

Cross-lingual variation of light verb constructions: using parallel corpora and automatic alignment for linguistic research

Tanja Samardžić

Linguistics Department

University of Geneva

Tanja.Samardzic@unige.ch

Paola Merlo

Linguistics Department

University of Geneva

Paola.Merlo@unige.ch

Abstract

Cross-lingual parallelism and small-scale language variation have recently become subject of research in both computational and theoretical linguistics. In this article, we use a parallel corpus and an automatic aligner to study English light verb constructions and their German translations. We show that parallel corpus data can provide new empirical evidence for better understanding the properties of light verbs. We also study the influence that the identified properties of light verb constructions have on the quality of their automatic alignment in a parallel corpus. We show that, even though characterised by limited compositionality, these constructions can be aligned better than fully compositional phrases, due to an interaction between the type of light verb construction and its frequency.

1 Introduction

Fine-grained contrastive studies traditionally belong to the field of applied linguistics, notably to translation and second language acquisition studies. Recently, however, interest for contrastive studies has been renewed due to developments in the general theory of language (the notion of micro-parameters (Kayne, 2000)) on the one hand, and due to advances in natural language processing based on the exploitation of parallel corpora, on the other hand.

Parallel corpora are collections of translations with explicit alignment of sentences. They are important resources for the automatic acquisition of the cross-linguistic translation equivalents that are needed for machine translation. There is also interest in using parallel corpora to automatically develop new annotated linguistic resources by projecting the annotation that already exists in one

language (usually English) (Padó, 2007; Basili et al., 2009). Such resources can be used for training systems for automatic parsing for different languages. Recently, parallel multilingual corpora have also been used to improve performance in mono-lingual tasks (Snyder et al., 2009).

For most of these applications, the aligned sentences in the parallel corpora need to be analysed into smaller units (phrases and words), which, in turn, need to be aligned. Although crucial for successful use of parallel corpora, word (and phrase) alignment is still a challenging task (Och and Ney, 2003; Collins et al., 2005; Padó, 2007).

Our research concentrates on one type of construction that needs a special treatment in the task of aligning corpora and projecting linguistic annotation from one language to another, namely light verb constructions. These constructions, usually identified as paraphrases of verbs (e.g. *have a laugh* means *laugh*, *give a talk* means *talk*), are frequent, cross-lingually productive forms, where simple-minded parallelism often breaks down. Their meaning is partially uncompositional, formed in a conventional way, which means that they cannot be analysed as regular constructions and that they cannot be translated to another language directly word by word. Unlike collocations and idioms, however, these constructions are formed according to the same “semi-productive” pattern in different languages. Due to their cross-lingual analysability, they can be expected to be aligned at the word level in a parallel corpus, even if their components are not direct word-to-word translations of each other. This means that word alignment of these constructions, needed for automatic translation and transferring annotations, is possible, but it is not straight-forward.

An in-depth study of these constructions in the specific context of parallel corpora and alignment can cast new light on the correlation of their linguistic and statistical properties. On the one hand,

the statistical large-scale analysis of the behaviour of these constructions as the output of an alignment process provides novel linguistic information, which enlarges the empirical base for the analysis of these constructions, and complements the traditional grammaticality judgements. On the other hand, the linguistically fine-grained analysis of the statistical behaviour of these constructions provides linguistically-informed performance and error analyses that can be used to improve aligners.

2 Two Types of Light Verb Constructions and their Alignment

Light verb constructions have already been identified as one of the major sources of problems in transferring semantic annotation between languages as close as English and German (Burchardt et al., 2009). Light verb constructions introduce two kinds of divergences that can pose a problem for automatic word alignment. In the case of *true light verb constructions* (Kearns, 2002), English phrases such as *have a laugh*, *give [stg.] a wipe*, and *take a look* typically correspond to German single words, *lachen*, *wischen*, and *blicken* respectively. Such correspondences can be expected to result in actual parallel sentences where English verbs *have*, *give*, and *take* would be either aligned with the verbs *lachen*, *wischen*, and *blicken* respectively or would have no alignment at all. Such alignments are not common cases and can be expected to pose a problem to an automatic aligner.

