To Cause Or Not To Cause: Cross-Lingual Semantic Matching for Paraphrase Modelling

Sebastian Padó and **Katrin Erk***
Computational Linguistics, Saarland University
Saarbrücken, Germany
{pado, erk}@coli.uni-sb.de

Abstract

This paper presents a manual pilot study in cross-linguistic analysis at the predicate-argument level. Looking at translation pairs differing in their parts of speech, we find that predicate-argument structure abstracts somewhat from morphosyntactic language idiosyncrasies, but there is still considerable variation in the distribution of semantic material over predicates. We propose an algorithm for automatically identifying matching predicate-argument structures (*frame paraphrases*). The resulting data allows an analysis of semantic differences e.g. in the expression of different degrees of causality.

1 Introduction

Aligned parallel corpora allow for many different types of cross-lingual knowledge transfer. They have been employed e.g. for transferring part of speech tags (Yarowsky et al., 2001), syntactic structure (Smith and Smith, 2004), and semantic classes (Padó and Lapata, 2005). While these methods have exploited alignment on different linguistic levels, there is no work so far on analysing the matches and mismatches between the predicate-argument level structures of parallel sentences.

Such an enquiry has much to offer: Representations on the predicate-argument level such as PropBank (Palmer et al., 2005) and FrameNet (Fillmore et al., 2003) are being increasingly used for all kinds of NLP applications that require a deeper level of text understanding than syntax, such as Question Answering (Narayanan and Harabagiu, 2004) and Information Extraction (Moschitti et al., 2003), especially in cross-linguistic settings such as CLEF (Peters and Braschler, 2001).

In this paper, we use frame semantics (Fillmore, 1982) to represent the predicate-argument structures (frame structures) of parallel sentences. Viewing an aligned sentence pair as a cross-lingual paraphrase, we try to identify matching parts in the frame structures. Matching parts, which we call *frame paraphrases*, contain semantically equivalent material, which may however be distributed differently across frames. Frame paraphrases can also be used for identifying *monolingual* paraphrases.

This paper reports a manual pilot study with the following contributions: (a) on a small sample, we assess the degree of parallelism and nonparallelism in frame structures for a translation pair with differing parts of speech; (b) we propose an algorithm for finding matching parts of frame structures in the nonparallel cases; and (c) we find different realisation possibilities for degrees of causality on a continuum from causative to inchoative.

2 Frame semantics

Frame Semantics (Fillmore, 1982) models the meaning of a word or expression by reference to a *frame* which describes the background and situational knowledge necessary for understanding what the predicate is "about". Each frame provides its specific set of semantic roles, called *frame elements (FEs)*, which represent the participants and props of the situation. Table 1 shows the definitions of

*This study was conducted within the SALSA project. We acknowledge the funding of the DFG (Grant PI 154/9-2)

	CHANGE_POSITION_ON_A_SCALE (CPOS)
Def	This frame consists of words indicating the change of an ITEM'S position on a scale.
FEs	ITEM The **tea price** rose.
FEEs	advance.v, decline.n, decline.v, decrease.n decrease.v, diminish.v, double.v, increase.v, rise.v

	CAUSE_CHANGE_OF_SCALAR_POSITION (CCOSP)
Def	This frame consists of words indicating that an AGENT or CAUSE affects the position of an ITEM on a scale.
FEs	AGENT **Lipton's** increased the tea price. CAUSE **The draught** increased the tea price. ITEM Lipton's increased the **tea price**.
FEEs	cut.n, cut.v, decrease.v, diminish.v, growth.n, increase.v, lower.v, move.v, raise.v, reduce.v

Table 1: Frames CPOS and CCOSP

the frames CHANGE_POSITION_ON_A_SCALE (in the following abbreviated to CPOS) and CAUSE_CHANGE_OF_SCALAR_POSITION (CCOSP), respectively, which are distinguished by the (non-)existence of an AGENT/CAUSE role.

The Berkeley FrameNet project (Fillmore et al., 2003) is building a semantic lexicon for English describing the frames and linking them to the words and expressions that can *evoke* them, called *frame-evoking elements (FEEs)*. FEEs can be verbs as well as nouns, adjectives, prepositions, adverbs, and multiword expressions. Currently, FrameNet contains over 600 frames with about 8,900 FEEs, exemplified in more than 135,000 annotated sentences from the British National Corpus, which are being used to train systems for the automatic assignment of frames and frame elements.

