Querying multilevel annotation and alignment for detecting grammatical valence divergencies

Oliver Čulo
FTSK, Universität Mainz
An der Hochschule 2, 76726 Germersheim
E-mail: culo@uni-mainz.de

Abstract

The valence concept has been used in machine translation as well as didactics on order to build up valence dictionaries for the respective uses. Most valence dictionaries have been built up manually, but given the growing number of parallel resources, it would be desirable to automatically exploit them as basis for building up bilingual valence dictionaries. The present contribution reports on a pilot study on a German-English parallel corpus. In this study, patterns of verb plus grammatical functions were extracted from parallel sentences. The paper reports on some of the basic findings of this extraction, regarding divergencies both in valence patterns as well as syntactic realisations of the predicate, i.e. the verb. These findings set the agenda for further research, which should focus on how to detect semantic shifts of valence carriers in translation and how this affects valence.

Keywords: valence, valence extraction, parallel corpora, translation

1. Introduction

The concept of valence (Tesnière 1959) has been endorsed in multilingual research domains in various ways. Various machine translation systems use some notion of valence in the core of their analysis and transfer structures (see relevant descriptions e.g. for EUROTRA (Steiner, Schmidt, and Zelinsky-Wibbelt 1988), METAL (Gebruers 1988), Verbmobil (Emele et al. 2000) or TectoMT (Žabokrtský, Ptáček, and Pajas 2008)). For didactic purposes, various bilingual valence dictionaries have been compiled (D. Rall, Rall, and Zorrilla 1980; Engel and Savin 1983; Bianco 1996; Simon-Vandenbergen, Taeldeman, and Willems 1996).

Most of the valence resources mentioned are based on manually compiled valence dictionaries. Nowadays, as ever more and larger parallel corpus resources are available, it is desirable to exploit these in order to gain more data for bilingual valence dictionary creation. There have been various attempts at extracting bilingual valence dictionaries from parallel corpora. In some cases, the extraction process is tackled from a high-level semantic level, as in the case of bilingual frame semantic dictionaries (Boas 2002; 2005). Other approaches choose a syntactic annotation, as in the case of the Prague Czech-English Dependency Treebank (Čmejrek et al. 2004). In both cases, the semantic or „deep“ dependency (or *tectogrammatical*, see (Sgall, Hajičová, and Panevová 1986)) annotation abstracts away from syntactic variation, making the extraction task somewhat less complex. In the course of the FUSE-project (Cyrus 2006), predicate-argument annotation and alignment between German and English sentences serves as basis for the study of both syntactic and semantic valence divergencies. Padó (2007) investigates the (frame) semantic dimension of valence divergencies. In the former case, the annotation is very specifically tailored to the project itself, making the methods harder to reproduce when applied to other corpora. In the latter study, the level of investigation again abstracts away from syntactic variation.

The study presented here focusses on grammatical differences in valence pattern between German and English. Both for the detection and description of differences, top-level grammatical function like subject,

direct object etc. are used. This follows the tradition of using grammatical functions rather than syntactic categories as e.g. in the previously listed bilingual valence dictionaries. Grammatical functions abstract away from syntactic variation but as compared to e.g. the tectogrammatical approach of (Čmejrek et al. 2004), no deep annotation is needed in order to retrieve grammatical functions of a sentence.

The corpus used in the study is annotated and aligned on multiple linguistic levels, but not with a specific focus on valence. Also, the method of querying multiple annotation and alignment levels at once is outlined. On top of that, valence divergencies are discussed with respect to factors like contrastive differences, register or translation properties and strategies.

2. Study setup

2.1. The corpus

The corpus used in the study was built to investigate contrastive commonalities and differences between English and German as well as peculiarities in translations. It consists of English originals (EO), their German translations (GTrans) as well as German originals (GO) and their English translations (ETrans). Both translation directions are represented in eight registers with at least 10 texts totalling 31,250 words per register. In the present paper, examples are taken from the registers SHARE (corporate communications), SPEECH (political speeches) and FICTION (fictional texts). Altogether, the corpus comprises one million words. Additionally, register-neutral reference corpora are included for German and English including 2,000 word samples from 17 registers.