Another type of divergence concerns *constructions with vague action verbs* (Kearns, 2002). In this case, English phrases such as *make an agreement*, *make a decision*, and *give a talk* correspond to German phrases *einen Vertrag schliessen*, *eine Entscheidung treffen*, and *einen Vortrag halten*, respectively. Parallel sentences containing these constructions should be aligned so that English nouns *agreement*, *decision*, and *talk* are aligned with German nouns *Vertrag*, *Entscheidung*, and *Vortrag*. At the same time, English verb *make* should be aligned with German *schliessen* in the first example, with *treffen* in the second, and *give* should be aligned with *halten* in the third example. Aligning the nouns should not pose any problem, since these alignments are direct lexical translations (c.f. (LEO, 2006 9) online dictionary, for example) and they can be expected to be aligned in many different sentences. However, aligning the

verbs is necessarily more complicated, since they are not direct translations of each other and cannot be expected to be aligned in other contexts.¹

However, the difference between the two types of light verb constructions is not clear cut. They are better seen as two ends of a continuum of verb usages with different degrees of verbs' lightness and different degrees of compositionality of the meaning of constructions. (Stevenson et al., 2004; Butt and Geuder, 2001; Grimshaw and Mester, 1988). Even though several English verbs have been identified as having light usages (e.g. *take*, *make*, *have*, *give*, *pay*), there has been little research on the influence that the properties of the heading light verb can have on the degree of semantic compositionality of the construction.

The purpose of the present research is to examine the German translation equivalents of the range of different English light verb constructions occurring in a parallel corpus and study the differential performance of a standard aligner on this language pair for these constructions.

3 Experiments

Our study is based on the assumption that the quality and bijectivity of the alignment are proportional to the corpus frequency and linguistic compositionality of the construction. Therefore, we identify two aspects of the alignment of these constructions as the relevant objects of study.

First, we quantify the amount and nature of correct word alignments for light verb constructions compared to regular verbs, as determined by human inspection. Given the described divergences between English and German, it can be expected that light verb constructions will be aligned with a single word more often than constructions headed by a regular verb. Assuming that the properties of the heading light verbs do influence semantic compositionality of the constructions, it can also be expected that light verb constructions headed by different verbs will be differently aligned to the German translations, constituting different types of constructions.

¹Direct word-to-word English translations of *schliessen* listed in the LEO dictionary, for example, are: *infer*, *comprise*, *imply*, *close*, *close down*, *conclude*, *consummate*, *draw up*, *lock*, *shut*, *shutdown*, *sign off*, *quit*, while *make* is only listed within the phrase that is translation for this particular collocation. Similarly, English word translations for *treffen* are: *encounter*, *hook up*, *cross*, *get together*, *meet*, *meet up*, *hit*, *hurt*, *score*, *strike*, while *make* can only be found as a part of the phrase-to-phrase translations.

Second, we evaluate the quality of automatic word alignments of light verb constructions.

Current word alignment models are based on the assumption that the best word alignments are composed of the best word-to-word translations (as an effect of using Expectation-Maximisation for training). Factors in the translations that deviate from one-to-one alignments are often lexically specific (fertility) and require sufficient statistics. Because of the interaction of these properties of the alignment model and the semi-compositionality of light verb constructions, these constructions can be expected to pose a problem for automatic word alignment. Specifically, we expect lower overall quality of word alignment in the sentences containing light verb constructions than in the sentences that contain corresponding regular constructions.

As indicated, however, we also expect that the quality of automatic word alignment will be influenced by different distributional phenomena that are not necessarily related to the linguistic properties of parallel texts, in particular related to frequency of some of the components of the construction.

These predictions about the alignment of light verb constructions in English and German and their realisations in a corpus are examined in an experiment.

3.1 Materials and Methods

A random sample of instances of each of the defined types of construction was extracted from a large word-aligned parallel corpus and manually examined.

3.1.1 Corpus

The instances of the phrases were taken from the English-German portion of the Europarl corpus (Koehn, 2005) that contains the proceedings of the sessions held in 1999, irrespective of the source language and of the direction of translation. Before sampling, the corpus was word-aligned using GIZA++ (Och and Ney, 2003). Alignments were performed in both directions, with German as the target language and with English as the target language.