Frame semantic analysis. Frame-semantic analysis models the predicate-argument structure of a sentence, ignoring information such as negation, modality and tense. It abstracts from different realisations of roles, as in (1) and (2), and over FEEs that belong to different parts of speech (3); also, support constructions and multiword expressions receive the same analysis as single-word FEEs (4). This makes frame semantics interesting for the study of paraphrases, which may differ in syntactic realisation but agree in their semantics.

(1) $[\text{I}]_{\text{Donor}}$ $[\text{gave}]_{\text{FEE}}$ $[\text{Mary}]_{\text{Receiver}}$ [a book]$_{\text{Theme}}$.
$[\text{I}]_{\text{Donor}}$ $[\text{gave}]_{\text{FEE}}$ [a book]$_{\text{Theme}}$ [to Mary]$_{\text{Receiver}}$.

(2) $[\text{He}]_{\text{Speaker}}$ $[\text{demanded}]_{\text{FEE}}$ [a reimbursement]$_{\text{Message}}$.
$[\text{He}]_{\text{Speaker}}$ $[\text{demanded}]_{\text{FEE}}$ [to be reimbursed]$_{\text{Message}}$.

(3) $[\text{I}]_{\text{Self-mover}}$ $[\text{hurried}]_{\text{FEE}}$ [to the library]$_{\text{Goal}}$.
$[\text{I}]_{\text{Self-mover}}$ $[\text{trudged}]_{\text{FEE}}$ [to the library]$_{\text{Goal}}$.
The $[\text{walk}]_{\text{FEE}}$ [to the library]$_{\text{Goal}}$ is quite pleasant.

(4) $[\text{John}]_{\text{Cognizer}}$ $[\text{analyzed}]_{\text{FEE}}$ [the data]$_{\text{Phenomenon}}$.
$[\text{John}]_{\text{Cognizer}}$ $[\text{did}]_{\text{Support}}$ an $[\text{analysis}]_{\text{FEE}}$ [of the data]$_{\text{Phenomenon}}$.

Cross-lingual frame semantics. Resources mirroring the English FrameNet are currently being developed for Spanish, Japanese, German and Chinese, with a prospect of other languages to follow soon. All resources refer to the same frames, listing language-specific FEEs and complementing the frame set with language-specific additions.

This allows us to study paraphrases across languages in the same way as paraphrases within a language, as *cross-lingual frame paraphrases*: single frames or frame groups that have the same or similar meaning[1]. Rather than paraphrases consisting of monolingual word sequences, these frame paraphrases are (partially) language independent and can be realised in different, language-specific ways.

This immediately raises the question of whether a sentence and its translation will always receive the same frame-semantic analysis – if so, FrameNet would constitute an interlingua representation of the predicate-argument structure. In the following, we

[1]Note that due to the granularity of FrameNet, two phrases with the same frame-semantic analysis may even be opposite in meaning, e.g. the frame MORALITY_JUDGMENT is evoked by both "good" and "evil". To distinguish these two words, an additional, more fine-grained resource is necessary.

will see that this is not always the case, and investigate the mismatches we find in a corpus sample.

3 Data

In order to find an interesting sample for the identification of cross-lingual paraphrases, we inspected the list of English FEEs for the two related frames CPOS and CCOSP, and their German translations. The English verb "increase" has both a CCOSP reading (transitive instances) and a CPOS reading (intransitive and noun instances):

(5) There is a desire to **increase** public spending. (CCOSP)

(6) By 2010, emissions will **increase** by 6%. (CPOS)

We also found that English "increase" is frequently translated into German as one of the adjectives "höher" (higher) and "größer" (larger). Translation pairs across parts of speech constitute an interesting test case for frame-semantic accounts of paraphrases: Even though FrameNet generalises over parts of speech, it distinguishes frames on the basis of realisable roles, which means that the adjective "höher", which cannot realise the CAUSE/AGENT itself, cannot evoke the CCOSP frame that the transitive usage of "increase" can.