All texts are annotated with part-of-speech information using the TnT tagger (Brants 2000), morphology using MPRO (Maas, Rösener, and Theofilidis 2009), and grammatical functions and chunk categories, manually annotated with MMAX2 (Müller and Strube 2006).

Furthermore, all texts are aligned on word level using GIZA++ (Och and Ney 2003), on chunk level indirectly by mapping the grammatical functions onto each other, on clause level manually again using MMAX2, and on sentence level using the WinAlign component of the Trados Translator's Workbench (Heyn 1996) with additional manual correction.

2.2. A format independent API for multilevel queries

The API designed for the corpus is made up of three parts. On top, there is the interface, containing control methods with basic read/write and iteration calls for the corpus. Under the hood, a package called CoReTool is used to represent linguistic structures in stratified layers, and the parallel structures (e.g. aligned words, sentences, etc.) as sets of pairs. The intermediate level handles the XML-based data format of the corpus. Queries are mainly written using the format-independent CoReTool data structures and are thus re-usable for other corpora as well. The layers dealing with corpus management and format handling can, in theory, be exchanged depending on the corpus used. This stratificational approach is a major difference between this corpus API and other APIs, where programming data structures and underlying data format are more closely linked.

Fundamental within CoReTool is the notion of Text. A Corpus is made up of an ordered collection of Texts, which again is made up of an ordered collection of Sentences, which again is made up of an ordered collection of Tokens. This structure is so to speak the backbone of CoReTool and the minimum of data that we expect in a corpus. In addition, a Corpus can be divided into Registers which also relate to collections of Texts (from the Corpus). Likewise, a Sentence can contain Clauses or Chunks which relate to the Tokens of the Sentence. For each of these sub-units of a text (including Tokens), it is possible to have aligned counterparts. Every single alignment is represented as a pair; so if unit U is aligned with U' and U'', there will be two pairs $<U,U'>$ and $<U,U''>$.

The CoReTool Java package uses simple data structures like ordered lists to organize the linguistic content it represents. In addition, a couple of basic methods for calculating statistics – e.g. numbers on chunk categories or grammatical functions – are included. The package so far lacks a proper backend-enabled design, so that IO

```
for every wordPair in wordPairs

     slWord := getSlWord(wordPair)
     tlWord := getTlWord(wordPair)
     slChunk := getChunkForWord(slWord)
     tlChunk := getChunkForWord(tlWord)

     if (not mappable(getGramFunc(slChunk), getGramFunc(tlChunk))
     then markCrossingLine(slWord, tlWord, slChunk, tlChunk)
     end if

end for
```

Figure 1: Pseudo-Code of the query for crossing lines between grammatical functions and words

methods could be easily plugged in on demand. Also, the linguistic representation of CoReTool is currently restricted to syntactic structures. However, the need to extend the package with further functionalities, e.g. in to be able to operate with semantic annotation as well, may or will hopefully soon be rendered unnecessary by latest developments of query tools like e.g. ANNIS2[1].

2.3. Querying for *empty links* and *crossing lines*

Two concepts are used to detect instances of valence divergencies. These concepts are based on well-known concepts from translation studies. Elements which have no alignment exhibit an e*mpty link*. Such 0:1-equivalents have been described e.g. by Koller (2001). Elements which are aligned, but which are embedded in higher units that are not aligned, result in *crossing lines*. This would e.g. be the case for two aligned words which are embedded in different grammatical functions. Crossing lines relate to the concept of shifts (in the given example a shift in grammatical function) as described e.g. by Catford (1965).

The corpus is queried for empty links and crossing lines using the CoReTool package. Empty links can be detected by simply querying one alignment level. For crossing lines, querying combinations of both annotation and alignment levels is necessary. A query for a shift in function requires (1) going through pairs of aligned words, (2) for each pair: getting the chunks the aligned words are embedded in, and (3) checking the mapping of these chunks, i.e. check whether the grammatical functions they've been assigned are compatible (cf. figure 1). As in this study setup the same set of grammatical functions was used for German and English, mapping was straightforward.

