. Johansson, & S. Oksefjell (eds.), *Corpora and Cross-linguistic Research*. Rodopi, Amsterdam: 343-376.

--- 1999. The polysemous cognates Swedish *gå* and English *go*. Universal and language-specific characteristics. *Languages in Contrast*, 2(2): 89-115.

--- 2002. Polysemy and disambiguation cues across languages. The case of Swedish *få* and English *get*. In B. Altenberg & S. Granger (eds.) *Lexis in contrast*. Benjamins, Amsterdam: 119-150.

--- 2004. Physical contact verbs in English and Swedish from the perspective of crosslinguistic lexicology. In: K. Aijmer & B. Altenberg (eds.) *Advances in corpus linguistics*. Rodopi, Amsterdam/New York: 327-352.

--- 2005. The lexical typological profile of Swedish mental verbs. *Languages in Contrast,* 51(1): 121-157.

--- 2006. Towards a lexical profile of the Swedish verb lexicon. *Sprachtypologie und Universalienforschung.* 59(1): 103-129.

Winer, G., Cottrell, J., Gregg, V., Fournier, J. & Bica, L. 2003. Fundamentally misunderstanding visual perception: Adults' belief in visual emissions. *American Psychologist Vol. 57:6-7*, 417-424.

Electronic resources

The Bank of Swedish:
http://spraakbanken.gu.se/

FrameNet:
http://framenet.icsi.berkeley.edu/.

WordNet:
http://wordnet.princeton.edu/

Global WordNet and EuroWordNet:
http://www.globalwordnet.org/

Swedish WordNet:
http://www.lingfil.uu.se/ling/swn.html.

Medical Frames as Target and Tool

Lars Borin, Maria Toporowska Gronostaj, Dimitrios Kokkinakis
Göteborg University
Department of Swedish Language
Språkdata/Språkbanken
Sweden
{first.last}@svenska.gu.se

Abstract

In this paper we present a pilot study on the development of a FrameNet-like annotation of a sample of Swedish medical corpora, for a selected set of verbal predicates. We explore and exploit a number of linguistic tools for the provision of much of the necessary annotations required by such a semantic scheme. Particular attention is paid to the syntactic and semantic roles of scheme elements. We discuss in detail methodological issues and take up the relevance of our research for natural language processing (NLP) tasks.

1 Introduction

The conviction that enrichment of corpora with annotation layers of syntactic and semantic information will provide valuable support for refined text mining has been the main impetus for this corpus oriented pilot study. We have explored cumulative morphosyntactic text processing as a preliminary stage in semantic tagging. The main goal of our study has been to examine whether such integration of information can in a significant way contribute to semi-automatic acquisition and extraction of semantic schemes from corpora, in particular in the medical domain. By semantic schemes we mean frame-like constructions analogous to those in FrameNet. Formally, "FrameNet annotations are constellations of triples that make up the frame element realization for each annotated sentence" (Ruppenhofer et al., 2006:6), i.e. grammatical function [e.g. Subject]; frame-element [e.g. HUMAN]; phrase type [e.g. NP]. FrameNet resources have been recently developed for a number of languages, e.g., Spanish, German and Japanese. The FrameNet project (Baker et al., 1998) builds upon the theory of semantic frames formulated by Fillmore (1976), supported by corpus evidence. It is assumed here that access to such formalized semantic schemes can significantly improve the semantic component of a number of NLP tasks requiring semantic processing, including question-answering, automatic semantic role labelling, natural language generation, and information extraction (IE), in which there is a direct correspondence between frame-like structures and templates. Templates in the context of IE are frame-like structures with slots representing the basic components of events (cf. Surdeanu et al., 2003).

Related work is presented in section 2. The methodology underlying the morphological, syntactic and semantic pre-processing is outlined in section 3. Section 4 deals with the issues concerning lexical annotation of medical corpora. In section 5 we discuss the possibility of semi-automatic acquisition of frames based on qualitative and quantitative criteria. We end the article with conclusions and discussion.

2 Related Work

There are a number of approaches to FrameNet-like annotation including the influential work by Gildea and Jurafsky (2002) and Gildea and Palmer (2002), who point to the necessity of using syntactic information for the semantic annotation task and for predicting semantic roles based on the FrameNet corpus; the use of named-entity recognition by Pradhan et al. (2004) and others; see for instance the CONLL 2004 and CONLL 2005 shared tasks for semantic role labeling[1] and the SemEval-2007 Frame semantic structure extraction

[1] <http://www.lsi.upc.edu/ ~srlconll/>

task.[2] In our context, the work by Johansson & Nugues (2006) on Swedish is of particular relevance. In their work a corpus was annotated using cross-language transfer from English to Swedish. However, closer to our goals has been the work described by Wattarujeekrit et al. (2004); Huang et al. (2005), Cohen and Hunter (2006) and Chou et al. (2006) within the (bio)medical domain.

3 Methodology

3.1 Corpus Sampling and Annotation

We started by sampling a large number of sentences from the MEDLEX corpus (Kokkinakis, 2006), a large collection of articles from the medical domain, currently comprising about 45,000 documents. The sampling was performed after the identification and selection of a set of 30 important verbs, according to their significance compared to general newspaper corpora and which indicate events containing medical entities. Examples of such verbs are *operera* 'to operate', *behandla* 'to treat', *injicera* 'to inject', *vaccinera* 'to vaccinate' and *palpera* 'to palpate. The medical entities were supplied by the use of a Swedish MeSH tagger[3] for the categories *anatomy* (A), *organisms* (B), *diseases* (C), *chemicals and drugs* (D), *analytical, diagnostic and therapeutic techniques and equipment* (E), and *psychiatry and psychology* (F). Although MeSH is a valuable resource, it is rather limited in coverage considering the wealth of terminology in medical language. Therefore, we have complemented the MeSH annotations by developing a module that recognizes important types of (medical) terms, particularly *names of pharmaceutical products, drugs, symptoms* and (anatomical) *Greek and Latin terms*. Named entity tags were also added to the sample. A generic named entity tagger was applied which recognizes and annotates eight main types of named entities; *person, location, organization, object/artifact, event, work, time* and *measure expressions*; for details see Kokkinakis 2004.

[2] <http://framenet.icsi.berkeley.edu/semeval/FSSE.html>

[3] The Medical Subject Headings (MeSH) is the controlled vocabulary thesaurus of the U.S. National Library of Medicine (NLM), widely used for indexing medical data. The MeSH is a hierarchical thesaurus. The Swedish MeSH tagger is based on the Swedish translation made by staff at the Karolinska Institute Library (<http://mesh.kib.ki.se/swemesh/>) which contains 22,325 entries. MeSH is the central vocabulary component of the UMLS, frequently used as a provider of lexical medical information for biomedical natural language processing tasks (bio-NLP).

The net effect of the preprocessing described in this section is that the NPs in the sample sentences are annotated with their semantic classes, which turns out to be a very useful piece of information to have when parsing the sentences.

3.2 Streamlining Parsing with Semantic Classes

Grammatical functions are one of the main features and prerequisites for the realization of FrameNet annotations. Therefore, the semantic class annotations described above, together with part-of-speech tags, were merged into a single representation format and fed into the syntactic analysis module, which is based on the Cass parser (*Cascaded analysis of syntactic structure*; see Abney 1997). The Cass parser is capable of annotating grammatical functions and is designed for use with large amounts of (noisy) text. Cass uses a finite-state cascade mechanism and internal transducers for inserting actions and roles into patterns. The Swedish grammar used by the parser has been developed by Kokkinakis and Johansson Kokkinakis (1999), and has been modified and adapted in such a way that it is aware of the features provided by the pre-processors, particularly the medical terminology.

The annotations produced by the entity and terminology taggers significantly reduce the complexity of the sentence content, which in turn reduces the complexity of the parsing task, since the sentences contain fewer tokens, with less complex phrases, and thus can be more reliably parsed. Consider the example in figure 1, which, after the pre-processing stages, has been reduced from 26 to 10 tokens and 6 annotations, while a complex noun phrase, *cancer coli Duke's B*, has been replaced by a single label, '<DISEASE>'.

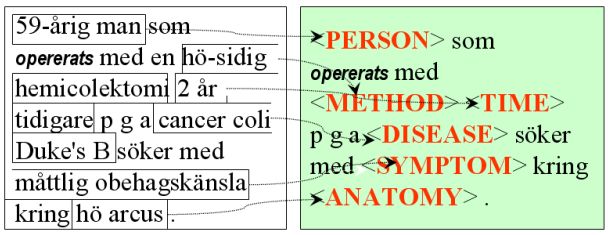


Figure 1. Simplification of input sentences

The syntactic analyses produced by the parser were in turn transformed into the TIGER-XML interchange format (König & Lezius, 2003), a flexible graph-based architecture for storage, indexing and querying of syntactically analyzed texts (appendix

1a). Our main purpose for doing this was that we wanted to apply existing software for manual frame annotation and for the analysis and inspection of the results, namely the SALSA/ SALTO tool (Burchardt *et al.*, 2006), which requires TIGER-XML input, thus minimizing the software development overhead (appendix 1b). Using this method we are now in the process of developing a semantically annotated sample that can be further used for experiments with machine learning algorithms.

4 Medical Frames as Target

4.1 Medical Frames in FrameNet

Access to multilayered lexical and grammatical information representing the content of texts is one of the prerequisites for an efficient understanding and generation of natural language. The FrameNet approach, with roots in Fillmore's case roles, offers an interesting approach to the study of lexical meaning described in terms of semantic frames. Semantic frames are generalisations of conceptual scenarios evoked by predicates and their frame elements. According to Ruppenhofer et al. (2006) there are roughly 780 semantically related frames (10,000 word senses/lexical units) accounted for in FrameNet. For each frame, there is a set of lexical units listed and exemplified with semantically and syntactically tagged examples from the British National Corpus (BNC). A small subset of these frames pertain directly to medical scenarios, like *Medical conditions, Experience bodily harm, Cure, Health response, Recovery, Institutionalization, Medical instrument.* Other, more general ones like *Placing* and *Removing,* do this in an indirect way by including lexical units of medical terminology dealing with notions of implanting or removing body parts. An overview of a repository of medically related frames in FrameNet with specification of core and non-core frame elements is provided in appendix 2b. The core frame elements, capturing the semantic valence of predicates, are obligatory ones, while the non-core ones add optional information.

The semantic salience of the types of core elements listed in appendix 2b applies also to Swedish. However, whenever designing frame-like schemes for specific sub-domains, further descriptive detail might be called for. Conflation of conceptually similar frame elements, e.g. Ailment and Affliction, semantic role overlap between general and specific roles as for example Agent and Healer, and postulation of new medical schemes are some of the issues which need to be considered when building a similar resource with focus on medical scenarios for Swedish.

4.2 From Frame Elements to MeSH Categories and Scheme Elements

Mapping medical frame elements onto the corresponding concepts in a thesaurus-based lexicon turns a relatively information-poor lexical resource into a more expressive and robust one and hence more useful for semi-automatic semantic annotation of corpora. For annotating the Swedish corpus, we have used our thematically sorted lexicons with medical vocabulary and the Swedish data from MeSH.

Since the MeSH vocabulary is sub-classified according to topics like anatomy, diseases etc., there is a possibility of mapping between some medical core concepts in the FrameNet and the top nodes in MeSH classification including their hyponyms. The results of this mapping are indicated in table 1:

Core frame elements in FrameNet	MESH thesauristic nodes
Ailment, Affliction	Diseases
Body_parts	Anatomy
Medication	Chemicals and Drugs
Treatment	Analytical, Diagnostic and Therapeutic Techniques and Equipment
Patient	Persons

Table 1. Mapping core frame elements onto MeSH top nodes

As already mentioned above (section 3.1), the tag set based on the MeSH top nodes has been further enlarged with thematic lists for both medical concepts like *symptoms* and supplementary named entities such as *time*, *location*, *measure* etc. All of these occur frequently in combination with the verbs selected for our sample. Since the sample came from a medical corpus, the instantiated uses of the verbs represent predominantly their medical senses. To make the semantic medical schemes appear more distinct the corpus sentences have been syntactically pre-processed, i.e., complex syntactic phrases containing syntactic dependences have been analysed to find their semantic heads, which have been subjected to semantic annotation, with the exception of noun phrases containing two or more medical tags. The latter will undergo further analysis for detecting types of medical

collocations. Examples (i) and (ii) below illustrate the annotated corpus.

(i) *<TIME> har <PERSON-GRP> opererats i <PLACE> för sina <SYMPTOM> i <ANATOMY> . (Original sentence: "Sedan 1987 har cirka 7 000 personer opererats i Sverige för sina svettningsproblem i händerna")*

(ii) *<DISEASE> i <ANATOMY> - <DISEASE> - kan <TIME> opereras med utmärkt resultat om durationen är <TIME> . (Original sentence: "Bristning i centrala retina - makulahål - kan idag opereras med utmärkt resultat om durationen är under 4-6 månader .")*

As follows from the above, the focus in our work is on the semantic types of referents, and thus our methodology contrasts with the FrameNet approach which takes the predicate and the evoked role scenario as the point of departure for determining a set of frame elements. The tags in our corpus are meant to provide a first approximation of medical semantic schemes by naming the types of annotated elements. To make the distinction between FrameNet and our approach clear, the terms *semantic schemes* and *scheme elements* are used henceforth in our study. A quantitative overview of semantic tags in the sample sentences (700 000 tokens) is given in the table 2.

Semantic labels	# in the whole sample (# with *operera*)
DISEASE	22 100 (1 346)
ANATOMY	11 080 (1 528)
CHEMICAL	10 450 (186)
METHOD	2 276 (467)
ORGANISM	4 090 (7)
PERSON	12 434 (1460)
PERSON-GRP	11 810 (829)
LOCATION	3 024 (216)
TIME	19 131 (897)
MEASURE	3 732 (319)

Table 2. Semantic annotations in the sample sentences

4.3 Case Study: Medical Senses of *operera* 'to operate'

To assess the correctness of our assumptions and the possible advantages or disadvantages of the chosen methodology, we have taken a closer look at the Swedish verb *operera*, whose medical sense ('perform surgery') is not described in FrameNet. The verb *operera* is polysemous in both Swedish and English, but only its medical senses are considered below, as the corpus and the pilot study is restricted to the medical sub-domain. In the following we select some of the frequent schemes instantiated in the corpus in order to examine the types of the medical scenarios this verb can evoke (appendix 1c illustrates dependency concordances with *operera*). The verb *operera* in its medical readings occurs in the corpus as either a simplex, reflexive or particle verb (phrasal verb) followed by the particles *bort* or *ut* (away, out) or *in* (in), as illustrated below:

- **simplex** *operera*: two sub-senses and thus two partly different schemes are represented in the corpus:

(i) to give consent to and undergo a surgical procedure with PERSON used in the double role of both semi-Agent and Experiencer, with ANATOMY and DISEASE as possible core arguments;

*e.g. <PERSON (semi-Agent & Experiencer)> **har** precis **opererat** <ANATOMY> i <ANATOMY> (Original sentence: Jag **har** precis **opererat** min laterala menisk i vänster knä)*

(ii) to perform a surgical procedure, with one PERSON in the role of Patient, another PERSON in the role of Agent (Medical professional), DISEASE and BODY PART as possible core arguments

*e.g. <PERSON(Patient)> **opererades** <TIME> av <PERSON (Agent)> (Original sentence: Han **opererades** omedelbart av dr Piotr)*

*<PERSON (Patient)> som är <MEASURE> **har** både strålats och **opererats** för <DISEASE> (Original sentence: "min pappa som är 63 **har** både strålats och **opererats** för tonsillscancer")*

- **reflexive** *operera sig*: to give consent to have a surgical procedure performed with PERSON in the double role of semi-Agent and Experiencer and DISEASE

*e.g. <PERSON (Experiencer)> **har opererat mig** för <DISEASE> i <ANATOMY> som var <MEASURE> (Original sentence: Jag **har opererat mig** för malignt melanom i ryggen som var 1,2 mm)*

- **particle verb** with two sub-senses:

(i) to give consent to removing or implanting a body part or an implant with semi-Agent & Experiencer and ANATOMY or IMPLANT as possible scheme elements.

