Bilingual FrameNet Dictionaries for Machine Translation

Hans C. Boas

Department of Germanic Studies
E.P.Schoch 3.102, C 3300
University of Texas at Austin
Austin, TX 78712-1190, U.S.A.
hcb@mail.utexas.edu

Abstract

This paper describes issues surrounding the planning and design of GermanFrameNet (GFN), a counterpart to the English-based FrameNet project. The goals of GFN are (a) to create lexical entries for German nouns, verbs, and adjectives that correspond to existing FrameNet entries, and (b) to link the parallel lexicon fragments by means of common semantic frames and numerical indexing mechanisms. GFN will take a fine-grained approach towards polysemy that seeks to split word senses based on the semantic frames that underlie their analysis. The parallel lexicon fragments represent an important step towards capturing valuable information about the different syntactic realizations of frame semantic concepts across languages, which is relevant for information retrieval, machine translation, and language generation.

1. Introduction

The aim of GermanFrameNet (henceforth GFN) is to develop a corpus-based lexicon of several thousand German lexical units described in terms of Frame Semantics (Fillmore, 1982). The resulting database is planned to be structured along similar lines as the English-based FrameNet lexicon currently under development at the International Computer Science Institute in Berkeley, California (Lowe et al., 1997; Baker et al., 1998; Fillmore & Atkins, 1998; Johnson et al., 2001).[1] More specifically, GFN will create, based on word uses attested in large corpora, a database of lexical entries for German verbs, nouns, and adjectives taken from a variety of semantic domains. By following the database structure of FrameNet, a lexical entry in GFN will provide an exhaustive account of the syntactic and semantic combinatorial properties of each "lexical unit" (i.e., one word in one of its uses).

In its first phase, GFN will reproduce lexicon fragments that are parallel to existing English FrameNet fragments. The results will show to what degree existing descriptions of English-based semantic frames employed by FrameNet can be used to describe the semantic and syntactic combinatorial properties of German lexical units belonging to the same semantic domain. Of particular interest will be the question of granularity when it comes to deciding whether a number of lexical units associated with a German word should be split or lumped along similar lines as those of the corresponding lexical units associated with their English counterpart(s).

In its second phase, GFN will link the German and English modules, thereby creating a bilingual lexicon on frame semantic principles. By means of accessing the different syntactic realizations of a lexical unit described with reference to a semantic frame, it will thus be possible to go from the German component to the English component to compare how semantic frames are realized differently across languages.

The paper is structured as follows. Section 2 illustrates the structure of lexical entries in the FrameNet database for semantically related sets of words in the Communication-Statement domain. Section 3 discusses the creation of parallel German lexicon fragments by showing how the semantic and syntactic combinatorial properties of German lexical units can be described in terms of the same semantic frames used for the corresponding English lexical units. Section 4 proposes mechanisms designed to link German lexical units to their English translation equivalents by making reference to a common semantic frame in combination with a numerical indexing mechanism. Section 5 is concerned with problems related to polysemy.

2. The FrameNet database

The structure of the FrameNet database is committed to the concepts of Frame Semantics (Fillmore, 1982), a descriptive and analytical framework based on semantic frames containing frame elements (semantic roles). In this approach, word meanings are best understood in reference to the conceptual structures that support and motivate them. This means that any description of word meanings must begin by identifying such underlying conceptual structures, namely the individual frames. Then, words are collected in semantically related sets, representing specific instances of semantic frames, and described with reference to these frames.[2]

2.1. The structure of lexical entries

An example of a set of semantically related words are the verbs *admit*, *announce*, *comment*, *mention*, *remark*, *tell*, etc., together with a set of corresponding nouns such as *admission*, *announcement*, *comment*, *mention*, and *remark*. The description of each lexical unit consists of several parts.

[1] FrameNet is funded by the National Science Foundation through two different grants (IRI #9618838, and ITR/HCI #0086132). See: http://www.icsi.berkeley.edu/~framenet for details.

[2] For a detailed discussion of these ideas, see Fillmore (1982), Fillmore & Atkins (1992), Fillmore & Atkins (1994), and Fillmore & Atkins (1998). For a comprehensive overview of the main concepts of Frame Semantics, see Petruck (1996).