3. Divergencies in valence patterns for grammatical functions

The ideal situation for valence extraction from parallel corpora would be that of sentence pairs with equivalent verbs at their core and perfectly matching syntactic patterns. Minor shifts, e.g. in the type of grammatical functions governed by the verb, can easily be accounted for. However, besides differences in realisation of arguments, there may also be differences in the realisation of the predicate. Such a typical shift is the *head switch*, in examples like *Ich schwimme gern – I like swimming*, where the German adverb *gern* 'willingly, with pleasure' becomes the full verb *like* in English. As we will see, there may be other factors for different kinds of shifts in the verb. We will be looking at more semantically/pragmatically triggered shifts, for a more syntactic investigation especially of shifts in the realisation of the predicate, e.g. support verb constructions versus full verbs, see (Čulo 2010).

Probably the simplest case for a valence divergency on the level of grammatical functions is that of differences in the kinds of grammatical function as which an argument is realised. Compare, for instance, the sentence pair in figure 2, with the English original on top and the German translation at the bottom, and let us

[1]http://www.sfb632.uni-potsdam.de/d1/annis/

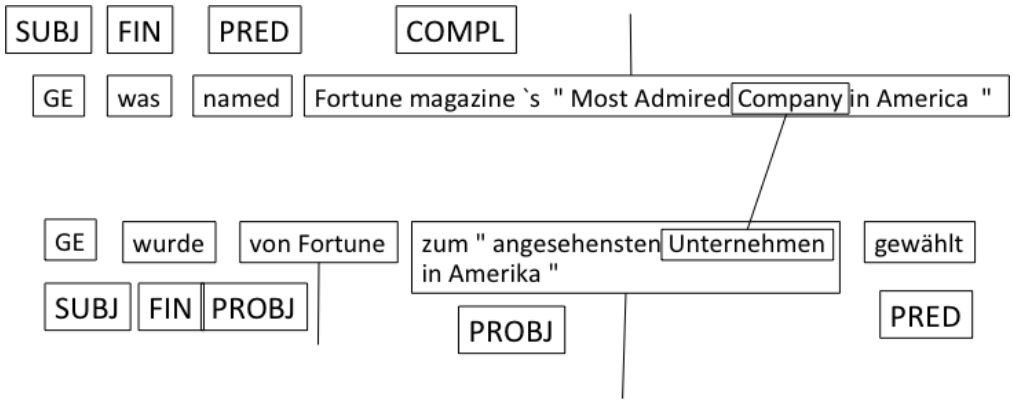

Figure 2: A crossing line for the words Company and Unternehmen and the grammatical functions COMPL and PROBJ

focus on the phrase *"Most admired Company in America"*. This phrase is embedded in a predicative complement (tag: COMPL) in English, as is governed by verbs like *name*, *appoint*, *elect* etc. The COMPL function has no equivalent in German, resulting in an empty link (indicated by the vertical lines with only linked to only one box). In order to understand, though, what is happening in that case, one has to evaluate the links from within the phrase: the word *Company*, for instance, is aligned with the equivalent word *Unternehmen* which is, however, embedded in a prepositional object (PROBJ) in German. The cause for this shift lies in a contrastive difference in the valence patterns of a whole class of verbs (namely the APPOINT class, following Levin (1993)). But, as there currently is no semantic annotation present in the corpus, there is no automatic way of linking the verb sense to this particular shift. We will come back to this point when discussing the last example.