*e.g. <PERSON (semiAgent & Experiencer)> **opererade bort** <ANATOMY> för <TIME> (Original sentence: "Jag **opererade bort** blindtarmen för ganska exakt 36 timmar sedan")*

(ii) to perform a surgical procedure aiming at removing or implanting a body part or an implant with PERSON in role of Agent (medical professional), ANATOMY, IMPLANT and optionally with PERSON being a Donor as

possible scheme elements. IMPLANT and Donor have not been annotated in the examined corpus. (The tag IMPLANT will be reserved for an artefacts, since organic implants are tagged as ANATOMY.)
e.g. <PERSON (Agent)> ***opererade in*** *en p-stav i den kvinnliga <ANATOMY> (Original sentence: "Läkaren* ***opererade in*** *en p-stav i den kvinnliga patientens arm")*

This specification of scheme elements captures some prototypical scenarios for the verb *operera*. The schemes can undergo certain modifications resulting in null instantiation of scheme elements, which can be either constructional, definite or indefinite (Fillmore et al. 2003).

5 Semi-automatic Acquisition of Semantic Schemes

Semi-automatic acquisition of semantic schemes on the basis of an annotated corpus is far from a trivial task for verbs such as *operera*, mainly due to the fact that the human subject, when used in active form can correspond to different semantic roles, ranging from the agentive ones, e.g. Agent usually manifested by medical professionals to a semi-agentive in Experiencer role and non-agentive in the Patient role. The question remains whether there are explicit supportive cues to distinguish between those role instances and whether other roles can be semi-automatically tagged. Some proposals which might be worth testing with respect to role identification for the examined verbs are:

Agent: Medical professional
- lexical criterion: checking the list of lexical units naming medical professionals;
- presence of a prepositional phrase introduced by *av* followed by a scheme element PERSON in a sentence in passive voice;
- presence of another np in the same scheme labelled as PERSON (Patient).

Experiencer:
- presence of a noun annotated as PERSON in a scheme and an inalienable noun annotated with the label ANATOMY having either a definite form (*Jag opererade bort blindtarmen)* or preceded by a possessive pronoun referring to the subject (*Jag har precis opererat min laterala menisk […])*;
- reflexive use of the verb (*Jag har opererat mig för malignt melanom).*

Patient:
- presence of an explicit Agent in the same scheme;
- presence of an implicit Agent in the same scheme (passive voice);
- object in an active sentence or subject in the passive sentence annotated with the tag PERSON.

Anatomy:
- lexical criterion: checking an available sub-lexicon.

Disease:
- lexical criterion: checking an available sub-lexicon;
- syntactic cue: use of preposition *för* in construction *operera någon för DISEASE* (*cf.* English *operate on sb (for sth*))

For a preliminary listing of schemes for the analysed verb senses see appendix 2a.

6 Conclusions

The advantages of the pre-processing and the consequences for lexical annotation have been illustrated and we believe that given the results of our case studies, the described methodology represents a feasible way to proceed in order to aid the annotation of large textual samples. As advantages of lexical annotation, the following needs mentioning:

- relevant semantic schemes can be retrieved from medical corpora
- integrated layers of syntactic and semantic annotation support the acquisition of semantic roles and thus enhance text understanding
- the semantic schemes provide input for various NLP tasks
- semantically annotated nouns promote disambiguation of predicates
- access to semantic schemes can support classification of lexical units carrying related meaning (e.g. *operera bort, avlägsna, ta bort*)

The quantitative analysis of the examined corpus has shown that the importance of many linguistically optional scheme elements needs to be

reassessed when viewed from a medical pragmatic perspective. For example Time, Measure and Method provide relevant data for diagnosing patients' health condition. Another issue that may need special attention in future annotating tasks is that of tagging pronouns. It seems that these should not be tagged before anaphoric relations and their semantic roles have been established. This is particularly important for distinguishing between patients and health care providers. The figures in table 2 illustrate clearly the importance of identifying and annotating different entity types, particularly for the annotation of FrameNet non-core elements such as Time, Measure and Method, but also a strong indication of the frequency of important core elements such as Disease and Anatomy.

References

S. Abney. 1997. Part-of-speech tagging and partial parsing. *Corpus-Based Methods in Language and Speech Processing*. S. Young and G. Bloothooft (eds), 118–136. Kluwer AP.

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet Project. *Proc. of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL 1998)*. Montreal.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, and S. Pado. 2006. SALTO - A versatile multi-level annotation tool. *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

W.-C. Chou, R.T. Tsai, Y.-S. Su, W. Ku, T.Y. Sung, and W.-L. Hsu. 2006. A semi-automatic method for annotating a biomedical proposition bank. *Proc. of the Workshop on Frontiers in Linguistically Annotated Corpora.* 5–12. Sydney, Australia

K.B. Cohen and L. Hunter. 2006. A critical review of Pasbio's argument structures for biomedical verbs. *BMC Bioinformatics*

C.J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the NY Acad. of Sciences Conference on the Origin and Development of Lnguage and Speech*. Vol. 280.

C.J. Fillmore, C.S. Johnson, and M.R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3).

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

D. Gildea and M. Palmer. 2002. The necessity of parsing for predicate argument recognition. *Proc. of ACL 2002*, Philadelphia, PA.

M. Huang, X. Zhu, and M. Li. 2005. A hybrid method for relation extraction from biomedical literature. *Journal of Medical Informatics*.

R. Johansson and P. Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. *Proc. of Coling/ACL 2006*. Sydney, Australia

D. Kokkinakis. 2004. Reducing the effect of name explosion. *Proc. of the Beyond Named Entity Recognition, Semantic Labelling for NLP Tasks*. workshop at LREC. Lisbon, Portugal

D. Kokkinakis. 2006. Collection, encoding and linguistic processing of a Swedish medical corpus – The MEDLEX experience. *Proc. of the 5th LREC*. Italy.

D. Kokkinakis and S. Johansson Kokkinakis. 1999. A cascaded finite-state parser for syntactic analysis of Swedish. *Proc. of the 9th European Chapter of the Association of Computational Linguistics (EACL)*. Bergen, Norway.

E. König and W. Lezius. 2003. The TIGER language – A description language for syntax graphs, Formal definition. Technical report. Institut für Maschinelle Sprachverarbeitung, University of Stuttgart.

S. Pradhan, W. Ward., K. Hacioglu., J. Martin, and D. Jurafsky. 2004. Shallow semantic parsing using support vector machines. *Proc. of the Human Language Technology Conference/North American chapter of the ACL (HLT/NAACL)*, Boston, MA.

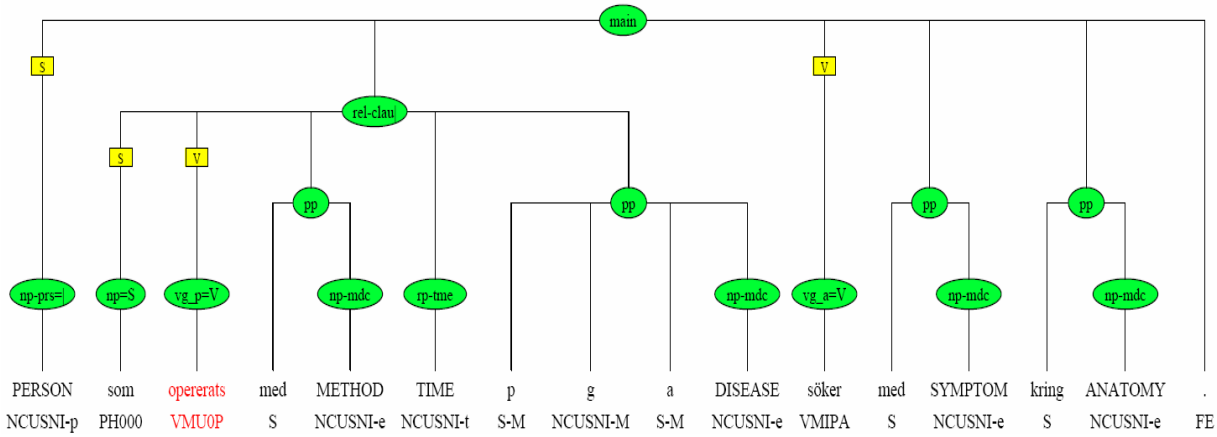
J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice.* Available from <http://framenet.icsi.berkeley.edu/book/book.pdf>

M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction. *Proc. of the 41st Annual Meeting of the Assoc. of Comp. Ling,* 8–15.

T. Wattarujeekrit, P.K. Shah, and N. Collier. 2004. Pasbio: Predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics* 2004, 5:155

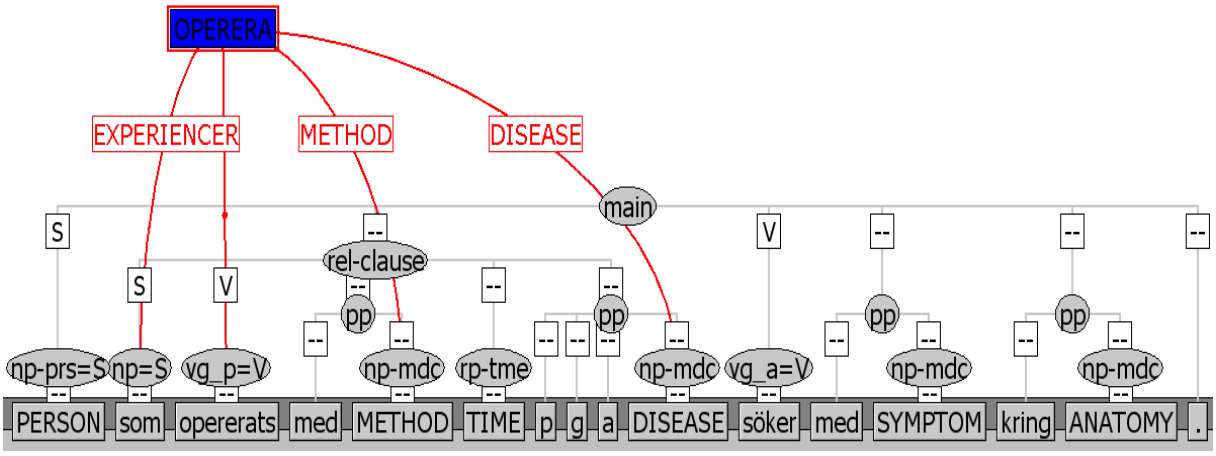
Appendix

1a



Syntactic analysis

1b



Role Assignment

1c

livskvalitetsstudie av <PERSON-GRP> som **opererats** för <DISEASE> är hämtade från journala
DISEASE> sedan <MEASURE> . /<PERSON> har **opererats** för <DISEASE> . /<PERSON> har **opererat**
opererats för <DISEASE> . /<PERSON> har **opererats** för <DISEASE> <TIME> . /<PERSON> har **op**
ats för <DISEASE> <TIME> . /<PERSON> har **opererats** för <DISEASE> två gånger , andra gånge
för <DISEASE> två gånger . /<PERSON> har **opererats** för en sk <DISEASE> <DISEASE> av comed
PERSON> har påvisat att <PERSON-GRP> som **opererats** för <SYMPTOM> eller <DISEASE> <TIME> h
. /<TIME> vet <PERSON> att min <PERSON> **opererats** för s.k. <DISEASE> och att man vid ope
illverkaren kan ätas av <PERSON-GRP> som **opererats** för <DISEASE> , och inte anses påverka
är man rullstolsbunden <TIME> . /Har man **opererats** genom <ANATOMY> behöver man oftast int
<ANATOMY> eller <ANATOMY> . /Om man har **opererats** genom <ANATOMY> brukar man behöva vara
ehandling . /Förutom de 71 <ANATOMY> som **opererats** hittade <PERSON-GRP> 231 <ANATOMY> med
under sin ST-utbildning . /Av de 48 som **opererats** har knappt hälften **opererats** en gång o
ASE> . /Studien visar tydligt att de som **opererats** har fått väsentligt bättre livskvalite
t att komma ihåg att de <PERSON-GRP> som **opererats** har inga eller mycket <SYMPTOM> från s
digare uppföljningar av <PERSON-GRP> som **opererats** har visat på nedslående resultat . /Me
D> , men betydelsen av detta hos dem som **opererats** har länge varit oklar och omtvistad .
skötas väl när en mekanisk <METHOD> har **opererats** in . /Metoden att operera ASD har funn
er operationen . /<PERSON> hade tidigare **opererats** i <ANATOMY> och kom till sjukhuset på
N> , <MEASURE> , är <PERSON-GRP> och har **opererats** i <ANATOMY> . /<PERSON> , <MEASURE> ,
llan <TIME> och upp <TIME> . /Om man har **opererats** i <ANATOMY> brukar man vårdas på sjukh
SE> ; risken ansågs för hög . /Något som **opererats** in i <ANATOMY> i diagnostiskt , terape
någon form , till exempel kärlklämmor , **opererats** in i <PERSON> , och särskilt om operat
n opererande enheten . /<PERSON-GRP> som **opererats** i <ANATOMY> är fortfarande bättre <TIM
ndersökning av <PERSON-GRP> som tidigare **opererats** i <ANATOMY> visade att <CHEMICAL> halv
E> skickas <PERSON-GRP> från <PLACE> som **opererats** i <PLACE> till <PLACE> för eftervård .
ats i <PLACE> . /<TIME> har <PERSON-GRP> **opererats** i <PLACE> för sina <SYMPTOM> i <ANATOM
<PERSON> egna <ANATOMY> har flyttats och **opererats** införbi förträngningen i <ANATOMY> . /

Semantic Concordance

Appendix

2a

Scheme: V *operera*	**Exempel**
PERSON(Agent) **V** PERSON(Patient)	Vi har opererat två patienter med Budd-Chiaris syndrom; Även kirurgen som opererat henne tog sig tid för att deltaga
PERSON(Agent) **V** (an instance of indefinite null instantiation)	I dagsläget opererar fyra urologer vid hans klinik; När läkarna opererar, suger slangarna blodceller genom lasern
PERSON(Agent) **V** METHOD	Roboten opererar med fyra armar
PERSON(Agent) **V** DISEASE	De opererar aldrig näsfrakturer
PERSON(Agent) **V in/ut** IMPLANT	Oftast opererar man in en mekanisk klaffprotes Risken för ett nytt benbrott finns alltid när man opererar ut metallimplantatet
PERSON(Agent) **V bort/ut** ANATOMY	Man opererar bort hela njuren,
PERSON(Agent)**V bort** ORGANISM	När man opererar en pinoidalcysta
PERSON(semi-Agent&Experiencer) **V** ANATOMY	Jag har precis opererat min laterala menisk i vänster knä
PERSON(semi-Agent&Experiencer) **V sig för** DISEASE	Jag har opererat mig för malignt melanom i ryggen

Schemas for the verb *operera*

2b

Frame	**Core frame elements**	**Non-core frame elements**
Medical_conditions	Ailment, Patient	Body_part, Cause, Degree, Name, Symptom
Experience_bodily_harm	Body part, Experiencer	Containing_event, Duration, Frequency, Injuring_entity, Iterations, Manner, Place, Severity, Time
Cure	Affliction, Body_part, Healer, Medication, Patient, Treatment	Degree, Duration, Manner, Motivation, Place, Purpose, Time
Health_response	Protagonist, Trigger	Body_part, Degree, Manner
Institutionaliztion	Authority, Facility, Patient	Affliction, Depictive, Duration_of_final state, Explanation, Manner, Means, Place, Purpose, Time
Recovery	Affliction, Body part, Patient,	Company, Degree, Manner, Means,
Medical_instruments	Instrument	Purpose
Medical_professionals	Professional	Affliction, Age, Body _system, Compensation, Contract_basis, Employer, Ethnicity, Origin, Place_of_employment, Rank, Type
Medical specialties	Specialty	Affliction, Body_system, Type
Observable_bodyparts	Body_part, Possessor	Attachment; Descriptor, Orientational_location, Subregion,
Placing	Agent, Cause, Theme, Goal	Area, Beneficiary, Cotheme, Degree, Depictive, Distance, Duration, Manner, Means, Path, Place, Purpose, Reason, Result, Source, Speed, Time
Removing	Agent, Cause, Source, Theme	Cotheme, Degree, Distance, Goal, Manner, Means, Path, Place, Result, Time, Vehicle

Medical frames in FrameNet

A Dependency-Based Conversion of PropBank

Susanne Ekeklint
Växjö University
sek@msi.vxu.se

Joakim Nivre
Växjö University and Uppsala University
nivre@msi.vxu.se

Abstract

As a prerequisite for the investigation of dependency-based methods for semantic role labeling, this paper describes the creation of a dependency-based version of the widely used PropBank, DepPropBank, and discusses some of the issues involved in the integration of syntactic and semantic dependency structures.