We therefore decided to analyse the occurrences of the translation pair "increase–höher" in the bilingual English/German part of the EUROPARL corpus (Koehn, 2002), the collection of the multilingual Proceedings of the European Parliament from 1997 through 2003. The sample we study includes occurrences of the verb "increase" as well as occurrences of the noun and of the past participle "increased", which is often used as an adjective. We word-aligned and parsed the sentences containing the aligned translation pair, resulting in 122 sentence pairs. In all sentences, we hand-corrected word alignment and syntactic structure and manually assigned FrameNet frames. When no appropriate frames were available in FrameNet, we constructed new ones in accordance with FrameNet frame construction principles. In the following, these frames are marked with an asterisk.

English	German	Count
CPOS (36 n, 13 v, 24 ppart)	CPOS (adj)	73
CCOSP (49 v)	CPOS (adj)	49

Table 2: Frames evoked by increase/höher

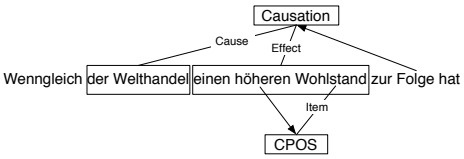
4 Experiment 1: Direct frame match

In a first experiment, we assess the simplest possible hypothesis, namely that there is a direct match between the frames evoked by "increase" in English and an aligned "höher" in German. Unsurprisingly, the results in Table 2 show that there is a significant amount of frame mismatch: The adjective in German can only be used inchoatively. This corresponds to the English noun and past participle cases; however, the majority of verb uses in English is transitive, and invokes the (mismatching) CCOSP frame.

However, it is not the case that all mismatches are true cases of different conceptualization. A number of English CCOSP instances are transitive, but used in subjectless constructions without CAUSE/AGENT; on the other hand, we observe that "höher" is often embedded in a second frame which adds information about the CAUSE/AGENT, as in (8). The aligned English sentence, shown in (7), evokes CCOSP.

(7) . . . though world trade can of course increase prosperity.

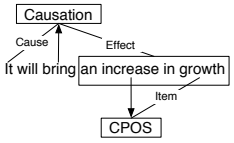
(8)



(...even if world trade has higher prosperity as result)

The phenomenon of CAUSE/AGENT roles which are contributed by other frames occurs in English as well, especially with noun or past participle instances of "increase":

(9)



According to these observations, we have to reject the hypothesis that we can always find direct

matches at the level of single frames. Direct frame match can only provide a baseline model which ignores the CAUSE problem altogether.

5 New Approach: Iterative matching

Since a match of translations on a single-frame level is obviously too rigid, we instead try to identify correspondences between larger frame structures: cross-linguistic frame paraphrases in the form of pairs of *frame groups*. For example, we would like to identify the pair of CAUSATION and CPOS in (8) as a frame paraphrase of the CCOSP in (7).

A simple strategy would be to consider each combination of CPOS and an embedding frame, as in (8), as a frame paraphrase of CCOSP; but this strategy runs into the problem of overgeneration. For example, it would identify the combination of CPOS (FEE "höhere") and OBJECTIVE_INFLUENCE (FEE "profitieren") in (11) as a paraphrase of the CCOSP in (10). This is not a valid match since in (11) the CPOS is not being caused, rather it is described as the cause of the influence.

(10) The market must be made more profitable for enterprises by increasing their exports.

(11) Die Betriebe sollten durch höhere Exporte mehr von dem Binnenmarkt profitieren.
(Enterprises should profit from the market through higher exports.)

We avoid overgeneration by explicitly aligning the semantic roles of two frame groups in a cross-lingual sentence pair using word alignments, and consider two frame groups as paraphrases if the total semantic material covered is the same in both languages, even if its distribution across roles is different. Our intuition is that this simple "rearrangement" model will cover many of the local changes due to language-specific realisation preferences.

Definitions. To avoid confusion, we will need to distinguish between a *frame* (meaning a frame type) and a *frame instance*, the occurrence of a frame in a sentence. The same holds true for *frame groups* and *frame group instances*: We write $\overline{f}$ for a frame group (type) and $\overline{r_1}, \overline{r_2}, \ldots$ for its role (types); on the token side, we use f for the corresponding frame group instance and $r_1, r_2, \ldots$ for its role instances.