3.1.2 Word alignment using GIZA++

The program for automatic word alignment, GIZA++, has been developed within a system for automatic translation. It implements a series of

statistical word-based translation models. In these models, word alignment is represented as a single-valued function, mapping each word in the target sentence to one word in the source sentence. To account for the fact that some target language words cannot be aligned with any source language word, a special empty word (“NULL”) is introduced in the source sentence.

The definition of word alignment does not allow many-to-many mappings between the words of two languages, needed for representing alignment of non-compositional multi-word expressions. However, it allows aligning multiple words in one language to a single word in the other language, which is needed for successful alignment of English light verb constructions.

3.1.3 Sampling phrase instances

To study light verb constructions in a parallel corpus systematically, we group the instances of the constructions into two types: light verb constructions headed by the verb *take*, as an example of true light verb constructions, and those headed by the verb *make*, as an example of vague action verbs. We compare both types of light verb constructions to regular constructions headed by the verbs which are WordNet synonyms of the verb *make* (*create*, *produce*, *draw*, *fix*, *(re)construct*, *(re)build*, *establish*) with the same subcategorization frame.

We analyse three samples of the constructions, one for each of the types defined by the heading verb. Each sample contains 100 instances randomly selected from the word-aligned parallel corpus. The constructions are represented as ordered pairs of words, where the first word is the verb that heads the construction and the second is the noun that heads the verb’s complement. Only the constructions where the complement is the direct object were included in the analysis.²

3.1.4 Data collection

The following data were collected for each occurrence of the English word pairs.

The word or words in the German sentence that are actual translation of the English words were identified. If either the English or German verb

²This means that constructions such as *take something into consideration* were not included. The only exception to this were the instances of the construction *take something into account*. This construction was included because it is used as a variation of *take account of something* with the same translations to German.

form included auxiliary verbs or modals, these were not considered. Only the lexical part of the forms were regarded as word translations.

We then determine the type of mapping between the translations. If the German translation of an English word pair includes two words too (e.g. take+decision $\leftrightarrow$ Beschluss+fassen), this was marked as the “2-2” type. If German translation is a single word, the mapping was marked with “2-1”. This type of alignment is further distinguished into “2-1N” and “2-1V”. In the first subtype, the English construction corresponds to a German noun (e.g. initiative+taken $\leftrightarrow$ Initiative). In the second subtype, the English construction corresponds to a German verb (e.g. take+look $\leftrightarrow$ anschauen). In the cases where a translation shift occurs so that no translation can be found, the mapping is marked with “2-0”.

We also collect the information on automatic alignment for each element of the English word pair for both alignment directions. These data were collected for the elements of English word pairs (verbs and nouns) separately. The alignment was assessed as “good” if the word was aligned with its actual translation, as “bad” if the word was aligned with some other word, and as “no align” if no alignment was found. Note that the “no align” label could only occur in the setting where English was the source language, since all the words in the sentence had to be aligned in the case where it was the target language.

For example, a record of an occurrence of the English construction “make+proposal” extracted from the bi-sentence in (1)³ would contain the information given in (2).

- (1) Target language German
 EN: *He made a proposal.*
 DE: *Er(1) hat(1) einen(3) Vorschlag(4) gemacht(3).*

Target language English
 DE: *Er hat einen Vorschlag gemacht.*
 EN: *He(1) made(5) a(3) proposal(4).*

- (2) English instance: made + proposal
 German alignment: Vorschlag + gemacht
 Type of mapping: 2-2

³Glosses:

Er hat einen Vorschlag gemacht.
 he has a proposal made
 The numbers in the brackets in the target sentences indicate the position of the automatically aligned source word.

		English		
		LVC take	LVC make	Regular
German translation	2 $\rightarrow$ 2	57	50	94
	2 $\rightarrow$ 1N	8	18	2
	2 $\rightarrow$ 1V	30	28	2
	2 $\rightarrow$ 0	5	4	2
Total		100	100	100

Table 1: Types of mapping between English constructions and their translation equivalents in German.