1 Introduction

The long-term goal of our research is to investigate the suitability of dependency-based representations for semantic role labeling (SRL). Our research also includes different ways of integrating semantic information into syntactic dependency structures. It has already been established that syntactic information is necessary for accurate SRL (Gildea and Palmer, 2002). It is however still an open issue which type of syntactic information should be used and how this information should be structured. The majority of published experiments on SRL are based on treebanks annotated with phrase structure. For the type of experiments that we wish to conduct, no suitable resource was available, so we decided to create one. In this paper we will therefore describe the creation of a dependency version of PropBank, called DepPropBank, and discuss some of the issues involved in the integration of syntactic and semantic dependency structures.

2 SRL and PropBank

The SRL that we consider is of the predicate-argument type and this type of semantic information can be used in order to improve quality in different natural language processing tasks, such as information retrieval, dialog management, translation or summarization. Typically, any application that needs to recognize entities answering to question words such as "Who", "When", and "Why" can benefit from this type of information. Figure 1 is an example of a sentence containing one predicate (*set*) and the arguments belonging to it. The SRL task can shortly be described as follows:

> Given a sentence the task consists of analyzing the propositions expressed by some target verbs of the sentence. In particular, for each target verb all the constituents in the sentence which fill a semantic role of the verb have to be recognized. It also includes determining which semantic role that each constituent has. (Carreras and Màrques, 2005)

Since it is important for us to be able to compare our experiments to previous work, we decided to create our data sets from PropBank (Palmer et al., 2005). PropBank is the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993), enriched with annotation of predicate-argument relations. PropBank is one of the most widely used resources for SRL experiments, popularized in particular by The CoNLL shared tasks in 2004 and 2005 (Carreras and Màrques, 2005), which have had a large impact on SRL and can be seen as representing the state of the art for this particular task. An annotation unit in PropBank is called a proposition and consists of a verb together with its semantic arguments, classified

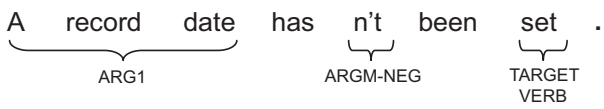


Figure 1: Sentence wsj02wsj_0202 from PropBank labeled with predicate argument-relations.

by numbered verb-specific roles or by general semantic modifier roles. The numbered verb-specific roles are ARG0-ARG5, where for example ARG0 in general corresponds to *agent* and ARG1 to *patient* or *theme*. The general semantic modifiers are adjuncts or functional labels that any verb may take optionally. There are 13 general semantic modifiers, e.g., ARGM-ADV for *general-purpose* and ARGM-NEG for *negation*. The roles are defined according to the role set for each verb, which defines the possible usage of each verb according to VerbNet (Levin, 1993). The PropBank data includes 44631 semantically annotated sentences, with an average of 2.53 propositions per sentence 3.21 arguments per proposition.

3 Dependency-Based SRL

A syntactic dependency graph is a labeled directed graph $G = (V, A_{syn})$, where V is a set of nodes, corresponding to the words of a sentence, and A_{syn} is a set of labeled directed arcs, representing syntactic dependency relations. The basic idea in dependency-based SRL is that we can construct an integrated syntactic-semantic representation by adding a second set A_{sem} of labeled arcs, representing semantic role relations, which gives us a multigraph $G = (V, A_{syn}, A_{sem})$, with two sets of labeled arcs defined on the same set of nodes. The SRL task can then be defined as the task of deriving A_{sem} given V and A_{syn}. In order to perform experiments based on PropBank, we therefore needed to convert the representations in the original Penn Treebank and PropBank to integrated syntactic-semantic dependency graphs. The result of this conversion is what we call DepPropBank.

4 DepPropBank

When designing the conversion from PropBank to DepPropBank we have had three different, partly conflicting requirements in mind:

1. We want to use the converted representations for machine learning experiments on dependency-based SRL, as described in the previous section. (Learnability)

2. We want to preserve the information in the original PropBank as precisely as possible. (Faithfulness)

3. We want to integrate syntactic and semantic relations as closely as possible. (Integration)

These requirements are not always compatible, and different trade-offs are possible. Therefore we have decided to create three different versions of DepPropBank, using three different models for integrating semantic information with syntactic dependency structures, investigating various degrees of tight and loose coupling in the integration. In formal terms, this amounts to three different algorithms for creating the set A_{sem} of semantic relations, given the set of nodes V, the set A_{syn} of syntactic relations, and the original PropBank annotation. The benefit of having three different versions is that we can empirically investigate the impact of different representational choices on SRL accuracy. We call the three different versions DepPropBank 1, 2, and 3.

However, before we could start to integrate the semantic information we needed to convert the syntactic phrase structures in the Penn Treebank to dependency structures. This was done using the freely available conversion program Penn2Malt.[1] This conversion is far from perfect but sufficiently precise for our current purposes. In the future we may instead decide to use the recently developed pennconverter (Johansson and Nugues, 2007),[2] which provides an improved conversion that, among other things, takes empty categories into account.

The next step was to relate the semantic annotation in PropBank to the phrase structure representations in the Penn Treebank. Figure 2 shows how the proposition for the target verb *set* from PropBank is integrated with the corresponding phrase structure from the Penn Treebank.

[1] http://w3.msi.vxu.se/∼nivre/research/Penn2Malt.html
[2] http://nlp.cs.lth.se/pennconverter/

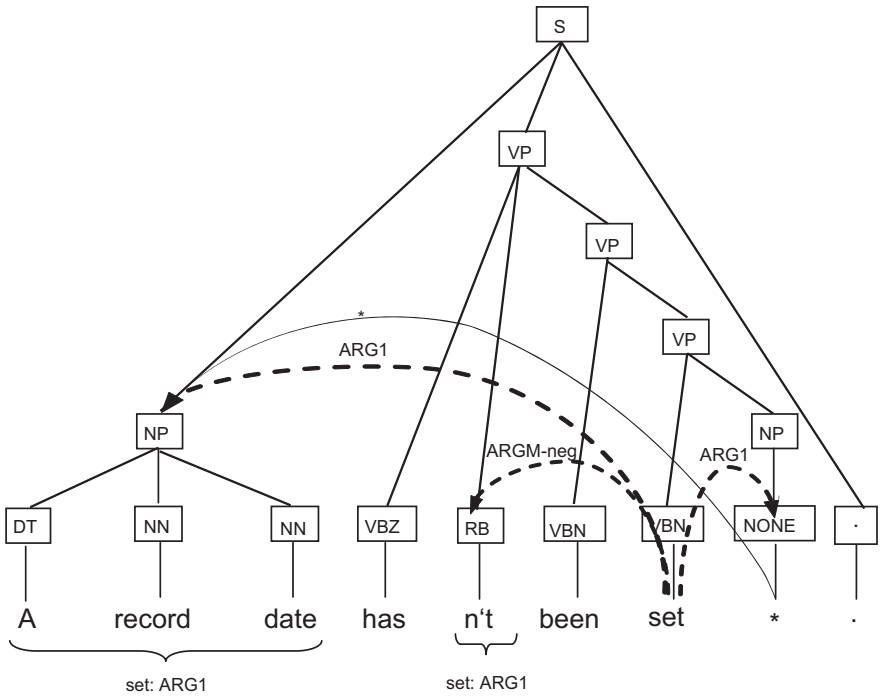


Figure 2: The phrase structure representation of sentence wsj02wsj_0202 in PropBank

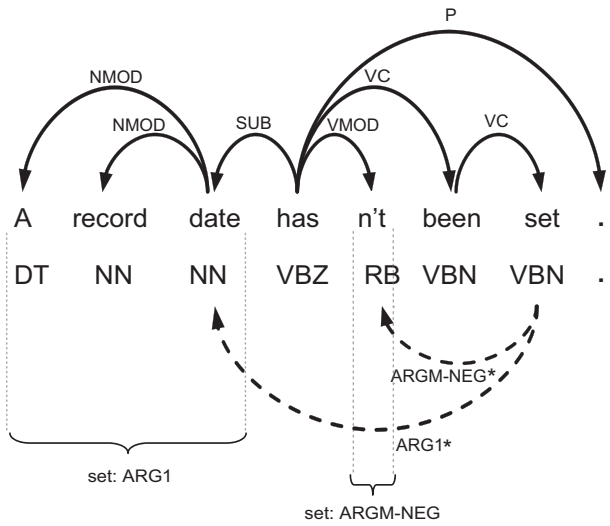


Figure 3: The dependency structure representation of sentence wsj02wsj_0202 in DepPropBank 1.

Since a dependency representation only contains terminal nodes (words), we needed to map these references to word sequences, also taking into account empty categories and co-indexation. The original PropBank annotation identifies predicates and arguments by referring to nodes in the syntactic annotation of the Penn TreeBank. An argument in the PropBank representation can be composed of several subtrees in the syntactic representation. We will refer to a sequence of words included in an argument as the *span* of that argument.

4.1 DepPropBank 1

Given that we have identified all the argument spans associated with a given predicate (and their semantic roles), we can extend the dependency graph generated by the syntactic conversion by adding arcs for semantic roles. In the first version of DepPropBank, this was done in the following way:

> Given an argument span s of predicate p with semantic role r:
>
> 1. For every word w within s that does not have its syntactic head within s, add an arc $p \xrightarrow{r^*} w$.
> 2. For every word w within s that has its syntactic head within s, assume that w belongs to the semantic spans of its syntactic head.

Figure 3 shows a dependency graph where the syntactic arcs in A_{syn}, drawn above the words, form a tree as usual, and where the semantic arcs in A_{sem} are represented by dotted arcs below the words. Note that the semantic arc labeled ARG1* only points to the syntactic head *date*, while the semantic argument span includes the whole syntactic subtree rooted at this node. We use the superscript $*$ on semantic arc labels to indicate that the argument relation extends transitively to syntactic descendants of the head.

Unfortunately, the first version of DepPropBank does not give an adequate representation of all the arguments in PropBank. The problem lies in the assumption that all syntactic dependents belong to the samt semantic spans as their head. This assumption holds for about 86% of all arguments in PropBank (given the current syntactic dependency conversion), but the remaining 14% require a more complex representation, where the internal semantic dependency structure of an argument does not necessarily coincide with its syntactic dependency structure. Figure 4 shows a sentence which has one correctly inherited syntactic subtree and one incorrect.

Looking at this result in a positive way, we can say that as many as 86% of the semantic subtrees have an exact match with the syntactic subtrees within their respective spans, in a representation where every semantic argument is represented by a single arc in A_{sem}. Experiments with this data set should therefore at least be interesting as a baseline for further experiments.

4.2 DepPropBank 2

The second version of DepPropBank 2 was created to solve the problem with the arguments that run outside the intended span. The semantic arcs in A_{sem} wee in this version simply added as follows:

> Given an argument span s of predicate p with semantic role r, add an arc $p \xrightarrow{r} w$ for every word w within s.

Figure 5 shows the same sentence fragment as figure 4, although this time with the representation of DepPropBank 2. Note the absence of the superscript $*$ on semantic role labels to indicate that each arc concerns only the word itself, not its syntactic descendants. The semantic representation has a very loose coupling to the syntactic structure in this version and the obvious drawback of version 2 is the flattening of the semantic structures. However, the representation has the advantage that there is always a single arc connecting each word in a semantic argument span to its predicate. Since there is an average of 2.5 propositions per sentence, of which several have partially or completely overlapping arguments, assigning hierarchical structures to semantic arguments would require a multigraph also for the semantic representation, where two nodes can be connected by more than one (semantic) arc. We could have solved this problem in several ways (for example by adding extra features to the labels and keeping the arcs as they were), but for machine learning experiments we found this particular representation promising.

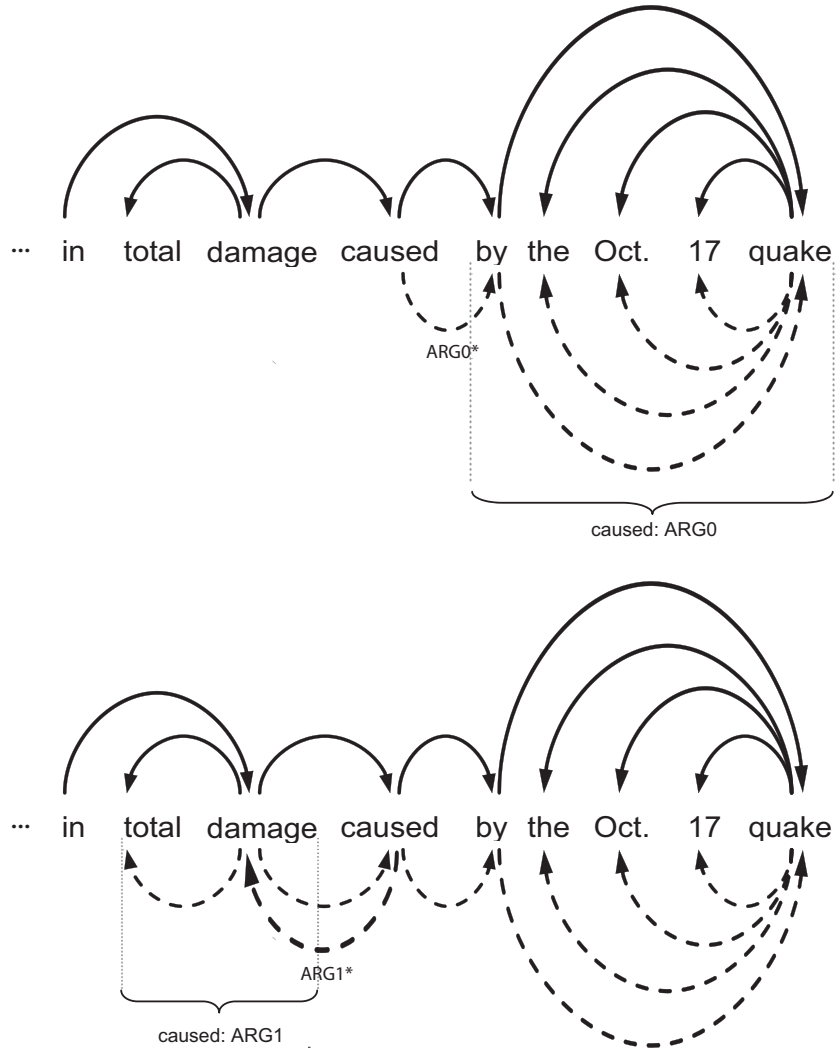


Figure 4: A good (top) and a bad (bottom) match between syntactic and semantic structure for arguments in DepPropBank 1.

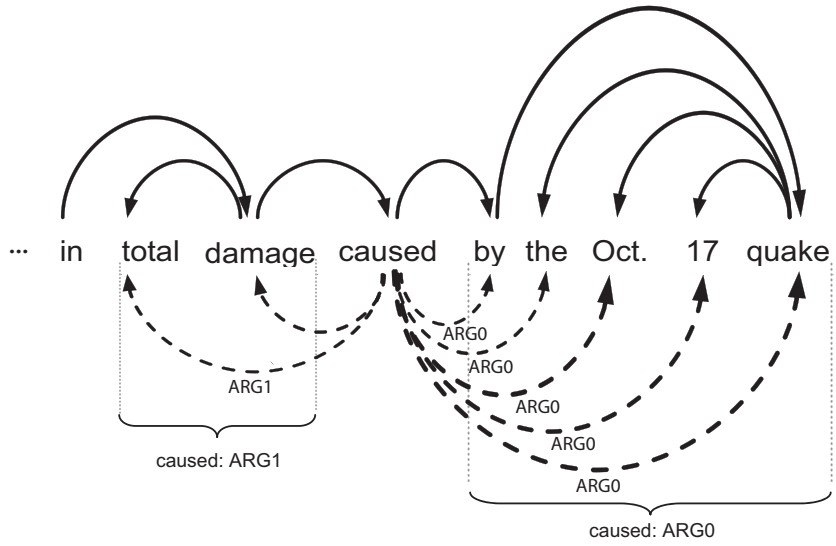


Figure 5: Flat semantic argument structure in DepPropBank 2.