For the moment, we will only look at simple frame groups, which we formalise as follows: A **frame group** $\overline{f}$ consists of either one or two frames. It has one **base frame**, $\mathrm{base}(\overline{f})$. If it has two frames, the second one is called the **embedding frame** and has a role designated as the **embedding role**. All roles of a frame group but the embedding role are called **free**.

A sentence s contains a **frame group instance f** of a frame group $\overline{f}$ if the following holds:

- There is an instance of base($\overline{f}$) in s, which we write as **base**($\mathbf{f}$).
- If $\overline{f}$ has an embedding frame, there is an instance of it in s, and the semantic head[2] of the embedding role either is the FEE of the base, or has the base FEE as a modifier.

We call two frame group instances **aligned** if the FEEs of their base frames are aligned.

As an example, consider the frame group with CPOS as base, CAUSATION as embedding frame and EFFECT as embedding role. There are instances of both frames in (8). "höher" modifies "Wohlstand", which is the headword of the EFFECT role. So (8) contains an instance of the frame group.

We describe correspondences between the roles of two frame groups by **role mappings**: Writing roles($\overline{f}$) for the set of free roles of a frame group $\overline{f}$, a role mapping $m : \mathrm{roles}(\overline{f_1}) \nrightarrow \mathrm{roles}(\overline{f_2})$ is a partial mapping from roles of a frame group $\overline{f_1}$ to roles of a frame group $\overline{f_2}$.

Given two aligned frame group instances f_1, f_2, the role mapping $m : \mathrm{roles}(\overline{f_1}) \nrightarrow \mathrm{roles}(\overline{f_2})$ **read off from word alignment of $\mathbf{f_1}$ to $\mathbf{f_2}$** maps a role $\overline{r_1}$ to a role $\overline{r_2}$ iff the head words of r_1 and r_2 in the sentence are word-aligned.

For example, in (7) the CCOSP frame has two roles, ITEM and CAUSE, while CPOS in (8) has one role, ITEM. However, the CAUSATION frame contributes a CAUSE role filled by the translation of the English CAUSE role. So the role mapping read off from CCOSP in (7) and from CPOS plus CAUSATION in (8) maps ITEM to ITEM and maps the CAUSE of CCOSP to the CAUSE of CAUSATION.

[2]Syntactic and semantic head can differ in the case of a transparent noun: In " he drank a pint of milk", the syntactic head of the Ingestible is "pint", a transparent noun, while the semantic head is "milk".

1: Given: target frame group $\overline{f_t}$, a set rel $\subseteq$ $roles(\overline{f_t})$ of relevant roles
2: Given: a set B of base frames with mappings m_t
3: Set the set of paraphrases $P = \{\overline{f_t}\}; m_t(\overline{f_t}) = \{(\overline{r}, \overline{r}) \mid \overline{r} \in \text{rel}\}$
4: **while** P changes **do**
5: **for** aligned frame group instances f_1,f_2 **do**
6: **if** $\overline{f_1} \in P$ and $\text{base}(\overline{f_2}) \in B$ **then**
7: Read off a role mapping m from the word alignment of f_1 to f_2.
8: **if** $range(m_t(\overline{f_1})) \subseteq dom(m)$ and $m_t(\overline{f_1}) \circ m$ and $m_t(\text{base}(\overline{f_2}))$ coincide on the intersection of their domains **then**
9: $P = P \cup \{\overline{f_2}\}$, where $m_t(\overline{f_2}) = m(\overline{f_1}) \circ m$
10: **end if**
11: **end if**
12: **end for**
13: **end while**

Figure 1: Incremental frame paraphrase acquisition

The algorithm we are about to present will start out with a frame group $\overline{f_t}$, the **target frame group**, for which frame paraphrases are to be derived. We will use role mappings $m_t(\overline{f}) : \text{roles}(\overline{f_t}) \nrightarrow \text{roles}(\overline{f})$ from the roles of the target frame group $\overline{f_t}$ to the roles of frame groups $\overline{f}$.

In the experiment of Section 6, the target frame group will be CCOSP, i.e. we will try to determine frame paraphrases of that frame.