2.1.1. Identification of Semantic Frames

The first part of the description is the identification of the semantic frame that underlies its analysis, in this case the Communication-Statement frame. It includes vocabulary for describing the elements of the frame (semantic roles). In the case of the Communication-Statement frame which involves situations in which a person produces a message, the list of frame elements includes the Speaker, the Addressee, the Message, and the Topic, among others. The Speaker is the person producing a (spoken or written) message, the Addressee is the person to whom the message is communicated, the Message identifies the content of what the Speaker is communicating to the Addressee, and the Topic is the subject matter to which the Message pertains.[3]

2.1.2. Semantic and syntactic combinatorial properties

The second part of the description of each lexical unit contains information about the manner in which syntactic constituents realizing semantic frame elements may or must be distributed within and around phrases headed by that word. This part of the description lists all possible combinations of frame elements, phrase types, and grammatical functions that realize elements of the frame. For example, for the Speaker frame element in the Communication-Statement frame it is important to record that it may be expressed as the external argument of predicative uses of the target word (e.g., [Nancy] *announced* her retirement, [Nancy] made a surprising *announcement*), or as the genitive modifier of the target noun (e.g., [Nancy's] *announcement* that she would retire). Similarly, the Addressee can be expressed in different ways, e.g. as a direct object (Nancy *told* [Collin] about what happened) or as a prepositional phrase introduced by *to* (Nancy *announced* the sale [to the staff]). For the Message frame element it is important to know that it can be expressed as a noun phrase (Michael *admitted* [two thefts between Monday and Thursday]) or a clause (Michael *admitted* [that he had committed two thefts between Monday and Thursday]), among other possibilities. For lexical units occurring in the Communication-Statement frame it is also important to know that the Topic may be expressed by a prepositional phrase headed by a number of different prepositions such as *about* (Miriam *talked* that night [about the people she was to give the money to]) and *of* (Joe *talked* [of buying more beer]), or by a direct object (She *explained* [her unhappiness] to a friend). For each combination of frame elements, grammatical functions, and phrase types, a collection of annotated corpus-attested sentences is given.[4]

Table 1 is a part of the lexical entry of the verb *announce* (the target) in the Communication-Statement frame as it is recorded in the FrameNet database. It summarizes for corpus-attested sentences (in which the Message precedes and the Speaker follows *announce*) the phrase types (e.g., Noun Phrase (NP)) and syntactic functions (e.g., Ext(ernal argument)) of the two frame elements.[5]

	message	TARGET	speaker
1	NP.Ext	announce.v	CNI
2	NP.Ext	announce.v	PP_by.Comp
3	QUO.Comp +QUO.Comp	announce.v	NP.Ext
4	QUO.Comp	announce.v	NP.Ext

Table 1: Subpart of the lexical entry for *announce*[6]

Table 2 lists the annotated corpus-attested example sentences linked to the information in Table 1 as a subpart of the lexical entry for *announce* in the FrameNet database.

1	On November 5 [_{message} her pregnancy] is officially *announced*[Tgt]. [_{speaker} CNI]
2	Extbuy records [_{message}acquisitioin criteria which] are *announced*[Tgt] in the press [_{speaker}by companies].
3	[_{message}'We now,"] *announced*[Tgt] [_{speaker}the Speaker], [_{message}'come to the debate on War Crimes."]
4	[_{message}'Martial law has been declared on Mars,"] *announced*[Tgt] [_{speaker}Derek Carlisle].

Table 2: Semantically annotated corpus sentences

Table 3 is a small subpart of the lexical entry of *announcement* and illustrates how the same frame elements are used to describe words belonging to the identical semantic frame, in this case Communication-Statement. Table 4 contains the relevant annotated corpus-attested example sentences.

	TARGET	speaker	message
1	announcement.N	PP_by.Comp	PP_of.Comp
2	announcement.N	PP_by.Comp	Sfin.Comp

Table 3: Subpart of the lexical entry for *announcement*

1	He was spurred to bring it to the attention of the scientific world by the *announcement*[Tgt] [_{speaker}by the Frenchman Daguerre][_{message}of his revolutionary invention, the Daguerreotype].
2	This decision coincides with an *announcement*[Tgt] [_{speaker}by Soviet scientists and ecologists] [_{message}that 1,382,000 square miles of the Soviet Union is an "ecological disaster area"].