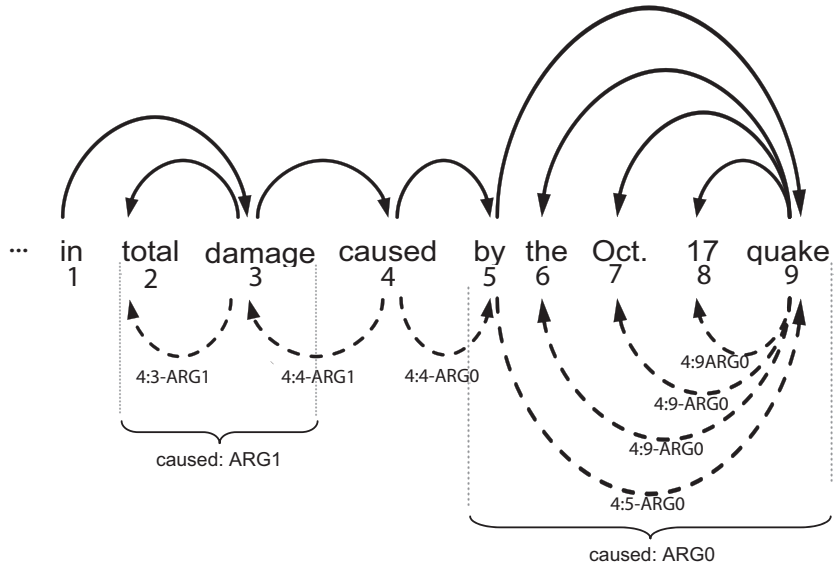


Figure 6: Hierarchical semantic argument structure in DepPropBank 3.

4.3 DepPropBank 3

Comparing DepPropBank 1 and 2 with respect to our three overall requirements, we can say that DepPropBank 1 maximizes syntactic-semantic integration (at the expense of faithfulness), while DepPropBank 2 maximizes faithfulness (at the expense of integration). From the point of view of learnability, both versions facilitate learning by minimizing path lengths in the semantic part of the graph (all paths being of length one), while DepPropBank 1 in addition minimizes the number of semantic arcs that need to be inferred (one arc per argument). In the third version, DepPropBank 3, the idea is to jointly maximize faithfulness and integration, possibly at the expense of learnability. The semantic arcs in A_{sem} were in this version added as follows:

> Given an argument span s of predicate p with semantic role r:
>
> 1. For each word w within s that does not have its syntactic head within s, add an arc $p \overset{i:i-r}{\rightarrow} w$, where i is the index (linear position) of p.
> 2. For each word w within s that has its syntactic head within s, add an arc $h \overset{i:j-r}{\rightarrow} w$, where h is the syntactic head of w, and i and j are the indices (linear positions) of p and h, respectively.

The advantage of this representation is that it has a strong integration of the semantic and syntactic structure without losing any of the information in the original annotation. The downside is the more complex graphs that we have to handle from a machine learning perspective. In fact, (V, A_{sem}) now needs to be a multi-graph, since it is possible to have two nodes connected by more than one arc. Moreover, the labels must encode the index of the predicate, which may be connected to a word by a path of arbitrary length. Figure 6 illustrates the more complex graphs of DepPropBank 3.

5 Conclusion

The three different versions of DepPropBank will allow us to empirically investigate the trade-off between integration, faithfulness and learnability in dependency-based SRL. Starting from the baseline of DepPropBank 1, which poses the simplest learning problem but where 14% of the arguments cannot be retrieved correctly, we can move on to the more faithful but also more complex representations in DepPropBank 2 and 3.

Since our data sets are derived from PropBank, we are also able to compare our results with the state of the art in SRL. In addition, we can investigate whether dependency-based representations give a better fit between argument spans and syntactic units than phrase structure representations. Finally, it is worth nothing that our models are applicable to languages that have treebanks annotated with dependency structure but not phrase structure, such as Czech (Böhmova et al., 2003) and Danish (Kromann, 2003), among others.

References

Alena Böhmová and Jan Hajič and Eva Hajičová and and Barbora Hladká. 2003. The Prague Dependency Treebank: A Three-Level Annotation Scenario In Anne Abeillé (ed.) *Treebanks: Building and Using Parsed Corpora*, Kluwer.

Xavier Carreras and Lluis Màrquez 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of CoNLL-2005*.

Dan Gildea and Martha Palmer 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of ACL-2002*.

Richard Johansson and Pierre Nugues 2007. Extended Constituency-to-Dependency Conversion for English. In *Proceedings of NODALIDA-2007*.

Matthias Trautner Kromann. 2003. The Danish Dependency Treebank and the DTAG Treebank Tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT)*.

Beth Levin 1993. *English Verb Classes and Alternations: A Preliminary Investigation* The University of Chicago Press

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics, 19*

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

Using WordNet to Extend FrameNet Coverage

Richard Johansson and **Pierre Nugues**
Department of Computer Science, Lund University, Sweden
{richard, pierre}@cs.lth.se

Abstract

We present two methods to address the problem of sparsity in the FrameNet lexical database. The first method is based on the idea that a word that belongs to a frame is "similar" to the other words in that frame. We measure the similarity using a WordNet-based variant of the Lesk metric. The second method uses the sequence of synsets in WordNet hypernym trees as feature vectors that can be used to train a classifier to determine whether a word belongs to a frame or not. The extended dictionary produced by the second method was used in a system for FrameNet-based semantic analysis and gave an improvement in recall. We believe that the methods are useful for bootstrapping FrameNets for new languages.

1 Introduction

Coverage is one of the main weaknesses of the current FrameNet lexical database; it lists only 10,197 lexical units, compared to 207,016 word–sense pairs in WordNet 3.0. This is an obstacle to fully automated frame-semantic analysis of unrestricted text.

This work addresses this weakness by using WordNet to bootstrap an extended dictionary. We report two approaches: first, a simple method that uses a similarity measure to find words that are related to the words in a given frame; second, a method based on classifiers for each frame that uses the synsets in the hypernym trees as features. The dictionary that results from the second method is three times as large as the original one, thus yielding an increased coverage for frame detection in open text.

Previous work that has used WordNet to extend FrameNet includes Burchardt et al. (2005), which applied a WSD system to tag FrameNet-annotated predicates with a WordNet sense. Hyponyms were then assumed to evoke the same frame. Shi and Mihalcea (2005) used VerbNet as a bridge between FrameNet and WordNet for verb targets, and their mapping was used by Honnibal and Hawker (2005) in a system that detected target words and assigned frames for verbs in open text.

1.1 Introduction to FrameNet and WordNet

FrameNet (Baker et al., 1998) is a medium-sized lexical database that lists descriptions of English words in Fillmore's paradigm of Frame Semantics (Fillmore, 1976). In this framework, the relations between predicates, or in FrameNet terminology, *target words*, and their arguments are described by means of *semantic frames*. A frame can intuitively be thought of as a template that defines a set of slots, *frame elements*, that represent parts of the conceptual structure and correspond to prototypical participants or properties. In Figure 1, the predicate *statements* and its arguments form a structure by means of the frame STATEMENT. Two of the slots of the frame are filled here: SPEAKER and TOPIC. The

As usual in these cases, [both parties]$_{\text{SPEAKER}}$ agreed to make no further **statements** [on the matter]$_{\text{TOPIC}}$.

Figure 1: Example sentence from FrameNet.

initial versions of FrameNet focused on describing situations and events, i.e. typically verbs and their nominalizations. Currently, however, FrameNet defines frames for a wider range of semantic relations, such as between nouns and their modifiers. The frames typically describe events, states, properties, or objects. Different senses for a word are represented in FrameNet by assigning different frames.

WordNet (Fellbaum, 1998) is a large dictionary whose smallest unit is the *synset*, i.e. an equivalence class of word senses under the synonymy relation. The synsets are organized hierarchically using the is-a relation.

2 The Average Similarity Method

Our first approach to improving the coverage, the Average Similarity method, was based on the intuition that the words belonging to the same frame frame show a high degree of "relatedness." To find new lexical units, we look for lemmas that have a high average relatedness to the words in the frame according to some measure. The measure used in this work was a generalized version of the Lesk measure implemented in the WordNet::Similarity library (Pedersen et al., 2004). The Similarity package includes many measures, but only four of them can be used for words having different parts of speech: Hirst & St-Onge, Generalized Lesk, Gloss Vector, and Pairwise Gloss Vector. We used the Lesk measure because it was faster than the other measures. Small-scale experiments suggested that the other three measures would have resulted in similar or inferior performance.

For a given lemma l, we measured the relatedness $\mathrm{sim}_F(l)$ to a given frame F by averaging the maximal relatedness, in a given similarity measure sim, over each sense pair for each lemma λ listed in F:

$$\mathrm{sim}_F(l) = \frac{1}{|F|} \sum_{\lambda \in F} \max_{\substack{s \in \mathrm{senses}(l) \\ \sigma \in \mathrm{senses}(\lambda)}} \mathrm{sim}(s, \sigma)$$

If the average relatedness was above a given threshold, the word was assumed to belong to the frame.

For instance, for the word *careen*, the Lesk similarity to 50 randomly selected words in the SELF_MOTION frame ranged from 2 to 181, and the average was 43.08. For the word *drink*, which does not belong to SELF_MOTION, the similarity ranged from 1 to 45, and the average was 13.63. How the selection of the threshold affects precision and recall is shown in Section 4.1.

3 Hypernym Tree Classification

In the second method, Hypernym Tree Classification, we used machine learning to train a classifier for each frame, which decides whether a given word belongs to that frame or not. We designed a feature representation for each lemma in WordNet, which uses the sequence of unique identifiers ("synset offset") for each synset in its hypernym tree.

We experimented with three ways to construct the feature representation:

```
Sense 1 (1 example)
{01924882} stagger, reel, keel, lurch, swag, careen
   => {01904930} walk
     => {01835496} travel, go, move, locomote

Sense 2 (0 examples)
{01884974} careen, wobble, shift, tilt
   => {01831531} move


1924882:0.67 1904930:0.67 1835496:0.67
1884974:0.33 1831531:0.33
```

Figure 2: WordNet output for the word *careen*, and the resulting weighted feature vector

First sense only. In this representation, the synsets in the hypernym tree of the first sense was used.

All senses. Here, we used the synsets of all senses.

Weighted senses. In the final representation, all synset were used, but weighted with respect to their relative frequency in SemCor. We added 1 to every frequency count.

Figure 2 shows the WordNet output for the word *careen* and the corresponding sense-weighted feature representation.

Using these feature representations, we trained an SVM classifier for each frame that tells whether a lemma belongs to that frame or not. We used the LIBSVM library (Chang and Lin, 2001) to train the classifiers.

4 Evaluation

4.1 Precision and Recall for SELF_MOTION

To compare the two methods, we evaluated their respective performance on the SELF_MOTION frame. We selected a training set consisting of 2,835 lemmas, where 50 of these were listed in FrameNet as belonging to SELF_MOTION. As a test set, we used the remaining 87 positive and 4,846 negative examples. Both methods support precision/recall tuning: in the Average Similarity method, the threshold can be moved, and in the Hypernym Tree Classification method, we can set a threshold on the probability output from LIBSVM. Figure 3 shows a precision/recall plot for the two methods obtained by varying the thresholds.

The figures confirm the basic hypothesis that words in the same frame are generally more related,

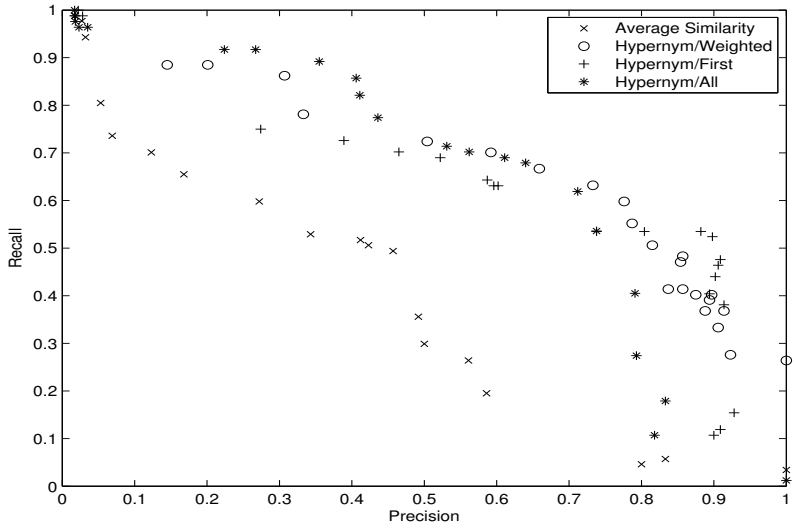


Figure 3: Precision/recall plot for the SELF_MOTION frame.

but the Average Similarity method is still not as precise as the Hypernym Tree Classification method, which is also much faster. Of the hypernym tree representation methods, the difference is small between first-sense and weighted-senses encodings, although the latter has higher recall in some ranges. The all-senses encoding generally has lower precision. We used the Hypernym Tree method with weighted-senses encoding in the remaining experiments.

4.2 All Frames

We also evaluated the performance for all frames. Using the Hypernym Tree Classification method with frequency-weighted feature vectors, we selected 7,000 noun, verb, and adjective lemmas in FrameNet as a training set and the remaining 1,175 as the test set – WordNet does not describe prepositions, and has no hypernym trees for adverbs. We set the threshold for LIBSVM's probability output to 50%. When evaluting on the test set, the system achieved a precision of 0.788 and a recall of 0.314. This can be compared to the result for from the previous section for the same threshold: precision 0.787 and recall 0.552.

4.3 Dictionary Inspection

By applying the hypernym tree classifiers on a list of lemmas, the FrameNet dictionary could be extended by 18,372 lexical units. If we assume a Zipf distribution and that the lexical units already in FrameNet are the most common ones, this would increase the coverage by up to 9%.

We roughly estimated the precision to 70% by manually inspecting 100 randomly selected words in the extended dictionary, which is consistent with the result in the previous section. The quality seems to be higher for those frames that correspond to one or a few WordNet synsets (and their subtrees). For instance, for the frame MEDICAL_CONDITION, we can add the complete subtree of the synset *pathological state*, resulting in 641 new lemmas referring to all sorts of diseases. In addition, the strategy also works well for motion verbs (which often exhibit complex patterns of polysemy): 137 lemmas could be added to the SELF_MOTION frame. Examples of frames with frequent errors are LEADERSHIP, which includes many insects (probably because the most frequent sense of *queen* is the queen insect), and FOOD, which included many chemical substances as well as inedible plants and animals.

4.4 Open Text

We used the extended dictionary in the Semeval-2007 task on Frame-semantic Structure Extraction (Baker, 2007). A part of the task was to find target words in open text and correctly assign them frames.

Our system (Johansson and Nugues, 2007) was evaluated on three short texts. In the test set, the new lexical units account for 53 out of the 808 target words our system detected (6.5% – this is roughly consistent with the 9% hypothesis in the previous section).

Table 1 shows the results for frame detection averaged over the three test texts. The table shows exact and approximate precision and recall, where the approximate results give partial credit to assigned frames that are closely related to the gold-standard frame. We see that the extended dictionary increases the recall – especially for the approximate case – while slightly lowering the precision.

Table 1: Results for frame detection.

	Original	**Extended**
Exact P	0.703	0.688
Exact R	0.504	0.528
Approx. P	0.767	0.758
Approx. R	0.550	0.581

5 Conclusion and Future Work

We have described two fully automatic methods to add new units to the FrameNet lexical database. The enlarged dictionary gave us increased recall in an experiment in detection of target words in open text. Both methods support tuning of precision versus recall, which makes it easy to adapt to applications: while most NLP applications will probably favor a high F-measure, other applications such as lexicographical tools may require a high precision.

While the simple method based on SVM classification worked better than those based on similarity measures, we think that the approaches could probably be merged, for instance by training a classifier that uses the similarity scores as features. Also, since the words in a frame may form disjoint clusters of related words, the similarity-based methods could try to measure the similarity to a subset of a frame rather than the complete frame. In addition to the WordNet-based similarity measures, distribution-based measures could possibly also be used.

More generally, we think that much could be done to link WordNet and FrameNet in a more explicit way, i.e. to add WordNet sense identifiers to FrameNet lexical units. The work of Shi and Mihalcea (2005) is an important first step, but so far only for verbs. Burchardt et al. (2005) used a WSD system to annotate FrameNet-annotated predicates with WordNet senses, but given the current state of the art in WSD, we think that this will not give very high-quality annotation. Possibly, we could try to find the senses that maximize internal relatedness in the frames, although this optimization problem is probably intractable.

We also think that the methods can be used in other languages. If there is a FrameNet with a set of seed examples for each frame, and if a WordNet or a similar electronic dictionary is available, both methods should be applicable without much effort.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING-ACL'98*.

Collin Baker. 2007. SemEval task 19: Frame semantic structure extraction. In *Proceedings of SemEval-2007*, forthcoming.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet detour to FrameNet. In *Proceedings of the GLDV 2005 workshop GermaNet II*, Bonn, Germany.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language*, 280:20–32.