Automatic alignment, target German, noun:
 good, verb: no align

Automatic alignment, target English, noun:
 good, verb: good

4 Results

In this section, we present the results of the analyses of both correct (manual) and automatic alignment of the three types of constructions, pointing out the relevant asymmetries.

4.1 Results of Manual Alignment

Table 1 shows how many times each of the four types of mapping (2-2; 2-1N; 2-1V; 2-0) between English constructions and their German translation equivalents occurs in the sample.

We can see that the three types of constructions tend to be mapped to their German equivalents in different ways. First, both types of light verb constructions are mapped to a single German word much more often than the regular constructions (38 instances of light verb constructions with *take* and 46 instances of light verb constructions with *make* vs. only 4 instances of regular constructions.). Confirming our initial hypothesis, this result suggests that the difference between fully compositional phrases and light verb constructions in English can be described in terms of the degree of the “2-1” mapping to German translation equivalents.

An asymmetry can be observed concerning the two subtypes of the “2-1” mapping too. The German equivalent of an English construction is more often a verb if the construction is headed by the verb *take* (in 30 occurrences, that is 79% of the 2-1 cases) than if the construction is headed by the verb *make* (28 occurrences, 61% cases).

		DE	EN
LVCs with <i>take</i>	Both EN words	5	57
	EN noun	63	79
	EN verb	6	57
LVCs with <i>make</i>	Both EN words	5	40
	EN noun	58	58
	EN verb	6	52
Regular construction	Both EN words	26	42
	EN noun	68	81
	EN verb	32	47

Table 2: Well-aligned instances of LVCs with *take*, with *make*, and with regular constructions (out of 100), produced by an automatic alignment, in both alignment directions (target is indicated).

In the case where the German translation equivalent for an English construction is a verb, both components of the English construction are included in the corresponding German verb, the verbal category of the light verb and the lexical content of the nominal complement. These instances are less compositional, more specific and idiomatic (e.g. *take+care* $\leftrightarrow$ *kümmern*, *take+notice* $\leftrightarrow$ *berücksichtigen*).

On the other hand, English constructions that correspond to a German noun are more compositional, less idiomatic and closer to the regular verb usages (e.g. *make+proposal* $\leftrightarrow$ *Vorschlag*, *make+changes* $\leftrightarrow$ *Korrekturen*). The noun that is regarded as their German translation equivalent is, in fact, the equivalent of the nominal part of the construction, while the verbal part is simply omitted. This result suggests that English light verb constructions with *take* are less compositional than the light verb constructions with *make*.

4.2 Results on Automatic Alignment

We evaluate the quality of automatic alignment of light verb constructions in comparison with regular phrases taking into account two factors, the alignment direction and the frequency of the elements of the constructions. The results are presented in the next two sections.

4.2.1 Direction of Alignment

Table 2 shows how the quality of automatic alignment varies depending on the direction of alignment, as well as on the type of construction. Recall that more than one target word can be aligned to the same source word and all words of the target have to be aligned.

It can be noted that all the three types of constructions are better aligned if the target language is English. However, the difference in the quality is bigger in light verb constructions than in regular constructions, clearly because in this direction the multi-word property of the English light verb constructions can be represented. Both words are well aligned in light verb constructions with *take* in 57 cases and with *make* in 40 cases if the target language is English, which is comparable with regular constructions (42 cases). However, if the target language is German, both types of light verb constructions are aligned well (both words) in only 5 cases, while regular constructions are well aligned in 26 cases.

Looking into the alignment of the elements of the constructions (verbs and nouns) separately, we can notice that nouns are generally better aligned than verbs for all the three types of constructions, and in both directions. However, this difference is not the same in all cases. The difference in the quality of alignment of nouns and verbs is the same in both alignment directions for regular constructions, but it is more pronounced in light verb constructions if German is the target. On the other hand, if English is the target, the difference is smaller in light verb construction than in regular phrases. These results suggest that the direction of alignment influences more the alignment of verbs than the alignment of nouns in general. This influence is much stronger in light verb constructions than in regular constructions.