A similar shift from COMPL to a different function is shown in figure 3. Here, however, the shift is not triggered by the fact that two equivalent verbs have different valence patterns, but by a change of the main verb which does not match known concepts like head switches.

	be → sein	*be → ~~sein~~*
E2G_SHARE	37 % (126)	63 % (215)
E2G_FICTION	45 % (138)	54 % (168)
E2G_SPEECH	60 % (224)	40 % (147)

Table 1: Proportions of be translated as either *sein* or with a different verb than *sein*

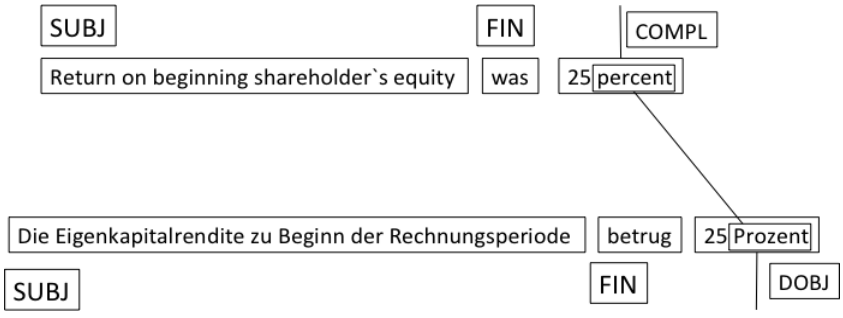

Figure 3: From English copular verb to German full verb

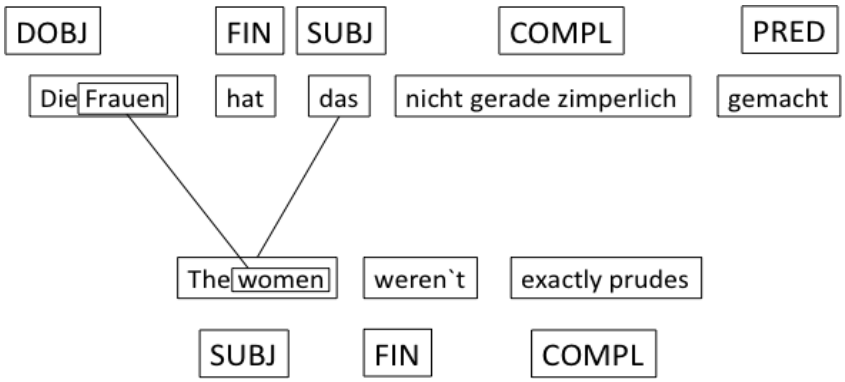

Figure 4: Multiple shifts as a result of translation strategies

The English copular verb *be* is translated with the transitive verb *betragen* in German. This particular kind of verb shift can be observed very often in the register SHARE, as shown in table 1. The reason for this lies in differences in style between English and German SHARE texts: English uses a more colloquial style where German puts rather formulaic expressions, using more full verbs than copular verbs.

Many of the shifts found in translations can be attributed to translation strategies as described e.g. by (Vinay and Darbelnet 1958) for French and English. An example of a modulation can be seen in figure 4. Here, what can be described by looking at the surface realisation, is that the word order from the German original has been kept in the English translation, probably to preserve the stress which is put on the phrase *Die Frauen* `the women`. But, while in German the first constituent is a direct object, this order of grammatical functions cannot be easily reproduced in English. A possible solution, as presented in the given example, is to shift the direct object to another function, here: the subject. In the given example, the verb is shifted, too, from transitive *gemacht* `made` to the copular *weren't*. One could hypothesise that this happens in order to adapt to the different configuration of functions and their semantic content. However, in order to really explain the more complex cases of multiple shifts in one sentence, further data /annotations may be needed.

If, for instance, we add frame semantic annotation, we may be able to describe the shift of the verb with relation to shifts in semantic content. In the example in figure 4, one could annotate the first sentence with the C*ause_change* frame (with *das* as *Cause* and *Die Frauen* as *Entity*), the second one with the *state_of_entity* frame. The English sentence could thus be interpreted as a translation of only a partial component of the sense of the original sentence: the English translation focusses on the outcome of the *Cause_change* process in the German original, giving more stress to the *Entity* (*the women*) in the *State_of_entity* by placing it to the sentence initial position. How to deal with such shifts – whether to include them in an extraction process or not – remains a matter of discussion. Data from process-based translation experiments may prove helpful for shedding light on the reasons for such a "partial" translation.