Matthew Honnibal and Tobias Hawker. 2005. Identifying FrameNet frames for verbs from a real-text corpus. In *Australasian Language Technology Workshop 2005*.

Richard Johansson and Pierre Nugues. 2007. Semantic structure extraction using nonprojective dependency tress. In *Proceedings of SemEval-2007*. To appear.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity – measuring the relatedness of concepts. In *Proceedings of NAACL-04*.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet, and WordNet for robust semantic parsing. In *Proceedings of CICLing 2005*.

Building a Large Lexicon of Complex Valency Frames

Karel Pala, Aleš Horák
Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
{pala,hales}@fi.muni.cz

Abstract

This paper describes the process of building and using a new comprehensive lexicon of Czech verb valency frames based on *complex valency frames*. The main features of the lexicon entries are designed to bring important semantic information to computer processing of predicate constructions in running texts. The most notable features include two-level semantic labels with linkage to the Princeton and EuroWordNet hierarchy and surface verb frame patterns used for automatic syntactic analysis. Some implications for other languages, particularly English, Bulgarian and Romanian, are reported.

1 Introduction

Semantic role annotation is usually based on the appropriate inventories of labels for semantic roles (deep cases, arguments of verbs, functors, actants) describing argument predicate structure of verbs. It can be observed that the different inventories are exploited in different projects (e.g. Vallex (Stranakova-Lopatkova and Zabokrtsky, 2002), VerbNet (Kipper et al., 2000), FrameNet (Fillmore et al., 2006), Salsa (Boas et al., 2006), CPA (Hanks, 2004), VerbaLex (Hlaváčková and Horák, 2005)).

With regard to the various inventories a question has to be asked: how adequately they describe semantics of the empirical lexical data as we can find them in corpora? From this point of view it can be seen that some of the inventories are more syntactic than semantic (e.g. Vallex 1.0). If we are to build verb frames with the goal to describe real semantics of the verbs then we should go 'deeper'. Take, e.g. verbs like *drink* or *eat*, – it is obvious that the role PATIENT that is typically used with them labels cognitively different entities – BEVERAGES with *drink* and FOOD with *eat*. If we consider verbs like *see* or *hear* we can observe similar differences not mentioning the fact that one can see anything. Then the role PATIENT has to be regarded as mainly syntagmatic though using subcategorization features can improve the situation, however, usually they are not exploited in other lexicons (e.g. in Vallex 1.0). If we are not able to discriminate the indicated semantic distinctions the use of the frames with such labels in realistic applications may not lead to convincing and reliable results.

These considerations led us to the design of the inventory of two-level labels which are presently exploited for annotating semantic roles in Czech verb valency frames in lexical database VerbaLex containing now approx. 11 000 Czech verbs.

1.1 Thematic Roles and Semantic types

A question may be asked what is the distinction between "shallow" roles such as AGENT or PATIENT and "deep" roles such as SUBS(food:1), as we use it in VerbaLex, see below. We already hinted that "shallow" roles seem to be very similar to syntagmatic functions. At the same time it should be obvious that information that a person functions as an agent who performs an action is not only syntagmatic. That was the main reason why we included them in our list of the roles. We do not think that SUBS(food:1) is a special case of of the deep role, rather, we would like to speak about a two-level role consisting of the ontological part, i.e. SUBS(tance), and the subcategorization feature part,

Figure 1: An example of a Complex Valency Frames for the verbs klesnout:1, klesat:1, padnout:1, padat:1, snést se:1, snášet se:1 (descend:1, fall:2, go down:1, come down:1).

```
who_nom*AGENT(human:1|animal:1) <eat:1/jíst:1> what_acc*SUBS(food:1)
     withwhat_ins*INS(cutlery:2)
who_nom*AGENT(human:1|animal:1|institution:1) <see:1/vidět:1> what_acc*ANY(anything:1)
who_nom*AGENT(human:1|animal:1) <hear:1/slyšet:1> what_acc|koho4*PHEN(sound:1)
     how*MAN(manner:1).
```

Figure 2: Translation of Czech CVFs to English.

e.g. beverage:1 which is also a literal in PWN 2.0 that can be reached by traversing the respective hyperonymy/hyponymy tree.

In the Hanks' and Pustejovsky's Pattern Dictionary (cf. (Hanks, 2004) and also (Hanks et al., 2007)) a distinction is made between semantic roles and semantic types: "the semantic type is an intrinsic attribute of a noun, while a semantic role has the attribute thrust upon it by the context." Also lexical sets are distinguished which are "clusters of words that activate the same sense of a verb and have something in common semantically."

Introduction of the mentioned notions is certainly very inspiring in our context, however, we think that at the moment the quoted 'definitions' as they stand do not seem to be very operational, they are certainly not formal enough for computational purposes. What is needed are the lists of the semantic roles and types but they are being created gradually along with building the necessary ontology. Thus for time being we have to stick to our two-level roles as they are, that are partly based on the TOP Ontology as used in EuroWordNet project (Vossen, 1998). For semantic roles and types Brandeis Shallow Ontology ((Pustejovsky et al., 2006)) has been used but it is not regarded a final solution at the moment. (Examples of the semantic roles and types can be found in the papers quoted above.)

2 VerbaLex and Complex Valency Frames

The design of VerbaLex verb valency lexicon was driven mainly by the requirement to describe the verb frame (VF) features in a computer readable form suitable for syntactic and semantic analysis. After reviewing actual verb frame repositories, we have developed *Complex Valency Frames* (CVFs) that contain:

- morphological and syntactic features of constituents
- two-level semantic roles
- links to PWN and Czech WordNet hyper/hyponymic (H/H) hierarchy
- differentiation of animate/inanimate constituents

produce, make, create – create or manufacture a man-made product
BG: {proizveždam} njakoj*AG(person:1)| neščo*ACT(plant:1)= neščo*OBJ(artifact:1)
CZ: {vyrábět, vyrobit} kdo*AG(person:1)| co*ACT(plant:1) = co*OBJ(artifact:1)
uproot, eradicate, extirpate, exterminate – destroy completely, as if down to the roots; "the vestiges of political democracy were soon uprooted"
BG: {izkorenjavam, premachvam} njakoj*AG(person:1)| neščo*AG(institution:2)= neščo*ATTR(evil:3)|*EVEN(terrorism:1)
CZ: {vykořenit, vyhladit, zlikvidovat} kdo*AG(person:1)|co*AG(institution:2) = co*ATTR(evil:3)|*EVEN(terrorism:1)
carry, pack, take – have with oneself; have on one's person
BG: {nosja, vzimam} njakoj*AG(person:1)= neščo*OBJ(object:1)
CZ: {vzít si s sebou, brát si s sebou, mít s sebou, mít u sebe} kdo*AG(person:1)= co*OBJ(object:1)

Figure 3: Common verb frame examples for Czech and Bulgarian

- default verb position
- verb frames linked to verb senses
- VerbNet classes of verbs.

An example of a CVF is displayed in the Figure 1.

3 Role Annotation and EWN Top Ontology

Presently, our inventory contains the general or ontological labels selected from the EuroWordNet Top Ontology (EWN TO), with some modifications, and the 2^{nd}-level subcategorization labels taken mainly from the Set of Base Concepts introduced in (EuroWordNet Project, 1999). The 2^{nd}-level labels (approx. 200) selected from the Set of Base Concepts (BCs) are more concrete and they can be viewed as subcategorization features specifying the ontological labels coming from EWN TO. The motivation for this choice is based on the fact that WordNet has a hierarchical structure which covers about 110 000 English lexical units (synsets). It is then possible to use general labels corresponding to selected top and middle nodes and go down the hyperonymy/hyponymy (H/H) tree until the particular synset is found or matched. This allows us to see what is the semantic structure of the analyzed sentences using their respective valency frames. The nodes that we have to traverse when going down the H/H tree at the same time form a sequence of the semantic features which characterize meaning of the lexical unit fitting into a particular valency frame. These sequences can be interpreted as quite detailed selectional restrictions.

The two-level labels contain ontological labels taken from EWN TO (about 40) that include roles like AGENT, PATIENT, INSTRUMENT, ADDRESSEE, SUBSTANCE, COMMUNICATION, ARTIFACT at the 1^{st} level. The 2^{nd}-level labels that are combined with them are literals from PWN 2.0 together with their sense number.

The notation allows us to handle basic metaphors as well. An example of CVFs for *drink/pít* may roughly take the form:

```
who_nom*AGENT(human:1|animal:1)
  <drink:1/pít:1>
  what_acc*SUBS(beverage:1)
```

4 Multilingual Aspects of CVFs – can CVFs be Universal?

We have started building VerbaLex database during the EU project Balkanet (Balkanet Project, 2002) when about 1500 Czech verb valency frames were included in Czech WordNet. They were linked to English and other languages within Balkanet through the Interlingual Index (ILI). In the Balkanet project an experiment took place in which CVFs developed for Czech verbs have been linked to the corresponding verbs of Bulgarian and Romanian (Koeva, 2004).

While the experience with Czech CVFs for Bulgarian and Romanian is positive (see below the Section 4.1), and the result can be generalized also for other Slavonic languages like Slovak or Polish, the question remains whether CVFs developed for Czech can be applied to English as well. If we exploit ILI and have look at the VFs for Czech/English

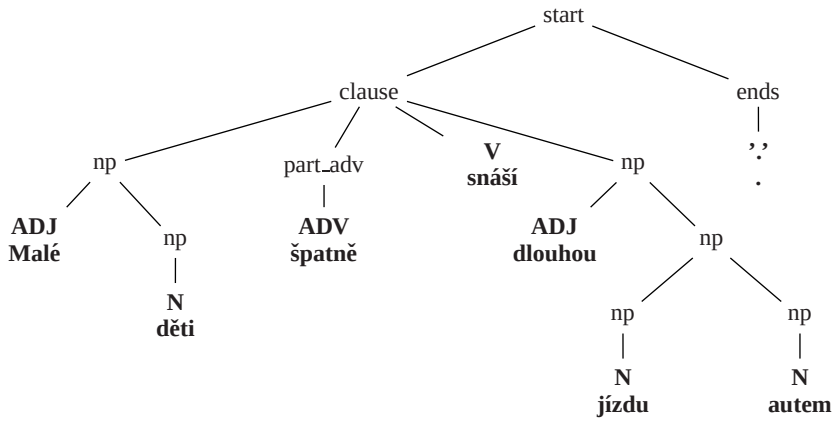


Figure 4: Syntactic tree of an example input sentence "`Malé děti špatně snáší dlouhou jízdu autem.`" (Small children badly withstand long journey by car.)

verbs like *pít/drink, jíst/eat* and apply them to their English translation equivalents we come to the conclusion that the Czech deep valencies certainly can describe their semantics. This conclusion is based on the simple assumption that we have the correct translation equivalents at our disposal. VerbaLex is incorporated into Czech WordNet and through ILI also to PWN 2.0, thus we have the necessary translation pairs at hand. This also can be applied for other WordNets linked to PWN. If the principle of translatability holds it means that the deep valencies developed for Czech can be reasonably exploited also for English (see the Figure 2).

In our view, the roles designed originally for the Czech verbs can serve for the corresponding English equivalents as well.

4.1 Bulgarian example

The enrichment of Bulgarian WordNet with verb valency frames was initiated by the experiments with Czech WordNet (CzWN) which already contained approx. 1500 valency frames (cf. (Koeva and others, June 2004)). Since both languages (Czech and Bulgarian) are Slavonic we assumed that a relatively great part of the verbs should realize their valency in the same way. The examples of Bulgarian and Czech valency frames in the Figure 3 show that this assumption has been justified (English equivalents come from PWN 1.7).

The construction of the valency frames of the Bulgarian verbs was performed in two stages:

1. Construction of the frames for those Bulgarian verb synsets that have corresponding (via Interlingual Index number) verb synsets in the CzWN and in addition these CzWN synsets are provided with already developed frames.

2. Creation of frames for verb synsets without analogues in the CzWN. The frames for more than 500 Bulgarian verb synsets have been created and the overall number of added frames was higher than 700. About 25% of the Bulgarian verb valency frames completely coincide with the Czech ones.

Similar results have been obtained also for Romanian where a good agreement was observed on the semantic level but the surface valencies had to be re-processed, Czech and Romanian are morphologically different.

In our view these experiments are convincing enough and they show sufficiently that it is not necessary to create the valency frames for the individual languages separately.

4.2 Levin's Classes and Czech Verbs

We have created semantic classes of Czech verbs that are inspired by Levin's classes (Levin, 1993) and VerbNet classes (Kipper et al., 2000). Since Czech is a highly inflectional language the patterns of alternation typical for English cannot be straightforwardly applied – Czech verbs require noun phrases in the morphological cases (there are 7 of them both in singular and plural). However, classes similar to Levin's can be constructed for

```
verb_rule_schema:  3 nterms, '#2'
   nterm 1:  k1gNnPc1
   nterm 2:  k5eAp3nPtPmIaI
   nterm 3:  k1gFnSc4
   group 1:  0, 2, +npnl -> .{ left_modif } np .  k1gMnSc1 ''malé děti''
   group 2:  2, 3, +ADV -> .'špatně' .  k6xMeAd1
   group 3:  4, 7, +npnl -> .{ left_modif } np .  k1gFnSc4 ''dlouhou jízdu autem''
possible subjects:  #1
Clause valency list:
     snášet <v>-#2:(1)hH#1:(0)hPTc1-#3:(2)hPTc4
     snášet(0) <v>#1:(1)hH-#2:(2)hPTc4
Verb valency list:
     snášet <v>#2:hH-#1:hPTc4
     snášet <v>#1:hPTc4
Matched valency list:
     snášet(0) <v>#2:(1)hH-#1:(2)hPTc4
```

Figure 5: The output of the verb frame extraction algorithm during the example sentence analysis.

Czech verbs as well but they have to be based only on the semantics of the verb classes. Before the starting the VerbaLex project we had compiled a Czech-English dictionary with Levin's 50 semantic classes and their Czech equivalents containing approx. 3000 Czech verbs.

In VerbaLex project we went further and linked Czech verbs with the verb classes as they are used in VerbNet – they are also based on Levin's classification extending it to almost 400 classes. This means that for each Czech verb in VerbaLex we mark the VerbNet semantic class a verb belongs to. We consider this information useful though it is known (according to our knowledge at least) that Levin's classes have not been extensively confronted with any corpus data. This certainly makes them less reliable.

The basic assumption in this respect is that the semantic classes of verbs should be helpful in checking the consistency of the inventory of semantic roles since in one class we can expect the roles specific only for that class. For example, with verbs of clothing the role like GARMENT and its respective subcategorizations can be reliably predicted, similarly it should work for other verb classes, such as verbs of eating, drinking, wearing, emotional states, weather and others. In the close future we plan to compare VerbNet semantic classes with the classes that we expect to obtain by sorting our valency frames according to the roles they occur with.

5 Application in Syntactic Analysis

We are currently testing the application in our syntactic analyzer `synt` that is designed for parsing real-text sentences. The verb frame extraction (VFE) process in `synt` is controlled by the metagrammar semantic actions. The parser builds a forest of values[1] to represent a result of the application of contextual constraints. The VFE actions are then executed on a different level (Horák and Kadlec, 2005) than the "usual" actions, which allows us to apply VFE actions on the whole forest of values.

If the analyzed verb has a corresponding entry in VerbaLex, we try to match the extracted frame with frames in the lexicon. When checking the valencies with VerbaLex, the dependence on the surface order is discharged. Before the system confronts the actual verb valencies from the input sentence with the list of valency frames found in the lexicon, all the valency expressions are reordered. By using the standard ordering of participants, the valency frames can be handled as sets independent on the current position of verb arguments. However, since VerbaLex contains an information about the *usual* verb position within the frame, we promote the standard ordering with increasing or decreasing the respective derivation tree probability.

The system processing can be presented on an example sentence – see the syntactic tree in the Figure 4 and the textual output of the part of the system

[1] a DAG (directed acyclic graph) structure that corresponds to the resulting chart structure supplemented with values computed during the semantic actions like feature agreement tests or verb frame extraction

that works on the VFE algorithm in the Figure 5. The system first identifies the verb rule constituents (`nterm`s), then the corresponding groups, i.e. the actual sentence constituents that will play the role as verb frame arguments, are extracted from the forest of values. Groups usually do not correspond to nterms one-to-one, since they are stored within nonterminals deeper in the forest and not directly in the verb rule. This part of the VFE algorithm has unfortunately exponential time complexity, however, for common sentences the depth of the verb frame constituents is not more than three levels, so the actual running times are usually within fractions of seconds. After the identification of the groups, the algorithm looks for possible subjects – this is not as easy as it may look at the first sight, since the sentence subject can be expressed not only by a noun phrase in nominative (which is the most frequent option in Czech), but also by e.g. prepositional phrase or verb infinitive. If no possible subject is found, the algorithm supplies a pronoun for an inexplicit subject with the gender corresponding to the verb. The Clause valency list displays all possible combinations of the translations of the verb arguments found into verb frame patterns. This list is then intersected with the list of lexicon entries for the verb to obtain the Matched valency list as a result of the VFE algorithm.