Finally, our initial hypothesis that the quality of alignment of light verb constructions is lower than the quality of alignment of regular constructions has only been confirmed in the case where German is the target language (both words well aligned in 26 cases, compared to only 5 cases in both types of light verb constructions). Regular verbs are especially better aligned than light verbs in this case (32 : 6). However, if the target is English, the quality of alignment of regular constructions is similar to that of light verb constructions with *make* (42 and 40 good alignments respectively), while the constructions with *take* are aligned even better than the other two types (57 good alignments). These results suggest that the type of construction which is the least compositional and the most idiomatic of the three is best aligned if the direction of alignment suits its properties.

Frequency	<i>take</i> LVC	<i>make</i> LVC	Regular
Low	12	25	62
High	76	35	8

Table 3: The three types of constructions partitioned by the frequency of the complements in the sample.

Freq		Well aligned					
		<i>take</i> LVC		<i>make</i> LVC		Regular	
Low Freq	Both	4	33	8	32	21	34
	N	8	66	8	32	47	75
	V	4	33	12	48	53	85
High Freq	Both	47	62	18	51	4	50
	N	64	84	27	77	8	100
	V	58	76	18	51	4	50

Table 4: Counts and percentages of well-aligned instances of the three types of constructions in relation with the frequency of the complements in the sample. The percentages represent the number of well-aligned instances out of the overall number of instances within one frequency range. English is the target language.

4.2.2 Frequency

Since the quality of alignment of the three types of constructions proved different from what was expected in the case where English was the target language, we examine further the automatic alignment in this direction. In particular, we study its interaction with frequency.

The frequency of the nouns is defined as the number of occurrences in the sample. It ranges from 1 to 20 occurrences in the sample of 100 instances. The instances of the constructions were divided into three frequency ranges: instances containing nouns with 1 occurrence were considered as low frequency items; those containing nouns that occurred 5 and more times in the sample were considered as high frequency items; nouns occurring 2, 3, and 4 times were considered as medium frequency items. Only low and high frequency items were considered in this analysis.

Table 3 reports the number of instances belonging to different frequency ranges. It can be noted that light verb constructions with *take* exhibit a small number of low frequency nouns. The number of low frequency nouns increases in the constructions with *make* (25/100), and it is much bigger in regular constructions (62/100). The opposite is true for high frequency nouns (LVCs

with *take*: 76/100, with *make*: 35/100, regular: 8/100). Such distribution of low/high frequency items reflects different collocational properties of the constructions. In the most idiomatic constructions (with *take*), lexical selection is rather limited which results in little variation. Verbs in regular constructions select for a wide range of different complements with little reoccurrence. Constructions with *make* can be placed between these two types.

Different trends in the quality of automatic alignment can be identified for the three types of constructions depending on the frequency range of the complement in the constructions, as shown in Table 4. The quality of alignment of both components of the constructions is comparable for all the three types of constructions in low frequency items (in 33% of instances of light verb constructions with *take*, 32% of light verb constructions with *make*, and 34% of regular constructions both the verb and the noun were well aligned). It is also improved in high frequency items in all the three types, compared to low frequency. However, the improvement is bigger in light verb constructions with *take* (62% well aligned cases) than in LVCs with *make* (51%) and in regular constructions (50%).⁴

Looking into the components of the constructions separately, we can notice interesting differences in the quality of automatic alignment of verbs. The proportion of well-aligned verbs increases with the frequency of their complements in light verb constructions with *take* (33% of low frequency items compared to 76% of high frequency items.) It stays almost the same in light verb constructions with *make* (48% of low frequency items and 51% of high frequency items), and it even decreases in regular items (85% of low frequency items compared to only 50% of high frequency items).

5 Discussion

The results reported in the previous section confirm both components of our first hypothesis (on the expected differences in cross-lingual mapping) and refine the conditions under which the second hypothesis (on the expected differences in the quality of automatic alignment) is true. We discuss

⁴Note that the high frequency regular items are represented with only 8 instances, which is why the trends might not be clear enough for this subtype.

these conclusions in detail here.

5.1 Manual Alignment

Recall that the first component of our first hypothesis indicated that it is expected that light verb constructions will be aligned with a single word more often than constructions headed by a regular verb.