Table 4: Semantically annotated corpus sentences

[3] See Johnson et al. (2001) for further details.

[4] The main corpus for FrameNet is the British National Corpus that is used under an agreement with Oxford University Press.

[5] Information about the phrase type precedes information about the grammatical function of a particular realization of a frame element. CNI stands for constructional null instantiation which indicates that a given frame element is not overtly realized, but licensed by a grammatical construction (in this case the passive construction). Other types of null instantiation include definite and indefinite null instantiation (DNI and INI).

[6] Table 1 only lists the different phrase type and grammatical function realizations of Message and Speaker in one particular order. Other corpus-attested combinations of frame elements occurring with *announce* in the Communication-Statement frame include [Medium, Target, Message], [Message, Speaker, Target], [Message, Speaker, Target, Addressee], [Message, Speaker, Target, Manner], among others.

2.1.3. Relations between semantic frames

The third part of the description of a lexical unit includes a record of other frames of which the target frame is an elaboration or a blend. For Communication-Statement, for example, it is important to know that the frame belongs to a larger domain of frames that all express situations in which information passes from a Speaker to an Addressee. Other frames belonging to the communication domain include Communication-Conversation, Communication-Questioning, and Communication-Response, among others.

2.1.4. The lexical entry

Combining the information discussed so far yields an entry for each lexical unit in the FrameNet database that contains semantically and syntactically annotated sentences from which reliable information can be reported on its semantic and syntactic combinatorial properties. Furthermore, it offers a coherent method of describing sets of semantically related words by making reference to a common structuring device, namely the semantic frame. Adding to this information the part of speech of the lexical unit as well as its definition then results in the complete lexical entry as it appears in the FrameNet database.

2.2. Comparing properties of lexical units

Besides listing a detailed corpus-attested inventory of the different syntactic and semantic combinatorial properties of individual lexical units, the FrameNet database also offers an efficient way of comparing how frame elements are realized differently by individual lexical units. Consider, for example, the multiple ways in which the frame elements Speaker, Addressee, Message, and Topic are realized by the following verbs in the Communication-Statement frame.

		speaker
1	*admit*	CNI, NP.Ext, PP_by.Comp
2	*announce*	CNI, NP.Ext, PP_by.Comp
3	*comment*	CNI, NP.Ext, PP_by.Comp
4	*mention*	CNI, NP.Ext
5	*proclaim*	CNI, NP.Ext, PP_by.Comp
6	*propose*	CNI, NP.Ext, PP_by.Comp
7	*remark*	CNI, NP.Ext
8	*tell*	CNI, DNI, NP.Ext, PP_by.Comp
9	*talk*	CNI, NP.Ext, PP_by.Comp

Table 5: Syntactic realization of *Speaker*

		addressee
1	*admit*	PP_in.Comp, PP_to.Comp
2	*announce*	PP_to.Comp
3	*comment*	PP_to.Comp
4	*mention*	PP_to.Comp
5	*proclaim*	PP_to.Comp
6	*propose*	PP_to.Comp
7	*remark*	PP_to.Comp
8	*tell*	AVP.Comp, NP.Ext, NP.Obj, PP_to.Comp, Sfin.Comp
9	*talk*	NP.Ext, PP_at.Comp, PP_to.Comp

Table 6: Syntactic realization of *Addressee*

		topic
1	*admit*	n.a.
2	*announce*	n.a.
3	*comment*	DNI, NP.Ext, PP_about.Comp, PP_on.Comp, PP_upon.Comp
4	*mention*	NP.Obj
5	*proclaim*	n.a.
6	*propose*	n.a.
7	*remark*	PP_about.Comp, PP_of.Comp, PP_on.Comp, PP_over.Comp
8	*tell*	NP.Obj, PP_about.Comp, PP_of.Comp, PPing_about.Comp, Sfin.Comp
9	*talk*	AVP.Comp, NP.Ext, PP_about.Comp, PP_in.Comp, PP_of.Comp, PPing_about.Comp, PPing_of.Comp, Sfin.Comp, VPing.Comp