The effectiveness of the syntactic analysis with the VFE algorithm was measured on approximately 4.000 Czech corpus sentences with the median of 15 words per sentence and the Clause valency list contained 11 possible verb frames with the running time of 0.07 seconds per sentence.

6 Conclusions

In the paper we report on the building the lexical database of Czech verbs VerbaLex with their surface (morphological) and deep (semantic) valencies. For labeling the roles in the valency frames we have developed a list (ontology) of the two-level labels which at the moment contains approx. 40 'ontological' roles and 200 subcategorization features represented by the literals taken from Princeton WordNet 2.0. At present VerbaLex contains approx. 11 000 Czech verbs with 28 000 frames. We also mention some multilingual implications and show how the CVFs can be exploited in syntactic analysis of Czech.

Acknowledgments

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the project 201/05/2781.

References

Hans C. Boas, Elias Ponvert, Mario Guajardo, and Sumeet Rao. 2006. The current status of German FrameNet. In *SALSA workshop at the University of the Saarland*, Saarbrucken, Germany.

C.J. Fillmore, C.F. Baker, and H. Sato. 2006. Framenet as a 'net'. In *Proceedings of Language Resources and Evaluation Conference (LREC 04)*, volume vol. 4, 1091-1094, Lisbon. ELRA.

Patrick Hanks, Karel Pala, and Pavel Rychlý. 2007. Towards an empirically well-founded semantic ontology for NLP. In *Workshop on Generative Lexicon*, Paris, France. in print.

Patrick Hanks. 2004. Corpus pattern analysis. In *Proceedings of the Eleventh EURALEX International Congress*, Lorient, France. Universite de Bretagne-Sud.

Dana Hlaváčková and Aleš Horák. 2005. Verbalex – new comprehensive lexicon of verb valencies for czech. In *Proceedings of the Slovko Conference*, Bratislava, Slovakia.

Aleš Horák and Vladimír Kadlec. 2005. New meta-grammar constructs in Czech language parser synt. In *Proceedings of Text, Speech and Dialogue 2005*, pages 85–92, Karlovy Vary, Czech Republic. Springer-Verlag.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class based construction of a verb lexicon. In *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin TX.

S. Koeva et al. June 2004. Restructuring wordnets for the balkan languages, design and development of a multilingual balkan wordnet balkanet. Technical Report Deliverable 8.1, IST-2000-29388.

S. Koeva. 2004. Bulgarian VerbNet. Technical Report part of Deliverable D 8.1, EU project Balkanet.

Beth Levin, editor. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

J. Pustejovsky, C. Havasi, R. Sauri, and P. Hanks. 2006. Towards a generative lexical resource: The brandeis semantic ontology. In *Language Resources and Evaluation Conference, LREC 2006*.

M. Stranakova-Lopatkova and Z. Zabokrtsky. 2002. Valency dictionary of czech verbs: Complex tectogrammatical annotation. In *LREC2002, Proceedings*, volume III, pages 949–956. ELRA.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

Translational Equivalence and Cross-lingual Parallelism: The Case of FrameNet Frames

Sebastian Padó
Computational Linguistics
Saarland University
Saarbrücken, Germany
pado@coli.uni-saarland.de

Abstract

Annotation projection is a strategy for the cross-lingual transfer of annotations which can be used to bootstrap linguistic resources for low-density languages, such as role-semantic databases similar to FrameNet.

In this paper, we investigate the main assumption underlying annotation projection, *cross-lingual parallelism*, which states that annotation is parallel across languages. Concentrating on the level of *frames*, we provide a qualitative and quantitative characterisation of the relationship between translation and cross-lingual parallelism on the basis of a trilingual English–French–German corpus. We link frame (non)-parallelism to different kinds of *translational shifts* and show that a simple heuristic can detect the majority of such shifts.

1 Introduction

Recent work in computational linguistics suggests that many applications could benefit from a representation of text on the level of *predicate-argument structure* which abstracts away from idiosyncrasies of the text's surface structure. A promising descriptive framework for predicate-argument structure is provided by theories of *semantic roles* such as Frame Semantics (1985), and semantic role representations have been shown to be beneficial for a number of tasks ranging from question answering (Narayanan and Harabagiu, 2004) and the representation of propositional information in biomedicine (Cohen and Hunter, 2006) to cognitive tasks like modelling human sentence processing (Padó et al., 2006).

A crucial prerequisite for the use of semantic roles in NLP is the availability of robust and accurate models for the assignment of frames and roles to free text, a task often called *shallow semantic parsing*. Starting with the seminal study by Gildea and Jurafsky (2002), much effort has been spent on developing data-driven models for this task. Unfortunately, state-of-the-art shallow semantic parsing techniques still rely heavily on large annotated corpora. While such a resource is available for English in the form of the FrameNet database (Fillmore et al., 2003), the high cost of manual semantic annotation (Burchardt et al., 2006) has impeded the development of comparable resources for almost all other languages.

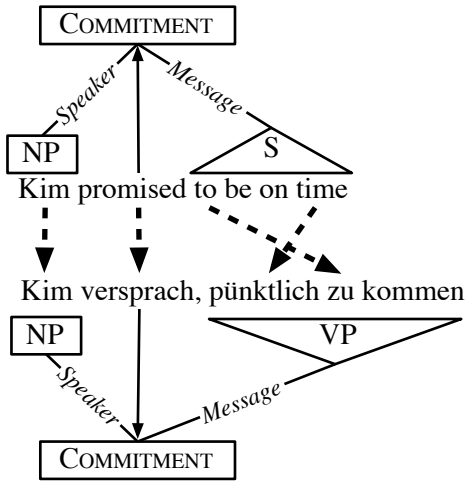


Figure 1: Annotation projection of frame-semantic annotation from English onto German.

An elegant solution to this resource scarcity problem is *annotation projection* (Yarowsky et al., 2001), a technique which uses parallel corpora to automatically transfer linguistic annotations from a source language onto a target language by following *translational equivalence links* in parallel sentence pairs (bisentences). Figure 1 illustrates this idea for frame-semantic annotation. We assume that the English side has been analysed – here, the verb *promise* introduces the frame COMMITMENT with the roles *Speaker* (assigned to the NP *Kim*) and *Message* (assigned to the

sentence *to be on time*). Annotation can now simply follow translational equivalence links (shown as dashed lines) to induce corresponding frame and role annotation for the German sentence: the frame COMMITMENT is now introduced by *versprach*, and its roles point to *Kim* (*Speaker*) and *pünktlich zu kommen* (*Message*), respectively. Thus, the manual work spent on the development of existing resources can be reused to create corresponding resources for other languages. Projection has been applied to the English FrameNet to induce corresponding resources for a number of languages, such as French (Padó and Pitel, 2007), German (Padó and Lapata, 2005b; Padó and Lapata, 2006), Spanish (Johansson and Nugues, 2005), and Swedish (Johansson and Nugues, 2006).

What must be kept in mind, however, is that the success of annotation projection relies on *cross-lingual parallelism*: The strategy proceeds by *copying* annotation directly across languages. When the translation does not preserve the linguistic analysis of the source sentence, projection is thus bound to assign an erroneous analysis to the target sentence.

At first glance, the parallelism assumption appears to be particularly problematic for frame-semantic analyses, since these consist of two levels, namely the frame assigned to the predicate and the realised roles. Due to the design of FrameNet, which defines semantic roles at the level of frames, roles can only be projected successfully if the frame is parallel. If a different frame is evoked by the target predicate, there is no guarantee that the projected roles are interpretable. It is therefore important to investigate the actual degree of frame and role parallelism in parallel corpora for different language pairs, to gauge the degree to which the parallelism assumption is warranted. Unfortunately, the studies listed above have concentrated on role parallelism, and either simply assumed frame parallelism, or limited their evaluation to cases of frame parallelism.

This paper addresses the question of cross-lingual parallelism on the *frame* level by providing a detailed, data-driven investigation. We base our discussion on the corpus of 1000 English–German bisentences with manual frame-semantic annotation described in Padó and Lapata (2005b).[1] We recently extended this corpus with a third language, by tagging the French translations of all original bi-sentences (Padó and Pitel, 2007). We re-used the original annotation guidelines, which allows us to verify our conclusions on two language pairs exemplifying different language families. In addition, the English–French bitext is not affected by possible biases introduced by the informed sampling strategy used for the creation of the English–German bitext.

[1]The corpus is available for download from `http://www.coli.uni-saarland.de/~pado/data.html`.

Plan of the paper. We proceed in three stages: In Section 2, we make the notion of cross-lingual parallelism more precise and investigate the scale of the problem. Section 3 then characterises the processes underlying frame non-parallelism using concepts from translation science. Finally, in Section 4, we sketch how the results motivate a simple heuristic to automatically detect affected instances.

2 Cross-lingual interpretability and parallelism

This Section provides a discussion of *cross-lingual parallelism*, the assumption that "translation preserves linguistic annotation". We introduced this assumption in Section 1 as essential for successful annotation projection, but have yet to develop a better understanding of it. Arguably, cross-lingual parallelism involves two steps, which we inspect in turn. First, it assumes that a linguistic theory for language A can be used "as is" for the analysis of language B (*cross-lingual interpretability*). Second, it assumes that the concrete translation process within each bisentence preserves the linguistic analysis of the source sentence (*cross-lingual parallelism* proper).

Cross-lingual interpretability. Cross-lingual interpretability is a claim about a linguistic theory T. It states that the *descriptive inventory* of T that is used to analyse some source language can be also used to analyse the target language in question.

While many early formal theories in linguistics (such as Chomsky's universal grammar or Katz and Fodor's theory of semantics) were aimed at perfect cross-lingual interpretability, this turned out to be infeasible in practice. However, these studies also yielded insights about properties of theories that lend themselves at least to high degrees of interpretability. One crucial factor is *granularity of description*: The

coarser the categories, the more likely they can be observed (and thus interpreted) cross-linguistically.[2]

Judging on these grounds, FrameNet frames can be expected to stand a good chance of cross-lingual interpretability. They can be seen as coarse-grained semantic classes of predicates which are *conceptually similar* by virtue of referring to schematised situations which can be expected to to apply across languages to a high degree (see also Boas (2005)). The cross-lingual interpretability of frames is limited by another factor, though. In addition to conceptual similarity, membership of a predicate in a frame has to be grounded *linguistically* by the predicate's syntactic ability to realise the frame's semantic roles. Thus, frames may not be interpretable in languages where the subcategorisation of predicates differs substantially from their English translations.

These expectations are borne out well by actual experiences from projects which directly re-use FrameNet frames for the semantic annotation of other languages (Subirats and Petruck, 2003; Ohara et al., 2004; Burchardt et al., 2006). Since cross-lingual interpretability is difficult to quantify, the evidence is qualitative in nature; however, several general tendencies have become apparent: (a), for any given language, a substantial majority of FrameNet frames is directly applicable; (b), the degree of interpretability is inversely related to the typological distance from English; (c), some semantic domains may be particularly problematic.

Cross-lingual parallelism. Even if a linguistic theory exhibits perfect cross-lingual interpretability, it can be completely unsuitable for annotation projection. The reason is that cross-lingual interpretability is not concerned at all with the analysis of *concrete utterances*. This point can be illustrated on the syntactic analysis of Figure 1, where we find that the syntactic category of the English phrase "to be on time" (sentence) diverges from the category of its German translation "pünktlich zu kommen" (verb phrase). Even though we can safely assume that the categories VP and S are interpretable in both languages, simple annotation projection would result in a wrong syntactic analysis for the German phrase.

[2]Naturally, this involves a trade-off: By concentrating on cross-lingual generalisations, the description provided by coarse-grained categories is, by definition, incomplete.

Language	Measure	Precision	Recall	F-score
EN→DE	FrameParal	0.72	0.72	0.72
	RoleParal	0.91	0.92	0.91
EN→FR	FrameParal	0.65	0.74	0.69
	RoleParal	0.88	0.87	0.88

Table 1: Cross-lingual parallelism of frame-semantic annotation on the frame and role levels

Thus, provided that interpretability is not an issue, the key question for the applicability of annotation projection is the degree of *cross-lingual parallelism* proper. We define cross-lingual parallelism to hold if a linguistic unit and its *translational equivalent* in a parallel corpus receive identical analyses.

The question of cross-lingual parallelism has been investigated by a number of studies for different levels of linguistic analysis, revealing an interesting trend: it appears that syntactic annotation, such as NP bracketings (Yarowsky et al., 2001) or dependency relations (Hwa et al., 2005) show only a quite low degree of cross-lingual parallelism (e.g., <40% for dependency relations).[3] In contrast, studies on lexical-semantic annotation in the widest sense, such as word sense (Bentivogli and Pianta, 2005) or coreference (Postolache et al., 2006), show a substantially higher degree of parallelism, often in excess of 80%.

These results give reason to hope that Frame Semantics, as an instance of lexical-semantic annotation, also shows a high degree of cross-lingual parallelism. To validate this hypothesis, we have analysed the corpus described in Section 1. The results are shown in Table 1. "FrameParal" measures how many frames simultaneously in both halves of a bisentence, and "RoleParal" measures the same number for semantic roles of parallel frames. To compute precision and recall, we treat the annotations of the target language (i.e., German and French) as gold standard against which we compare the English annotations. Since all major tendencies hold across both target languages, we discuss them jointly.

We first observe that provided that the frames are parallel, the roles are show a very high degree of parallelism (above 90%). This lends strong support to the amenability of semantic roles to annotation projection, and accounts for the favourable role projection results found by the studies listed in Section 1.

[3]Annotation projection for syntax therefore often employs post-projection rewriting steps to modify the source annotation.

Given our discussion above, however, this finding is not overly surprising, since due to the design principles of FrameNet, instances of frame parallelism draw from the same set of semantic roles. In fact, a more detailed analysis of the mismatch cases shows that most of the mismatches are cases of argument elision in one of the languages (e.g., passives).

It thus appears that the crucial factor in determining the prospects of annotation projection is in fact the degree of frame parallelism. Our data show that the situation is not as clear-cut as for role parallelism: The degree of frame parallelism is substantially lower, ranging around 70%. On the one hand, this number indicates that a majority of frame instances is preserved in translation, particularly taking into account that the *monolingual* inter-annotator agreement for frames on our dataset was not 100%, but 87%. On the other hand, it shows that there is a substantial fraction of instances of translated predicates where the frame changes, and thus annotation projection should not be applied.

3 Characterising Frame (Non)-Parallelism

In the last section, we have established that frame non-parallelism is in fact a substantial phenomenon. This section aims at characterising the circumstances of frame non-parallelism in parallel corpora.

3.1 Frame Parallelism and Translational Shifts

Recall from Section 2.2 that cross-lingual parallelism considers the relationship between the analyses of translationally equivalent linguistic units. Thus, a promising source for insight about cross-lingual parallelism is translation science. This field has known for a long time that *"translations deviate in many ways from their originals"* (Cyrus, 2006), and has investigated the linguistic changes arising from translation, termed *translational shifts* (Catford, 1965).

Cyrus' (2006) recent classification of translational shifts is particularly interesting for our study, since it is aimed specifically at investigating the relationship between predicates (and arguments) and their translations. It distinguishes two main classes of translational shifts. The first is *grammatical shifts*, such as change of voice, category change, or (de-) pronominalisation. The second consists of *semantic shifts*, the two most important of which are *modification* and *mutation*.[4] Modification is defined as "some type of semantic divergence, for example a difference in aktionsart", where the lexical meaning of the two predicates is still comparable. Modification has two subclasses, namely *explicitation* and *generalisation*, where more specific (or less specific) predicates are chosen as translations. The other class, mutation, covers cases of translation where the words "differ radically in their lexical meaning".

This classification throws some new light onto the difference between syntactic and lexico-semantic annotation observed in Section 2. Presumably, syntactic annotation is sensitive to grammatical shifts in translation, and by extension to semantic shifts which are often accompanied by grammatical shifts. In contrast, lexico-semantic annotation tends to abstract over grammatical properties, and thus can exhibit a higher degree of cross-lingual parallelism.