The Algorithm. The Algorithm, shown in Figure 1, starts out with a target frame group $\overline{f_t}$, which may be a single frame or a two-frame group, for which we want to find frame paraphrases. It assumes that we have a set of *base frames* that have already been identified as partial paraphrases of $\overline{f_t}$, i.e. they talk about the same "basic event", but do not fill all of the *relevant roles* of $\overline{f_t}$. The algorithm iteratively extends a set of known frame paraphrases of $\overline{f_t}$, along with role mappings linking paraphrase roles to the target roles.

The algorithm identifies pairs of aligned frame groups instances: one (f_1) which is already a known paraphrase of $\overline{f_t}$, and one (f_2) which isn't, but is based on a partial paraphrase of $\overline{f_t}$. The frame group $\overline{f_2}$ counts as a new frame paraphrase of $\overline{f_t}$ if two conditions hold: First, f_2 must realize all the roles that f_1 does. Second, we get the role mapping between the target $\overline{f_t}$ and the new group $\overline{f_2}$ by extending the $\overline{f_t}$-$\overline{f_1}$ role mapping through the word alignment of f_1 and f_2 – but this new mapping must consistently extend the mapping we already have for the partial paraphrase $\overline{\text{base}(f_2)}$.

6 Experiment 2: Applying iterative matching

We now apply the algorithm from Figure 1 to find paraphrases of CCOSP.

Initialisation. The target frame group is f_t = CCOSP, with relevant role set rel = { ITEM, CAUSE[3]}. (This set can be determined automatically as the set of roles typically found in the corpus instances of "increase" in the CCOSP reading.)

We have two base frames: CCOSP itself with the mapping {(ITEM, ITEM), (CAUSE, CAUSE)}, and the mapping found in Experiment 1: CPOS with the mapping {(ITEM, ITEM)}.

Iteration 1. We find new paraphrases only on the German side because the one paraphrase which is available for matching, CCOSP, occurs solely on the English side. We can identify 10 embedding frame types, of which 3 occur more than once (CAUSATION, GIVING, REQUIREMENTS).

Iteration 2. In the second iteration we find paraphrases on either side, which allows us to identify four more frame paraphrases. We now have 9 embedding frames which occur more than once (the additional ones are CAUSAL_CONNECTION*, COMMERCE_PAY, DECIDING, LIKELIHOOD, MEANS*, AND REQUEST) and 4 which occur once.

[3]For this experiment, we are conflating the roles Cause and Agent, which both describe the Cause but differ in whether a person or a force causes the change.

Cause	English	German	Freq.
0+0	CPOS	CPOS	45
	CCOSP n.c.	CPOS	20
1+1	CPOS FG	CPOS FG	22
	CCOSP	CPOS FG	14
	CCOSP FG	CPOS FG	9
1+0	CPOS FG	CPOS	4
	CCOSP	CPOS	2
	CCOSP FG	CPOS	3
0+1	CPOS	CPOS FG	2
	CCOSP n.c.	CPOS FG	1

Table 3: Cross-lingual frame (mis-)match in detail. FG: as base frame of frame group; n.c.: CAUSE not instantiated

Iteration 3. No new frame paraphrases are found; the algorithm has reached a fixpoint for P. However, we find some additional instances of known paraphrases, which leaves us with 10 frames attested more than once (new: PURPOSE), and 3 with one attestation.

Quantitative evaluation. Table 3 breaks down the data from Table 2 in terms of cross-lingual frame (mis-)matches. Since the identification of paraphrases for the current dataset hinges on the existence of a CAUSE frame element, we organise the table according on the existence of CAUSEs on the English and/or the German side.

The first group of rows (0+0) covers examples where a CAUSE exists neither in English nor in German. Slightly more than half of all sentences (65 of 122) fall into this category. About two thirds are matching CPOS cases, but one third is made up by instances with a CAUSE-less CCOSP on the English side; these are English infinitive and participle constructions without subject. Our algorithm ignores these instances since they do not have a CAUSE role.

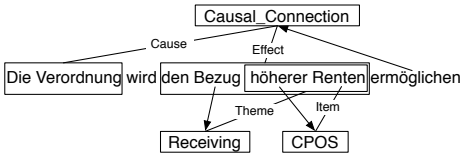
The second group of rows (1+1) contains examples where a CAUSE role exists in both languages (45, i.e. about 40%). These are the instances that our algorithm can maximally cover since it requires the existence of a CAUSE role on either side. In German, the CAUSE is always contributed by an embedding frame, since "höher" cannot introduce one itself; in English, the CAUSE is realised either directly in CCOSP, or is provided by an embedding frame. In 39 of the 1+1 instances, our algorithm established a paraphrase by identifying directly matching CAUSE roles. In the remaining 6 instances the CAUSE roles are not directly word-aligned, but have to be recovered by anaphora resolution.