The analysis of corpus data has shown that there is a clear difference between English regular phrases and light verb constructions in the way they are mapped to their translation equivalents in German. Regular constructions are mapped word-by-word, with the English verb being mapped to the German verb, and the English noun to the German noun. A closer look into the only 4 examples where regular constructions were mapped as “2-1” shows that this mapping is not due to the “lightness” of the verb. In two of these cases, it is the content of the verb that is translated, not that of the noun (produce+goods $\leftrightarrow$ Produktion; establishes+rights $\leftrightarrow$ legt). This never happens in light verb constructions.

On the other hand, light verb constructions are much more often translated with a single German word. In both subtypes of the “2-1” mapping of light verb constructions, it is the content of the nominal complement that is translated, not that of the verb. The noun is either transformed into a verb (take+look $\leftrightarrow$ anschauen) or it is translated directly with the verb being omitted (take+initiative $\leftrightarrow$ Initiative).

This difference provides empirical grounds for distinguishing between semantically full and semantically impoverished verbs, a task that is often difficult on the basis of syntactic tests, since they often exhibit the same syntactic properties.

The second component of the first hypothesis indicated that it was expected that the two types of light verb constructions be differently aligned.

The finding that English light verb constructions with *take* tend to be aligned more often with a single German verb and less often to a single German noun than the constructions with *make* justifies classifying the instances into the types based on the heading verb, which is not a common practice in the linguistic literature. It suggests that some semantic or lexical properties of these verbs can determine the type of the construction. More precisely, the meaning of the constructions with *take* can be regarded as less compositional than the

meaning of the constructions with *make*. This difference is also supported by the findings of a preliminary study of Serbian translation equivalents of these constructions (Samardžić, 2008). English constructions with *take* tend to be translated with a single verb in Serbian, while the constructions with *make* are usually translated word-by-word.⁵

5.2 Automatic alignment

The second hypothesis conjectured that we would find lower overall quality of word alignment in the sentences containing light verb constructions than in the sentences that contain corresponding regular constructions. The findings of this research show that the interactions between alignment and types of constructions is actually more complicated than this simple hypothesis, in some expected and some unexpected ways. To summarise, we found, first, better alignment of regular constructions compared to light verb constructions only if the target language is German; second, overall, alignment if English is target is better than if German is target; and thirdly, we found a clear frequency by construction interaction in the quality of alignment.

The quality of automatic alignment of both regular constructions and light verb constructions interacts with the direction of alignment. First, the alignment is considerably better if the target language is English than if it is German, which confirms the findings of (Och and Ney, 2003). Second, the expected difference in the quality of alignment between regular constructions and light verb constructions has only been found in the direction of alignment with German as the target language, that is where the “2-1” mapping is excluded. However, the overall quality of alignment in this direction is lower than in the other.

This result could be expected, given the general morphological properties of the two languages, as well as the formalisation of the notion of word alignment used in the system for automatic alignment. According to this definition, multiple words in the target language sentence can be aligned with a single word in the source language sentence, but not the other way around. Since English is

⁵The difference in the level of semantic compositionality of the constructions with *take* and *make* could follow from some semantic properties of these verbs, such as different aspectual properties or argument structures. However, establishing such a relation would require a more systematic semantic study of light, as well as full lexical uses of these verbs.

a morphologically more analytical language than German, multiple English words often need to be aligned with a single German word (a situation allowed if English is the target but not if German is the target).

The phrases in (3) illustrate the two most common cases of such alignments. First, English tends to use functional words (the preposition *of* in (3a)), where German applies inflection (genitive suffixes on the article *des* and on the noun *Bananensektors* in (3b)). Second, compounds are regarded as multiple words in English (*banana sector*), while they are single words in German (*Bananensektors*). This asymmetry explains both the fact that automatic alignment of all the three types of constructions is better when the target language is English and that the alignment of light verb constructions is worse than the alignment of regular phrases when it is forced to be expressed as one-to-one mapping, which occurs when German is the alignment target.

- (3) a. the infrastructure of the banana sector
- b. die Infrastruktur des Bananensektors

Practically, all these factors need to be taken into consideration in deciding which version of alignment should be taken, be it for evaluation or for application in other tasks such as automatic translation or annotation projection. The intersection of the two directions has been proved to provide most reliable automatic alignment (Padó, 2007; Och and Ney, 2003). However, it excludes, by definition, all the cases of potentially useful good alignments that are only possible in one direction of alignment.