Figure 2 illustrates the case of FrameNet as a type of lexico-semantic annotation. In the figure, the translation process is modelled as consisting of an interpretation step, which recovers an underlying state of affairs from a source language expression, and a generation step, which re-expresses this state of affairs in a new language. This leads to an upside-down version of the well-known Vauquois triangle (Vauquois, 1975), where frames can be seen as an intermediate, partly language-independent layer.

The graph on the left shows the case of grammatical shifts. These do not involve a change in the frame, since all possible reformulations of a state of affairs which involve only a grammatical shift share a common frame. Even category change is unproblematic, since frames can be evoked by predicates of different parts of speech. For example, the frame COMMITMENT can be evoked by the verbs *promise, vow, pledge* as well as by the nouns *promise, oath*, and others. In contrast, the graph on the right displays the typical case of semantic shifts: when the translator decides to express the state of affairs in the target language with an expression that deviates considerably from the source expression, frame non-parallelism may arise. Note, though, that due to the fairly coarse granularity of frame-semantic classes, not every semantic shift results in frame non-parallelism. We will

[4] Cyrus lists two additional semantic shifts, namely addition and deletion, which we disregard since we only consider the case where two corresponding predicates exist.

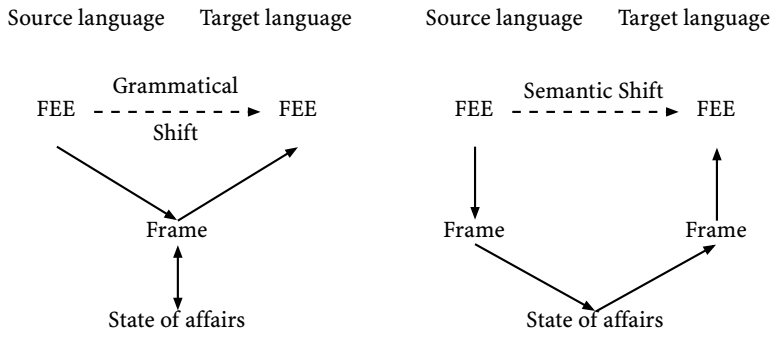


Figure 2: The connection between translation and frame parallelism (left) and frame non-parallelism (right)

discuss this point in detail in the next section.

3.2 Which Semantic Shifts Break Frame Parallelism?

In this section, we investigate what types of semantic shifts lead to frame non-parallelism. Optimally, such a study would be grounded in an small, but exhaustive, inventory of generic semantic shifts. Unfortunately, it seems that such an inventory is difficult to develop. Cyrus herself does not further subdivide her classes, noting that "it is rather difficult to find objective criteria" to distinguish even for the shifts she describes. The same issue has come up in the equivalent monolingual task of characterising the semantic relation between paraphrases in linguistic terms. For example, Barzilay and McKeown (2001) found that existing resources like WordNet have an insufficient inventory of relations and that "non-classical" relations are necessary. While a number of such relations have been investigated in the context of relation extraction over the last years, we are not aware of any successful efforts to construct a complete set of generic, "non-classical" semantic relations.

In the absence of such a resource, we use Cyrus' binary modification/mutation distinction as a basis, and provide an overview of the phenomena for which we find support in our corpus. While this method is clearly not exhaustive, it should provide an interesting insight into the types of semantic shifts that occur. For convenience, we always describe translation as taking place from English into another language.

Modification. Recall that modification assumes that lexical meaning is preserved to a large extent. Thus, frames can be parallel for "mild" cases of modification, while "serious" cases can result in non-parallelism. Modification is "mild" when the relation between the two predicates is one of synonymy, near-synonymy, or "mild" explicitation/generalisation. For example, both *say* and *wiederholen 'reiterate'* evoke the same frame, STATEMENT, even though *reiterate* is clearly an explicitation of *say*.

In contrast, a frequent phenomenon which leads to frame non-parallelism is translation that is sensitive to the predicates' arguments. For example, the predicates from the frame CAUSE_CHANGE_OF_SCALAR_POSITION, such as *increase* or *raise*, are used in English to very generally express processes of change. In contrast, French has a tendency to systematically use more specific frames, depending on the semantic type of the changing ITEM. Since FrameNet assumes that frame choice is determined lexically (by the predicate), frame non-parallelism can ensue, as in the following example:

(1) Extending the Community's legal competence within the framework of the third pillar has **increased** the burden.

Le fardeau s'est **alourdi** avec une extension de la compétence juridictionnelle communautaire dans le cadre du troisième pilier.

Here, the combination *increase [weight]* is translated with the more specific French *alourdir 'to make heavier'* which should presumably evoke the frame CAUSE_EXPANSION.

However, particularly in the case of explicitation/generalisation, the boundary to "serious" modification is hard to draw. Arguably, this mirrors a problem on the FrameNet side, namely the difficulty

to define in a precise manner the degree of *conceptual similarity* necessary for predicates to evoke the same frame, and to do so independently of the semantic domain (cf. Ellsworth et al. (2004)). In fact, it appears that the granularity of FrameNet frames is not completely uniform across all frames. In addition, there seems to be a tendency in FrameNet over time towards constructing more fine-grained frames which require a higher degree of conceptual similarity. This development is problematic from a cross-lingual point of view, since it leads to a higher number of instances with frame non-parallelism.

As a last prominent modification phenomenon, consider *change in aktionsart*, which is a clear-cut case of "serious" modification. In the following example, the causative English *raise* is translated by the inchoative French *monter 'rise'*:

(2) [. . .] The employment rate within the EU can be **raised** to 70%.

Le niveau d' emploi pourrait **monter** à 70% dans l' UE.

The resulting change in valency means that the English frame CAUSE_CHANGE_OF_SCALAR_POSITION cannot be evoked by the translation. Rather, *monter* evokes CHANGE_OF_SCALAR_POSITION.

As a final remark on modification, note that the frame non-parallelism introduced by this class of shifts is "benign" in that the semantic relation between original and translation almost always corresponds to a frame-to-frame relation in the FrameNet frame hierarchy. Examples are inheritance (for explicitation) or causative-of/inchoative-of (for change of aktionsart). When this hierarchy, which is still mostly exemplary, reaches a more complete state, it may be possible to treat all instances of modification as cases of (generalised) frame parallelism.

Mutation. Recall that in the case of mutation, the lexical meaning of the translation differs substantially from the original. Thus mutation, as a rule, results in frame non-parallelism. The right hand side of Figure 2 furthermore illustrates that mutation involves the usage of two different frames to describe the underlying state of affairs. The fact that this happens is not particularly surprising: Frames do not describe the complete meaning of the predicates they describe, but only its most salient meaning aspect. Since all but the most simple real-world states of affairs combine more than one meaning aspect, there are almost always "several ways of putting it", with different frames competing for the linguistic realisation.

The instances of mutation which we find in our corpus form a very inhomogeneous group, and are located along a continuum of *genericity*. On one end, we find cases which can be characterised well in terms of generic lexical relations such as causation, event–subevent, process–result, or perspectivisation. In this sense, they are similar to modification cases. However, translations further along the continuum become more and more idiosyncratic. The extreme is formed by instances whose interpretation involves a lot of world knowledge, and which are presumably very difficult to classify in terms of a general set of lexical relations.[5] The following examples illustrate different points on this continuum.

First, Example (3) shows an instance of a clear generic relation, where English expresses a process (*increase*), while French expresses the resulting end state (*atteindre 'reach'*):

(3) Why, for example, was the proposal to **increase** Europe's active population to 75% of the total population removed?

Pourquoi a-t-on retiré par exemple la proposition prévoyant que la population active devait **atteindre** 75% en Europe?

The relation between source and target expression becomes more elusive in the following example:

(4) The legal issue should take second place to consumer protection and **preventing** the public from harm.

La question juridique doit venir après la protection des consommateurs et les **précautions** pour nos citoyens.

What exactly is the relation between *prevent* and *précaution 'precaution'*? One possible interpretation is as a weaker version of process–result relation from above: since a precaution only *typically* implies that something is prevented, the relation might be characterised as process–typical result. However, this is not

[5] Arguably, metaphors form a prominent class of idiosyncratic mutations. They do not figure prominently in our study, through, since the guidelines asked our annotators to annotate "understood" rather than "literal" meanings.

the only possible characterisation: It is also possible to argue that a precaution is introduced *in order to* prevent something, and that the example is thus an instance of a means–purpose relation.

The last example illustrates the far end of the genericity continuum:

(5) Questions that were not **answered** during Question Time shall be answered in writing.

Les questions qui ne sont pas **examinées** pendant l' heure des questions recevront une réponse écrite.

The relation between *answer* and its translation *examiner 'examine'* can only be understood in the particular context of queries, where a response typically involves examining the issue at hand. This specificity makes it very hard to classify the translation pair *answer–examiner* in terms of generic relations.

4 Detecting Frame Non-Parallelism

We now come back to our original motivation, namely annotation projection of frame-semantic annotation. As we argued in Section 2, we need to detect instances of frame parallelism and non-parallelism to limit projection to parallel cases.

Our data analysis in Section 3, however, has left us with the impression that this distinction is difficult to make in a linguistically informed way. The inhomogeneity of the instances of non-parallelism combined with the difficulty of consistently delimiting FrameNet frames makes it difficult to relate this distinction in a straightforward manner to linguistic properties of the translation. We therefore propose to identify frame parallelism *distributionally* by observering properties of the translation and its context, a strategy which we have found to be effective for a related task, namely the cross-lingual induction of frame-semantic lexicons (Padó and Lapata, 2005a).

In this paper, we limit ourselves to outlining a naive heuristic on the type level, namely the *predominant sense heuristic* originally proposed for word sense disambiguation (McCarthy et al., 2004). This heuristic, which always assigns the most frequent (predominant) sense, is a serious rival for deeper methods, due to the skewed frequency of word senses. In our scenario, the predominant sense is the frame that is most often assigned to a predicate in the target language, and the heuristic consists in performing projection only if the source language frame is the predominant sense. Let t be a target predicate, s a source predicate, and f a frame. The decision rule is

$$\text{Project } f \text{ onto } t \text{ iff } \quad P(f|t) = \operatorname*{argmax}_f P(f|t)$$

where

$$P(f|t) = \frac{P(f,t)}{P(t)} = \frac{\sum_s P(f,t,s)}{\sum_f \sum_s P(f,t,s)}$$

We estimated the joint probabilities $P(f,t,s)$ from the complete EUROPARL corpus, using word alignments as indicators of translational equivalence, and testing two strategies for counting frames. The first was completely unsupervised and simply treated all frames listed in FrameNet for some source predicate s as seen for each instance of s (AllFrames). The second used a state-of-the-art frame disambiguation system (Erk, 2005) to assign the single most probable frame to each instance of s (BestFrame).

Language	Measure	Precision	Recall	F-score
EN→DE	Unfiltered	0.72	0.72	0.72
	AllFrames	0.89	0.51	0.65
	BestFrame	0.91	0.48	0.63
EN→FR	Unfiltered	0.65	0.74	0.69
	AllFrames	0.88	0.53	0.66
	BestFrame	0.90	0.49	0.63

Table 2: Impact of "predominant sense" filtering on cross-lingual frame parallelism

Table 1 shows the evaluation of these strategies on our 1000-sentence test corpus. The two language pairs again behave similarly. The results are encouraging: The predominant sense heuristic is able to detect the majority of instances of non-parallelism even without disambiguation (AllFrames), thus substantially improving precision. Frame disambiguation (BestFrame) reaps an additional small benefit, with final precision figures around 90%.

The practical applicability of this filtering scheme depends on the application, though. While the deterioration in recall that results from filtering (from 70% to around 50%) is presumably not a large problem, considering the size of parallel corpora available, predominant sense filtering results in a dataset for the target language where each target predicate is tagged with only one frame, namely the predominant one, and where all "minority readings" are discarded.

5 Conclusions

In this paper, we have discussed the question of cross-lingual parallelism of FrameNet frame instances in a parallel corpus. This problem is relevant in the context of using annotation projection to create frame-semantic resources for new languages. We have first discussed the concepts of cross-lingual interpretability and parallelism, and then characterised the cross-lingual parallelism of semantic frames quantitatively (in a parallel corpus) and qualitatively (in relation to translational shifts). Finally, we have sketched a strategy for identifying non-parallel instances.

We see two main avenues of future research. The first is a task-based evaluation of non-parallelism detection in the context of inducing shallow semantic parsers for the target language (Johansson and Nugues, 2006). The second is the replacement of the simplistic type-level "predominant sense" heuristic used in this paper by a token-level model of parallelism based on the lexical context. Such strategies work well in monolingual contexts (Erk, 2006) and have the potential of alleviating both the recall and the monosemy problem.

Acknowledgments. The work reported in this paper was supported by the DFG (grant Pi-154/9-2).

References

R. Barzilay, K. R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th ACL*, 50–57, Toulouse, France.

L. Bentivogli, E. Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. *Journal of Natural Language Engineering*, 11(3):247–261.

H. C. Boas. 2005. Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18(4):445–478.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, M. Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th LREC*, Genoa, Italy.

J. Catford. 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. Oxford University Press.

K. B. Cohen, L. Hunter. 2006. A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics*, 7(Suppl. 3):S5.

L. Cyrus. 2006. Building a resource for studying translation shifts. In *Proceedings of the 5th LREC*, Genoa, Italy.

M. Ellsworth, K. Erk, P. Kingsbury, S. Padó. 2004. PropBank, SALSA and FrameNet: How design determines product. In *Proceedings of the LREC Workshop on Building Lexical Resources From Semantically Annotated Corpora*, Lisbon, Portugal.

K. Erk. 2005. Frame assignment as word sense disambiguation. In *Proceedings of the 6th International Workshop on Computational Semantics*, Tilburg, The Netherlands.

K. Erk. 2006. Unknown word sense detection as outlier detection. In *Proceedings of the joint HLT and NAACL*, 128–135, New York City, NY.

C. J. Fillmore, C. R. Johnson, M. R. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

C. J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, IV(2):222–254.

D. Gildea, D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Special Issue of the Journal of Natural Language Engineering on Parallel Texts*, 11(3):311–325.

R. Johansson, P. Nugues. 2005. Using parallel corpora for automatic transfer of FrameNet annotation. In *Proceedings of the 1st ROMANCE FrameNet Workshop*, Cluj-Napoca, Romania.

R. Johansson, P. Nugues. 2006. A FrameNet-Based Semantic Role Labeler for Swedish. In *Proceedings of the joint ACL and COLING*, 436–443, Sydney, Australia.

D. McCarthy, R. Koeling, J. Weeds, J. Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42th ACL*, 279–286, Barcelona, Spain.

S. Narayanan, S. Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the 20th COLING*, 693–701, Geneva, Switzerland.

K. H. Ohara, S. Fujii, T. Ohori, R. Suzuki, H. Saito, S. Ishizaki. 2004. The Japanese FrameNet project: An introduction. In *Proceedings of the LREC Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon, Portugal.

S. Padó, M. Lapata. 2005a. Cross-lingual bootstrapping for semantic lexicons. In *Proceedings of the 22nd AAAI*, 1087–1092, Pittsburgh, PA.

S. Padó, M. Lapata. 2005b. Cross-lingual projection of role-semantic information. In *Proceedings of the joint HLT and EMNLP*, 859–866, Vancouver, BC.

S. Padó, M. Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of the joint ACL and COLING*, 1161–1168, Sydney, Australia.

S. Padó, G. Pitel. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. In *Proceedings of TALN*. To appear.

U. Padó, F. Keller, M. W. Crocker. 2006. Combining syntax and thematic fit in a probabilistic model of sentence processing. In *Proceedings of the 28th CogSci*, 657–662, Vancouver, BC.

O. Postolache, D. Cristea, C. Orasan. 2006. Tranferring coreference chains through word alignment. In *Proceedings of the 5th LREC*, Genoa, Italy.

C. Subirats, M. R. L. Petruck. 2003. Surprise! Spanish FrameNet! In *Proceedings of the Workshop on Frame Semantics, XVII. International Congress of Linguists*, Prague, Czech Republic.

B. Vauquois. 1975. *La traduction automatique à Grenoble*. Dunod, Paris.

D. Yarowsky, G. Ngai, R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st HLT*, 161–168, San Francisco, CA.

Frame-semantic Annotation on a Parallel Treebank

Martin Volk and Yvonne Samuelsson
Stockholm University
Department of Linguistics
106 91 Stockholm, Sweden
volk@ling.su.se

Abstract

This paper reports on experiments in frame-semantic annotation of a parallel treebank. Selected English and Swedish sentences that contained verbs of motion and communication were annotated independently by two annotators. We found that they assigned the same frame to corresponding sentences in 52% of the cases. This leads us to the conclusion that parallel treebanks can save considerable effort when building semantically annotated resources.