Table 7: Syntactic realization of *Topic*

		message
1	*admit*	NP.Ext, NP.Obj, PP_to.Comp, PPing_to.Comp, QUO.Comp, Sfin, Sfin.Comp, Swh.Comp, VPing.Comp
2	*announce*	NP.Ext, NP.Obj, QUO.Comp, Sfin.Comp
3	*comment*	AVP.Comp, PP_on.Comp, QUO.Comp, Sfin.Comp
4	*mention*	NP.Ext, NP.Obj, QUO.Comp, Sfin.Comp, Swh.Comp
5	*proclaim*	AVP.Comp, NP.Comp, NP.Ext, NP.Obj, PP.Comp, QUO.Comp, Sfin.Comp
6	*propose*	AVP.Comp, NP.Ext, NP.Obj, PP.Comp, PP_as.Comp, QUO.Comp, Sfin.Comp, VPing.Comp, VPto.Comp
7	*remark*	AVP.Comp, QUO.Comp, Sfin.Comp, Swh.Comp
8	*tell*	AVP.Comp, DNI, NP.Comp, NP.Ext, NP.Obj, PP_about.Comp, PP_of.Comp, PPing_about.Comp, PPing_of.Comp, PPing_since.Comp, QUO.Comp, Sfin.Comp, Swh.Comp
9	*talk*	NP.Ext, NP.Obj

Table 8: Syntactic realization of *Message*

The data illustrate how nine verbs of the Communication-Statement frame realize the same frame elements in a variety of different combinations. Whereas the realization of Speaker (Table 5) and Addressee (Table 6) do not vary widely, the realization of Topic (Table 7) and Message (Table 8) exhibit great variation among the different verbs. This is valuable information when it comes to employing syntactic parsers for applications such as text understanding or machine translation. When faced with word-sense disambiguation tasks, parsers may rely on lists of a lexical unit's semantic and syntactic combinatorial properties to evoke a particular semantic frame (or even number of frames) in order to disambiguate between multiple possible parses. For example, the verb *talk* is listed as evoking at least two different frames in FrameNet, namely Communication-Statement and Communication-Conversation.

	NP.Ext	talk.v	NP.Obj
1	[<speaker>They]	talked	[<medium>the same language]
2	[<speaker>They]	talked	[<message>a lot of nonsense]
3	[<interlocutors>We]	talked	[<topic>price]

Table 9: Different senses of *talk* occurring with the same syntactic frame

Although the syntactic frames occurring with *talk* in all three examples in Table 9 are similar, the meanings conveyed by the individual sentences differ. In order to know what the individual sentences in Table 9 mean it is therefore not enough to know the syntactic frame(s) occurring with *talk*. Instead, it is important to have access to lexicographic data providing semantic information that goes beyond traditional syntactic subcategorization frames. In the case of the first two sentences, the frame description of Communication-Statement informs us about the full range of semantic frame elements that can occur as postverbal object NPs (Medium and Message) in combination with preverbal external argument NPs. For the third sentence in Table 9, the plural external argument NP as part of the syntactic frame associated with *talk* evokes the Communication-Conversation frame. The difference between the Statement and Communication frames is in the way in which information is communicated. Whereas the former involves only a single speaker, the latter involves a reciprocal exchange of information between two interlocutors. Once the external argument NP is identified as either singular or plural, the semantic frame is identified and the FrameNet database is capable of providing information about the semantic relationships that hold between the constituents in the sentence. The same holds for the data in Table 10 where the singular vs. plural distinction of the external arguments gives decisive information about the semantic frames underlying the two sentences.

	NP.Ext	Talk.v	PP_about.Comp
1	[<speaker>I]	talked	[<topic>about all sorts of things]
2	[<interlocs>We]	talked	[<topic>about a lot of books]

Table 10: Different senses of *talk* occurring with the same syntactic frame

3. Creating German lexicon fragments

As pointed out in the introduction, GFN will follow in its creation of parallel lexicon fragments for German the main procedures established by the original FrameNet project. Initially, this means that GFN will determine how many of the English lexical entries described with respect to semantic frames can be mapped to German ones using the same semantic frames.