4. Conclusion and outlook

As has been shown, empty links and crossing lines have proven to be reliable indicators for detecting and in some cases a basis for describing differences in grammatical valence patterns. Furthermore, it has been shown that annotation and alignment on multiple levels can be used for studying valence divergencies and possibly for extracting bilingual valence dictionaries, without resorting to an annotation scheme specialised on these purposes only.

Future work shall concentrate on a broader

categorisation of valence divergencies with respect to more factors than those listed in this paper. In order to be able to link verb senses and certain types of shifts, the next step is to add (frame) semantic annotation to the corpus. Also, the purely product based data presented here could be complemented by process-based studies in the future, which should yield a more sound explanation of shifts as depicted in figure 4.

5. References

Bianco, M. T. (1996): Valenzlexikon deutsch-italienisch. Deutsch im Kontrast 17. Heidelberg: Julius Groos.

Boas, H. C. (2002): Bilingual FrameNet dictionaries for machine translation. In Proccedings of the third international conference on language resources and evaluation, 4:1364-1371. Las Palmas, Spanien.

---- (2005): Semantic frames as interlingual representations for multilingual lexical databases. International Journal of Lexicography 4, no. 18: 445-478.

Catford, J. C. (1965): A linguistic theory of translation. an essay in applied linguistics. Oxford: Oxford University Press.

Čmejrek, M., Cuřín, J., Havelka, J., Hajič, J., Kubon. V. (2004): Prague Czech-English dependency treebank: syntactically annotated resources for machine translation. In Proceedings of LREC 2004, 5:1597-1600. Lisbon, Portugal.

Čulo, O. (2010): Valency, translation and the syntactic realisation of the predicate. In D. Vitaš and C. Krstev, Proceedings of the 29th International Conference on Lexis and Grammar (LGC), 73-82. Belgrade, Serbia.

Cyrus, L. (2006): Building a resource for studying translation shifts. In Proceedings of LREC 2006.

Emele, M. C., Dorna, M., Lüdeling, A., Zinsmeister, H., Rohrer, C. (2000): Semantic-based transfer. In W. Wahlster (ed.), Verbmobil, 359-376. Artificial intelligence. Berlin ; Heidelberg [u.a.]: Springer.

Engel, U., Savin, E. (1983): Valenzlexikon deutsch-rumänisch. Deutsch im Kontrast 3. Heidelberg: Julius Groos.

Gebruers, R. (1988): Valency and MT: recent developments in the METAL system. In Proceedings of the second conference on applied natural language processing, 168-175.

Koller, W. (2001): Einführung in die Übersetzungswissenschaft. Narr Studienbücher. Tübingen: Gunter Narr.

Levin, B. (1993): English verb classes and alternations. The University Chicago Press.

Padó, S. (2007): Translational equivalence and cross-lingual parallelism: the case of framenet frames. In Proceedings of the nodalida workshop on building frame semantics resources for scandinavian and baltic languages. Tartu, Estonia.

Rall, D., Rall, M., Zorrilla, O. (1980): Diccionario de valencias verbales: aleman-español. Tübingen: Gunter Narr.

Sgall, P., Hajičová, E., Panevová, J. (1986): The meaning of the sentence in its semantic and pragmatic aspects. Springer Netherland.

Simon-Vandenbergen, A.-M., Taeldeman, J. ,Willems, D. (eds) (1996): Aspects of contrastive verb valency. Studia Germanica Gandensia 40.

Steiner, E., Schmidt, P., Zelinsky-Wibbelt. C. (1988): From syntax to semantics: insights from machine translation. London: Francis Pinter.

Tesnière, L. (1959): Éléments de syntaxe structurale. Paris: Klinksieck.

Vinay, J.-P., Darbelnet, J. (1958): Stylistique comparée du français et de l'anglais. Méthode de translation. Paris: Didier.

Žabokrtský, Z., Ptáček, J., Pajas. P. (2008). TectoMT: highly modular MT system with tectogrammatics used as transfer layer. In Proceedings of WMT 2008.