1 The parallel treebank SMULTRON

We have developed a German-English-Swedish parallel treebank, consisting of around 1000 sentences in each language. The first part of our parallel treebank consists of chapters one and two of Jostein Gaarder's novel "Sophie's World". The second part contains economy texts, taken from a quarterly report by a multinational company, a bank's annual report and a text about a banana certification program.

The name treebank is derived from the fact that syntax structures are mostly encoded as tree graphs. In the annotation we followed the Penn Treebank guidelines for the English trees and the NEGRA/TIGER guidelines for the German trees. For Swedish we adapted the German guidelines. The syntactic annotation for all three languages was done with the ANNOTATE treebank editor. Language-specific chunkers suggested partial trees which were manually checked. This step was followed by automatic tree deepening and extensive consistency checking.

We then aligned the trees in our treebank on the word and phrase level across languages. The alignment is meant to capture translation correspondences in the sense that a phrase pair could be cut out of the trees and reused in an example-based translation system. We distinguish between exact alignment and approximate alignment. This distinction is often debateable but should help if multiple translation alternatives are available for the subsequent MT system. The alignment was done with the TreeAligner, a graphical tool that allows to quickly draw the different alignment lines. We have named our treebank SMULTRON (Stockholm MULtilingual TReebank) and described its development in (Volk and Samuelsson, 2004; Volk et al., 2006), and (Samuelsson and Volk, 2006).

Figure 1 shows an example of parallel trees with word and phrase alignment. The English phrase *"When she crawled through it"* is an exact translation equivalent of *"När hon kröp genom den"* and is therefore aligned with a green line. But the phrase *"a large cavity between the bushes"* is only roughly equivalent to *"en liten håla inne bland buskarna"* (which literally means "a little hole in between the bushes"). Note that we allow m:n sentence alignments and 1:n word and phrase alignments.

The monolingual treebanks are represented in TIGER-XML which defines unique identifiers for all tokens and nodes in the trees. Our alignment uses these identifiers and stores the alignment information in a separate XML file.

2 Frame-semantic Annotation of Parallel Trees

On top of the syntactic annotation we have started to annotate the trees with frame-semantic labels. This was undertaken in student projects for English (Ivantsova, 2006) and for Swedish (Otsa, 2006).[1] In these projects we have focused on frames for motion and communication. 50 trees were handpicked from the Sophie part of our parallel treebank. We made sure that the sentences

[1] Both reports are available at www.ling.su.se/DaLi [Publications].

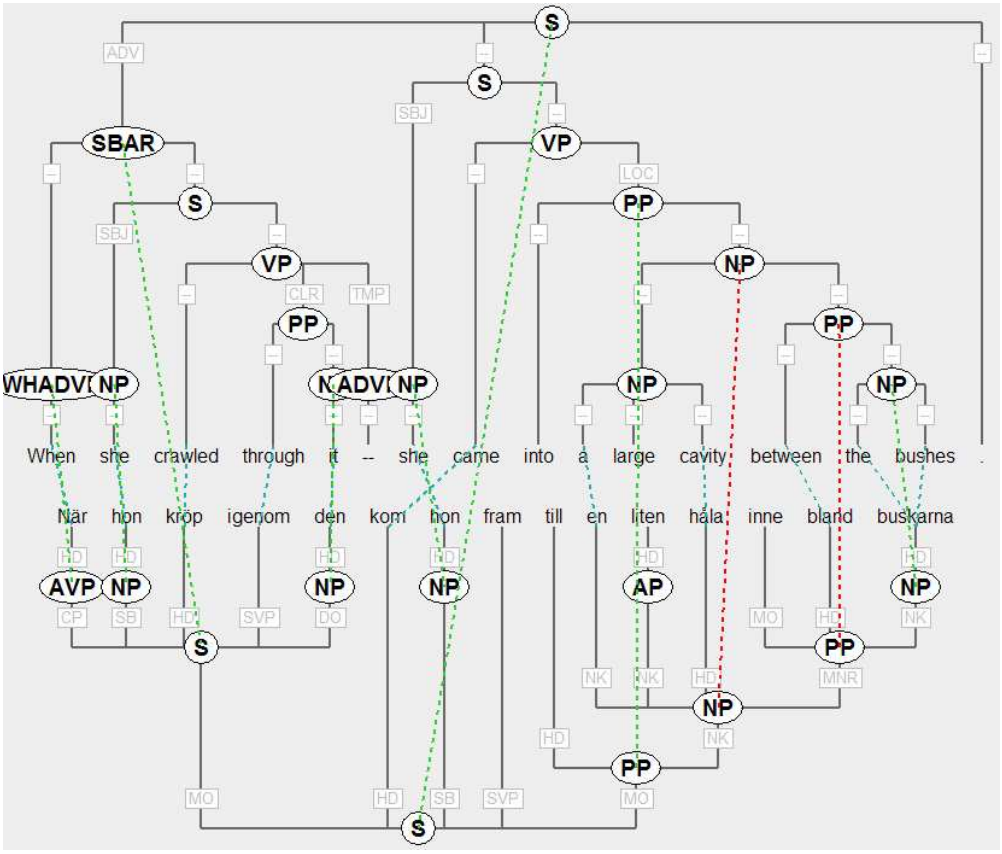


Figure 1: English-Swedish parallel trees with word and phrase alignment.

in both languages contained a verb of motion and communication. For example the English sentence *"She had walked the first part of the way with Joanna"* corresponds to the Swedish sentence *"Den första biten hade hon haft sällskap med Jorunn"*. But while the English sentence contains a motion verb *"walk"*, the Swedish has lost this aspect. It literally translates as *"The first part she had had company with Jorunn"*.

The selected sentences were then independently annotated by the two students in the English and Swedish treebank respectively. The goal of these projects was to see how often the two annotators would assign the same frames in parallel trees. Both used the SALSA tool which was developed for the frame-semantic annotation of German (Erk and Pado, 2004).[2]

Figure 2 shows the result of the frame semantic annotation of the English example tree. It contains the frames Self_motion and Arriving. The Self_motion frame has five elements.[3] The frame elements Self_mover and Path are realized in this sentence and are thus annotated, while Area, Source and Goal are left unattached.

The students used the FrameNet definitions (Fillmore et al., 2003) when they decided which frames and which frame elements to assign. For example, the description of the **Self_motion** frame includes the following definitions:

- The **Self_mover**, a living being, moves under its own power in a directed fashion, i.e. along what could be described as a Path, with no separate vehicle.
- **Goal** is used for any expression which tells where the Self_mover ends up as a result of the motion. E.g. *The children SKIPPED into the park.*
- **Path** is used for any description of a trajectory of motion which is neither a Source nor a Goal. E.g. *The scouts HIKED through the desert.*
- **Source** is used for any expression which implies a definite starting-point of motion. E.g.

[2] See http://www.coli.uni-saarland.de/projects/salsa/

[3] Frame elements are sometimes called "slots" or "roles" in the literature.

The cat RAN out of the house.

- Frame-evoking elements: *crawl, hike, run, skip, walk, ...*

The SALSA tool proved to be very useful for the frame-annotation of both the English and Swedish trees. It takes a TIGER-XML representation of the treebank as its input. It shows a graphical representation of one syntax tree at a time (with or without PoS tags and function labels) and allows the assignment of frames and frame elements. And it saves the result in an extended TIGER-XML file.

The annotator can preselect a set of frames from all defined FrameNet frames. We preselected all frames for motion and communication. The annotator can then assign a frame to a given tree by manually picking from a menu listing. We used eight different motion frames (Arriving, Source_Path_Goal, Body_movement, Cause_motion, Change_direction, Change_posture, Motion, and Self_motion) and six different communication frames (Communication, Communication_noise, Discussion, Questioning, Statement, Telling). Frames were mostly assigned to verbs but sometimes also to phrases (e.g. *was on her way* is assigned a motion frame).

3 Results

For the 50 English sentences 65 frames (17 frame types) were assigned. We list the frames and their frequencies in table 1. The 65 English frames come with 158 instantiated frame elements (26 frame element types). 34 English frames were identical to the frames annotated in the Swedish sentences (52%). In another 22 cases the annotators had assigned closely related frames (e.g. Motion vs. Self_motion) in the two languages. Clear annotation differences arose when the verb choice differed clearly. For example, the English sentence starting with *She was frequently told that ...* in our treebank corresponds to the Swedish *Hon fick ofta höra att ...* (literally: She often got to hear that ...).

This indicates that frame annotation done for one language can be automatically projected to a parallel text. For example, if the semantic frames are annotated for the English sentence "*When she crawled through it she came into a large cavity between the bushes*" (as in figure 2) and when the English syntax tree is aligned to its Swedish counterpart (as in figure 1), then we will be able to automatically transfer the semantic frames to the corresponding Swedish tree. This idea has also been explored by (Pado and Lapata, 2006) for German - English projections on automatic phrase alignments.

Motion	freq
Arriving	2
Body_movement	1
Cause_motion	3
Change_direction	1
Change_posture	1
Cotheme	1
Motion	10
Placing	1
Seeking	1
Self_motion	17
Source-Path-Goal	4
Communication	
Communication	3
Communication_noise	1
Discussion	1
Questioning	6
Statement	9
Telling	3

Table 1: Frames used in the annotation of the English sentences

When we transfer a frame from one sentence to a parallel sentence in another language, then we want both a correct anchoring of the frame in the target language and the correct assignment of the frame elements. This latter step adds to the complexity since some of the frame elements which are realized in the source sentence might not be realized in the target sentence and vice versa.

As a side effect we investigated whether the frames which were originally defined for English were also suitable for Swedish. We found that this was the case. Of course, the selection of the appropriate frames takes more time and effort for Swedish since the Frame-evoking elements (i.e. the verb or phrase triggering a certain set of frames) needs to be translated to English, but then it worked nicely. But we concede that our study was small and therefore we might have missed fine-grained distinctions as found for German-English by (Burchardt et al., 2006). They noticed, for instance, that the "use of dative objects

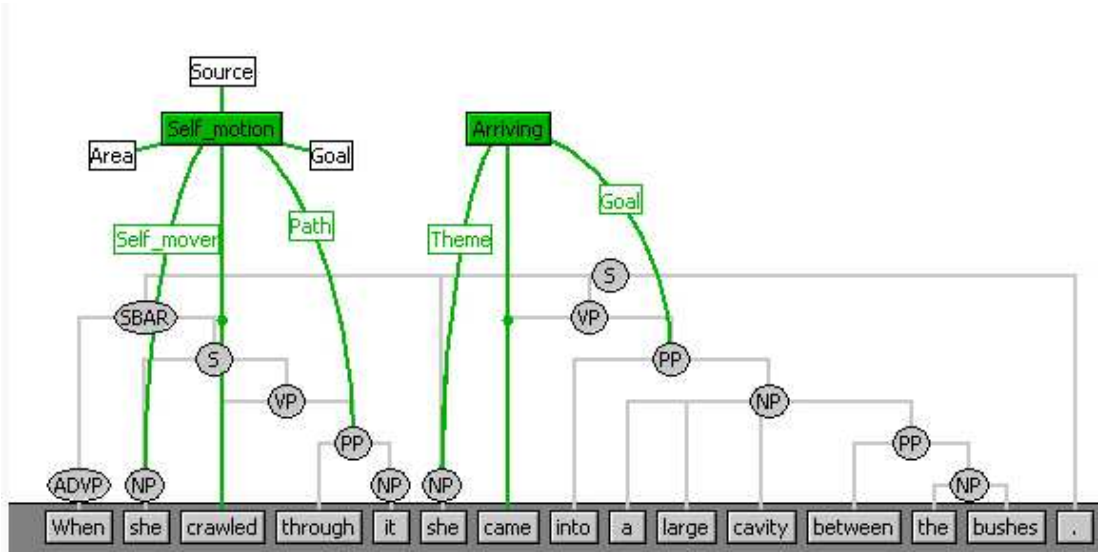


Figure 2: An English syntax tree with frame semantic annotations.

is much less restricted in German than in English". This meant that sometimes an English frame fitted a German sense, but lacked the necessary frame elements. We suspect that similar deviations will eventually arise when porting the English or German frames to Swedish.

4 Conclusions

Our study has demonstrated the usefulness of the SALSA tool and the English frame definitions for frame-semantic annotation of English and Swedish trees. But even more important, it indicates that automatic frame transfer across languages will work in more than 50% of the cases when given a good phrase-alignment. We have not investigated the correctness of the frame element transfer.

Our ultimate goal is to develop a methodology for the large scale annotation and interpretation of parallel texts which is both fast and accurate. Such a methodology will lead to valuable resources for Computational Linguistics, General Linguistics and Translation Studies.

Our parallel treebank provides unique annotation and evaluation material for such a project. We will focus on annotation projection, i.e. to transfer annotation that is computed with certainty for one language to the parallel languages.

References

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006. The SALSA corpus: A German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, pages 969–974, Genoa.

Katrin Erk and Sebastian Pado. 2004. A powerful and versatile XML format for representing role-semantic annotation. In *Proc. of LREC-2004*, Lisbon.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Natalya Ivantsova. 2006. Enriching a treebank with semantic information in the frame semantics paradigm. C-uppsats, Stockholm University, April.

Annika Otsa. 2006. Berikning av en trädbank med semantisk information. C-uppsats, Stockholms Universitet, April.

Sebastian Pado and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of ACL-COLING 2006*, pages 1161–1168, Sydney, Australia.

Yvonne Samuelsson and Martin Volk. 2006. Phrase alignment in parallel treebanks. In Jan Hajic and Joakim Nivre, editors, *Proc. of the Fifth Workshop on Treebanks and Linguistic Theories*, pages 91–102, Prague, December.

Martin Volk and Yvonne Samuelsson. 2004. Bootstrapping parallel treebanks. In *Proc. of Workshop on Linguistically Interpreted Corpora (LINC) at COLING*, Geneva.

Martin Volk, Sofia Gustafson-Capková, Joakim Lundborg, Torsten Marek, Yvonne Samuelsson, and Frida Tidström. 2006. XML-based phrase alignment in parallel treebanks. In *Proc. of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*, Trento, April.

Colophon

This workshop was organized by researchers of Lund University and was hosted by the University of Tartu. While they are now located into two distinct countries, both universities share common roots and were established while Sweden was an empire. Lund University is probably the largest in the Nordic/Baltic area with nearly 40,000 students, but Tartu has the precedence if we consider seniority. By date of foundation, Tartu's rank is third (1632) after Greifswald (1456), and Uppsala (1477), but before Åbo (1640) and Lund (1666).

The artwork on the cover represents a kannel and an Estonian belt pattern:

> **Kannel:** A kannel (Estonian) or kantele (Finnish) is a traditional plucked string instrument of the zither family. It is related to the Russian gusli, the Latvian kokle and the Lithuanian kanklės. Together these instruments make up the family known as Baltic Psalteries.
>
> Source: Wikipedia

> **Estonian belt pattern:** A typical feature of the creative activity of [Estonia] is an urge to kirjata 'to compose a pattern'. Throughout centuries people have used the Estonian term kiri 'writing' instead of the borrowed muster 'pattern' or ornament.
>
> Source: http://www.einst.ee/publications/crafts_and_arts/

Jonas Wisbrant designed the cover of these proceedings and produced the layout using Adobe InDesign CS2 using a Frutiger font.

Åke Viberg:
Wordnets, Framenets and Corpus-based Contrastive Lexicology

Lars Borin, Maria Toporowska Gronostaj, Dimitrios Kokkinakis:
Medical Frames as Target and Tool

Susanne Ekeklint, Joakim Nivre:
A Dependency-Based Conversion of PropBank

Richard Johansson, Pierre Nugues:
Using WordNet to Extend FrameNet Coverage

Karel Pala, Aleš Horák:
Building a Large Lexicon of Complex Valency Frames

Sebastian Padó:
Translational Equivalence and Cross-lingual Parallelism: The Case of FrameNet Frames

Martin Volk, Yvonne Samuelsson:
Frame-semantic Annotation on a Parallel Treebank

ISBN 978-91-976939-0-5
ISSN 1404-1200
Report 90, 2007
LU-CS-TR: 2007-240
Printed in Sweden, E-husets tryckeri, Lund 2007

LUND UNIVERSITY

Department of Computer Science
http://nlp.cs.lth.se

Institute of Computer Science
http://math.ut.ee

Production: Jonas Wisbrant • 2007