3.1. Mapping English lexical units to German

The project workflow will begin by targeting a list of English lexical units described in terms of a semantic frame. For each lexical unit, the first task will be to find a German lexical unit that is a translation equivalent to the English lexical unit. Ideally, this means that the German equivalent lexicalises the same (or close to the same) set of situations that are expressed by the English source lexical item.

3.1.1. Entries for English lexical units

Finding the range of possible German translation equivalents will first require a review of the syntactic and semantic combinatorial possibilities listed in the FrameNet database. Take, e.g., the verb *argue* in the Communication-Conversation frame. This frame describes situations in which one or more parties are exchanging information about a topic with another party. The frame elements include Interlocutors and Topic, among others. Table 11 illustrates a part of the lexical entry for *argue* in the Communication-Conversation frame. The corresponding annotated examples are given in Table 12.

	interlocutors	TARGET	topic
1	NP.Ext	argue.v	INI
2	NP.Ext	argue.v	PP_over.Comp
3	NP.Ext	argue.v	PP_about.Comp
4	NP.Ext	argue.v	PPing_about.Comp
5	NP.Ext	argue.v	Swhether.Comp

Table 11: Partial lexical entry of *argue* in Communication-Conversation

1	[<interlocutors>Mr and Mrs Popple] *argued*^{Tgt} once a week. [<topic>INI]
2	[<interlocutors>Auction houses and buyers] *argue*^{Tgt} [<topic>over compensation]
3	[<interlocutors>They] *argued*^{Tgt} [<topic>about it].
4	Anne says [<interlocutors>they] *argue*^{Tgt} [<topic>about drinking beer].
5	[<interlocutors>One]can *argue*^{Tgt} [<topic> whether pizza is healthy].

Table 12: Semantically annotated corpus sentences

3.1.2. Identifying German translation equivalents

The next step in finding German translation equivalents for the entries listed in the FrameNet database will involve working with bilingual and monolingual dictionaries as well as electronic corpora. For each combination of semantic and syntactic information recorded for an English lexical unit in the FrameNet database, a German equivalent will have to be identified that matches its meaning as closely as possible. For example, in cases when the Interlocutors and Topic frame elements are realized as an external argument and an indefinite null instantiation as in (1) in Tables 11 and 12, the closest translation equivalents include the two sentences in Table 13. Note that both reflexive and non-reflexive usages of German *streiten* are possible equivalents expressing the same type of situation as that expressed by *argue* in the context of (1) in Table 12.

1a	[<interlocutors>Herr und Frau Popple] *stritten*^{Tgt} ein mal pro Woche [<topic> INI].
1b	[<interlocutors>Herr und Frau Popple] *stritten*^{Tgt} [<ref>sich ein mal pro Woche [<topic> INI].

Table 13: German equivalents for example 12(1)

Similarly, the meanings expressed by *argue* in 12(2) – 12(3) can be expressed by reflexive and non-reflexive usages of *streiten* as Table 14 illustrates.[7]

2a	[<interlocutors>Auktionshäuser und Käufer] *streiten*^{Tgt} [<topic>um die Entschädigung].
2b	[<interlocutors>Auktionshäuser und Käufer] *streiten*^{Tgt} [<ref>sich][<topic> um die Entschädigung].
3a	[<interlocutors>Sie] *stritten*^{Tgt} [<topic>darüber].
3b	[<interlocutors>Sie] *stritten*^{Tgt} [<ref>sich] [<topic>darüber].

Table 14: German equivalents for examples 12(2) – 12(3)

Once a set of German translation equivalents is identified on the basis of traditional dictionaries (e.g., *Duden* or *Wahrig*), the next step will require a thorough search for corpus evidence to support our findings. At this stage, we will search a number of German electronic corpora to find attested usages for each syntactic frame associated with a German lexical unit. For example, based on the list of sentences contained in Tables 13 and 14, a corpus search for *streiten* is conducted to see (a) whether we find corpus attestations for each of the syntactic frames listed for the verb by traditional dictionaries, and (b) whether we find any other syntactic frames associated with *streiten* that are not mentioned by traditional dictionaries. By supporting our search for corpus-attested example sentences with native speaker intuitions, we expect this stage of the workflow to reveal the full range of syntactic frames associated with a lexical unit.[8]