The lower half of the table of our rows show the 12 cases (10%) with a CAUSE role on one side only. 11 of these are instances of proper cross-lingual divergence: In 8 cases, an English CAUSE does not have a German counterpart, and in 3 cases only the German side has a CAUSE. We interpret this asymmetry as a slightly stronger preference in German to conceptualise situations as events without an explicit CAUSE (e.g. "es entsteht X" - "X arises").

The last remaining instance shows the limits of the algorithm in its current form:

(12) This regulation will increase the size of pensions.

(13)



(The regulation will enable the drawing of higher pensions.)

Here, the identification of a paraphrase fails since the FEE of the base frame (CPOS) is not the head word of the EFFECT role, but its complement; the head word itself ("Bezug" - drawing) evokes another frame. While nothing in our algorithm precludes its generalisation to allow for frame groups consisting of more than two frames, this requires a more general definition of embedding (see Sec. 7).

Qualitative evaluation. Table 4 lists the (embedding frames of) frame paraphrases of CCOSP that our algorithm has identified. To evaluate how well these frames express causation, we compare them against a cognitive semantic account of causativity in Talmy (2000). Talmy's *basic causative situation* – one event results from another event – is expressed by the CAUSATION frame, the most frequent embedding frame in our corpus sample. Five of the frames in Table 4 are related to Talmy's *agentive causation*: The frame ACHIEVING* (an AGENT achieves a goal, the EFFECT) is just *agentive causation*; ATTEMPT_SUASION and REQUEST can be classified as *caused agency*; PURPOSE and ALLOTMENT*

(some THEME is allotted to an intended RECIPIENT through some ALLOTMENT_EVENT) can be grouped under *purpose and uncertain fulfilment*.

The frames MEANS* (something is a MEANS for achieving an EFFECT), REQUIREMENTS and DECIDING do not fall into any of Talmy's groups, but could maybe be seen as similar to the *purpose and uncertain fulfilment* class, with the difference that what is described is just an intention without a following causing event.

One of the most interesting frames in Table 4 is CAUSAL_CONNECTION* (a CAUSE contributes to an EFFECT). Talmy notes (p. 544) that "one of the more significant issues wanting attention pertains to the existence of gradience in causative concepts", but does not offer a detailed analysis. One possible use of the algorithm that we are proposing is that it offers the means for a corpus-based study of this gradient causativity: Table 5 shows the German and English expressions evoking CAUSAL_CONNECTION* that we found in our sample.

Three frames of Table 4 remain to be explained. Two of them can express the CAUSE of a scalar change in specific contexts: GIVING as in " DONOR gives RECIPIENT [increased] THEME", and LIKELIHOOD as in "Under certain CONDITIONS the DEGREE of likelihood of something is increased". The last one, COMMERCE_PAY, cannot offer good CAUSE roles and arises through free translations.

Frame	Cause FE	Embedding FE
Achieving*	Agent	Effect
Allotment*	Allotment_event	Theme
Attempt_suasion	Addressee	Content
Causal_connection*	Cause	Effect
Causation	Cause	Effect
Commerce_pay	Buyer	Money
Deciding	Cognizer	Decision
Giving	Donor	Theme
Likelihood	Conditions	Degree
Means*	Means	Effect
Purpose	Means	Goal
Request	Speaker	Message
Requirements	Dependent	Requirement

Table 4: Frame paraphrases for CCOSP identified by the algorithm

C helps to increase I; I is not unrelated to increasing C; C means increasing I
höheres I hat mit C zu tun (higher I is related to C); das mit C verbundene höhere I (the higher I related to C); C gewährleistet höheres I (C ensures higher I); C bedeutet höheres I (C means higher I)

Table 5: List of identified constructions evoking CAUSAL_CONNECTION (C = CAUSE; I = ITEM)

7 Summary and Discussion

Contributions. We have performed a manual pilot study investigating the use of frame semantics for parallel semantic analysis of aligned text. We have presented an algorithm which derives frame paraphrases (which are matched semantic substructures rather than word sequences) from a bilingual corpus. The algorithm provides a means of identifying and systematically studying cross-lingual mismatches in the distribution of semantic material (see Table 3).