Linguistically, the fact that the expected difference in the quality of alignment between regular constructions and light verb constructions has only been found in the direction where English constructions could not be aligned with single German words can be seen as another empirical indication of semantic impoverishment of light verbs in comparison with full lexical verbs.

Finally, we found an unexpected frequency by construction interaction (Table 4), which explains the finding that regular phrases are not better aligned than light verb constructions if English is the target language (opposite to our second hypothesis). This interaction, well known in language processing and acquisition, occurs in those cases where marked constructions are very frequent. In our case, the marked construction is the

semi-compositional light verb construction with *take*, which has frequent noun complements. In this case, despite the non-regularity of the construction, alignment is performed well if the direction of alignment allows its mapping to a single word. Also, with respect to this phenomenon, the constructions with *take* behave more markedly than those with *make*.

What is especially interesting about these data is the fact that the alignment is different not just between light verb constructions and regular constructions, but also between the two types of light verb constructions. The constructions with *take* exhibit more consistent properties of irregular items, while the constructions with *make* can be positioned somewhere between irregular and regular items. This additionally confirms the claim that these two types of constructions differ in the level of semantic compositionality, providing a basis for an improvement in their linguistic account.

6 Conclusions and Future Work

In this paper we have proposed a contrastive study of light verb constructions based on data collected through alignments of parallel corpora. We have shown how a linguistically refined analysis can shed light on particularly difficult cases for an alignment program, a useful result for improving current statistical machine translation systems. We have also shown how properties and behaviours of these constructions that can be found only in large parallel corpora and through sophisticated computational tools can shed light on the linguistic nature of the constructions under study.

Much remains to be done, both in this general methodology and for this particular kind of construction. As an example, we note that the fact that nouns are aligned better than verbs in all the three types of constructions deserves more investigation. What we do not yet know is whether this fact can be related to some known distributional differences between these two classes or not. It might also mean that nominal lexical items are more stable across languages than verbal ones. This can have implications for machine translations, as well as for annotation projection, since the stable words can be used as pivots for alignment and transfer algorithms.

References

- Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti. 2009. Cross-language frame semantics transfer in bilingual corpora. In Alexander F. Gelbukh, editor, *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 332–345, Mexico City, Mexico. Springer.
- Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2009. FrameNet for the semantic analysis of German: Annotation, representation and automation. In Hans Boas, editor, *Multilingual FrameNets in Computational Lexicography: Methods and applications*, pages 209–244. Mouton de Gruyter.
- Miriam Butt and Wilhelm Geuder. 2001. On the (semi)lexical status of light verbs. In Norbert Corver and Henk van Riemsdijk, editors, *Semilexical Categories: On the content of function words and the function of content words*, pages 323–370, Berlin. Mouton de Gruyter.
- Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 531–540, Ann Arbor. Association for Computational Linguistics.
- Jane Grimshaw and Armin Mester. 1988. Light verbs and theta-marking. *Linguistic Inquiry*, 19:205–232.
- Richard Kayne. 2000. *Parameters and Universals*. Oxford University Press, New York.
- Kate Kearns. 2002. Light verbs in English. Manuscript.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*, Phuket, Thailand.
- LEO. 2006-9. *LEO Online Dictionary*. LEO Dictionary Team, <http://dict.leo.org>.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.
- Sebastian Padó. 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Saarland University.
- Tanja Samardžić. 2008. Light verb constructions in English and Serbian. In *English Language and Literature Studies — Structures across Cultures*, pages 59–73, Belgrade. Faculty of Philology.
- Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2009. Adding more languages improves unsupervised multilingual part-of-speech tagging: a Bayesian non-parametric approach. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 83–91, Boulder, Colorado, June. Association for Computational Linguistics.
- Suzanne Stevenson, Afsaneh Fazly, and Ryan North. 2004. Statistical measures of the semiproductivity of light verb constructions. In *Proceedings of the ACL04 Workshop on Multiword Expressions: Integrating Processing*, pages 1–8. Association for Computational Linguistics.