3.1.2. Semantic annotation

The search for representative German corpus attestations will result in the creation of a number of individual subcorpora for each lexical unit. Each subcorpus will contain a list of sentences that exemplify a specific syntactic frame occurring with a given lexical unit. Out of each subcorpus, a number of canonical examples will be chosen for semantic annotation.[9] As in the original FrameNet project (Baker et al., 1998), annotators for GFN will use special annotation software to mark selected constituents in the extracted data according to the frame elements that they realize. This process will result for each subcorpus in a number of semantically annotated corpus sentences exemplifying how individual frame elements of a semantic frame are realized syntactically by the target word (see, e.g., Tables 13 and 14).

3.2. Lexical entries in GFN

The next step in the project workflow will be concerned with the writing of lexical entries. Based on semantically annotated corpus examples (cf. Table 14), we will write a database entry for each lexical unit that is parallel in structure to the lexical entries in the FrameNet database. The database entry will identify a word in one of its senses ("lexical unit"), a part of speech, and a frame. In addition to identifying the frame, we will list an indication of the higher-level frames which it inherits (e.g., *Conversation* inherits from *Communication*, which in turn inherits from a more general *Reciprocity* frame (see Fillmore & Atkins, 1998)). This will be augmented by a list with explanations of the frame elements used in the annotation together with the ways in which they can be syntactically realized.[10]

Furthermore, the lexical entry will include the collection of selected and annotated corpus sentences that exhibit every attested combinatorial pattern for the lexical unit. Tables 15 and 16 are preliminary examples of the structure of lexical entries that will be produced by GFN. They contain partial summaries of the semantic and syntactic combinatorial properties for the lexical entries of the non-reflexive and reflexive usages of *streiten* in the Communication-Conversation frame. They are based on annotated examples of the type contained in Tables 13 and 14.

	interlocutors	**TARGET**	**topic**
1a	NP.Ext	streiten.v	INI
2a	NP.Ext	streiten.v	PP_um.Comp
3a	NP.Ext	streiten.v	PP_über.Comp

Table 15: Partial lexical entry for *streiten*

	Interloc.	**TARGET**	**Ref.**	**topic**
1b	NP.Ext	streiten.v	sich	INI
2b	NP.Ext	streiten.v	sich	PP_um.Comp
3b	NP.Ext	streiten.v	sich	PP_über.Comp

Table 16: Partial lexical entry for reflexive *streiten*

4. Linking English and German lexicon fragments

Having sketched the procedure by which GFN is planning to create parallel lexicon fragments for German, we now address the issue of how to link the German entries to their corresponding English counterparts. One of the main problems when it comes to linking modular lexicon fragments has to do with cases in which the lexical semantics encoded by a lexical item in the source language does not exactly match the lexical semantics encoded by the corresponding lexical item in the target language.[11]

[7] Note that *sich streiten* is not a prototypical reflexive, but is only used reciprocally.

[8] This stage will require a detailed analysis of the semantics associated with a verb in combination with its various prepositional complements (cf. *streiten* {*um*/*über*/*für*/…}) as well as its English counterparts.

[9] See Fillmore & Atkins (1998) for the notion of lexicographic relevance.

[10] Another problem that GFN will have to deal with is different types and degrees of verb descriptivity (Snell-Hornby, 1983). That is, although corresponding German words may be closely related to their English counterparts, they may lack some of the semantic expressiveness of the English words (or vice versa). Compare, e.g., *bustle*, *grovel*, and *bummeln*, which exhibit high descriptivity, as opposed to *shout* and *wälzen*, which exhibit low descriptivity (Snell-Hornby, 1983).

[11] See Heid's (1994) classification of contrastive lexical problems based on his discussion of Kameyana et al.'s (1991) and Dorr's (1990) description of "mismatches" and "divergences." For an a possible solution to this problem that advocates the combination of Frame Semantics and Meaning-Text-Theory, see Fontenelle (2000).