While we have applied this algorithm only manually to a small dataset, the study has uncovered several new results. First, the application of the algorithm to the translation pair "increase-höher" offers a corpus-based view on expressions on a continuum between causation and non-causation and on the conditions of their use. Second, the study has provided a first impression of the degree of cross-lingual uniformity of frame-semantic structure.

Related work. Existing work on paraphrase identification has mostly focused on practical methods that exploit distributional evidence on different linguistic levels, from bag-of-words context to syntactic structures. Most studies have used monolingual data, except Bannard and Callison-Burch (2005), who also use bilingual data. Our study is situated at a deeper, semantic level of analysis, assessing the potential of using semantic structures for deriving paraphrases by means of structural matching.

Another related area is transfer-based machine translation, such as Dorna and Emele (1996), who model translations as relations between sets of semantic predicates in the source and target languages. Our aim is more modest: instead of identifying complete transfer rules on the logical level, we want to find sets of frame (i.e. predicate-argument structure) groups that can be used interchangeably.

Monolingual paraphrases. Since our matching algorithm obtains a set of equivalent frame groups, and since frame semantic representations are largely language-independent, such sets of frame groups

(e.g. Table 4) also constitute a compact, abstract model of monolingual paraphrases with a common semantics. This model can be instantiated in a particular language by using a semantic lexicon such as the FrameNet for English. The result of this is a list of concrete constructions, such as the ones in Table 5.

Frame Decomposition. One interesting result of our study is that the data does not support a decompositional analysis of CCOSP as a combination of CPOS and CAUSATION. In the embedding frames in Table 4, we find widely differing levels of causativity, as well as different perspectives on the situation, focusing on CAUSE/AGENT, INSTIGATOR, MEANS or PURPOSE. Our conclusion is that a decompositional analysis is too constrained to model this variance, and that a data-driven approach like ours, identifying paraphrases instead of meaning components, is more appropriate.

Scaling up. In this study we have limited ourselves to a single translation pair. We are confident that our algorithm is applicable to other translation pairs, given that it works even for the problematic "increase"-"höher". However, this will require automating the algorithm.

An automatic application of our algorithm will encounter errors in word alignment, parsing, and semantic analysis. Our algorithm performs strong checks on the alignment of roles, which can be expected to filter out errors in word alignment and syntax. Therefore, we anticipate shortcomings of the frame-semantic analysis to be the most serious problem to the automation of our algorithm. These can take two forms: the assignment of wrong frames, and the unavailability of appropriate frames due to the limited FrameNet coverage.

Another theoretical question to be solved in scaling up is that our current model of embedding is a syntax-based approximation which does not generalise to larger frame groups. A general semantic notion of frame embedding is nontrivial: It will have to address ambiguities such as intersective/nonintersective or de re/de dicto. An empirically motivated, but accurate definition of embedding is topic of our ongoing work.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL-05*, Ann Arbor, MI.

Michael Dorna and Martin C. Emele. 1996. Semantic-based transfer. In *Proceedings of ECAI-96*, Budapest, Hungary.

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

Charles Fillmore. 1982. Frame Semantics. *Linguistics in the Morning Calm*.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Draft.

Alessandro Moschitti, Paul Morarescu, and Sanda Harabagiu. 2003. Open-domain information extraction via automatic semantic labeling. In *Proceedings of the 14th Florida AI Conference*, pages 397–401, St. Augustine, FL.

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Sebastian Padó and Mirella Lapata. 2005. Cross-lingual bootstrapping for semantic lexicons. In *Proceedings of AAAI-05*, Pittsburgh, PA.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

C. Peters and M. Braschler. 2001. Cross-language system evaluation: the clef campaigns. *Journal of the American Society for Information Science and Technology*, 52(12):1067–1072.

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of EMNLP-2004*, Barcelona, Spain.

Leonard Talmy, 2000. *Towards a Cognitive Semantics*, chapter The Semantics of Causation. MIT Press, Cambridge, MA.

David Yarowsky, Grace Ngai, and Roger Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT 2001*.