GFN will aim to overcome this problem by employing semantic frames as structuring devices in order to link German lexical entries to their English counterparts. Since frames encode the semantic relationships and role constellations that hold between the individual frame elements, the inventory of frame elements will be used to compare how a given combination of semantic and syntactic information encoded by a lexical item in the source language is realized in the target language. This means that for each semantic and syntactic combinatorial property of a given lexical unit in the source language we will ideally have a correspondence link to its counterpart in the target language that makes use of the semantic frame as a structuring device.

The following figure is a schematic illustration showing how GFN is planning to make use of the notion of semantic frame to link corresponding subparts of parallel lexical entries to each other.[12] The actual implementation will use a variant of the typed feature structure system employed by the DELIS project (Emele 1994; Heid, 1994; Heid & Krüger, 1996) for linking monolingual dictionaries.

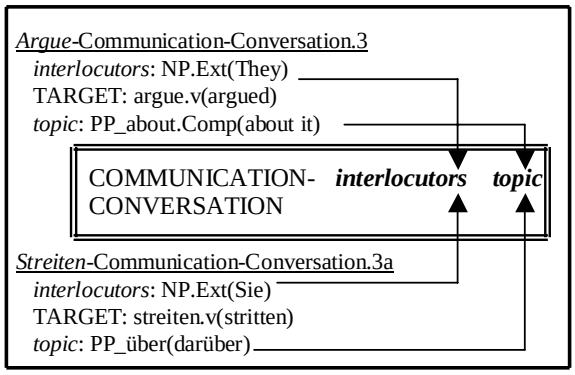

Figure 1: Semantic frame as a structuring device to link subparts of English and German lexical entries

As discussed in the previous sections, lexical entries contain exhaustive listings of the semantic and syntactic combinatorial properties. Assigning each subpart of a lexical entry a number makes it possible to identify a specific syntactic frame occurring with a given lexical item. When it comes to establishing correspondence links between English and German lexical entries, this numerical indexing system will allow us to precisely refer to a given subpart of a lexical entry in the source language when linking it to the corresponding subpart of a lexical entry in the target language.

For example, index "3" in Figure 1 indicates that a specific syntactic frame of *argue* is used to encode the semantics of the Communication-Conversation frame (cf. Table 11). The German equivalent is indexed with "3a" (cf. Table 15), referring to a specific subpart of the lexical entry for *streiten* in the Communication-Conversation frame and thereby indicating that this is the German translation equivalent.

[12] Similar proposals in favor of using semantic frames as structuring devices to link English lexical entries to German lexical entries have been made by Boas (2000) and Boas (2001).

An additional advantage of the numerical indexing system is that it allows for cross-referencing between subparts of multiple lexical entries across English and German lexicon fragments in combination with semantic frames. With respect to translation equivalents for *argue* in the Communication-Conversation frame in Figure 1, this means that we could add other links to the Communication-Conversation frame. One such option includes a link to a subpart of the lexical entry for the reflexive version of German *streiten*. In this case, this link would be established to the syntactic frame of the reflexive (reciprocal) usage of *streiten* that is indexed with "3b" in Table 16.[13]

5. Polysemy

Other multilingual database projects such as EuroWordNet aim at clustering word senses based on the concepts of generalization, metonomy, and diathesis alternation (Peters et al., 1998). These clustering methods rely crucially on relations between words described in terms of synsets and their basic semantic relations to each other (hyponomy, antonomy, meronomy, etc.) (Miller et al., 1990). The goal of word sense clustering is to "reduce the fine-grainedness of WordNet and express in a more systematic way the relations between its numerous sense distinctions." (Peters et al., 1998).

In contrast to the sense clustering approach advocated by EuroWordNet, GFN plans to take a more fine-grained approach to polysemy. Since GFN will initially create German counterparts to existing English FrameNet lexical entries, it will also employ semantic frames as structuring devices for its lexical database. This means that when a sense distinction is recorded in the FrameNet database for an English word, GFN will search for a similar sense distinction in German. Whenever a comparable sense distinction in German can be found in the corpus, GFN will then create a similar lexical entry for the German counterpart and link it to the English FrameNet entry as discussed in the previous section.

5.1. Multiple translation equivalents in the same frame

In more complicated cases, however, a given English lexical unit will have more than one equivalent counterpart in German. When such a case arises, it will require careful consideration, incorporating the consultation of native speaker intuition in addition to traditional dictionaries and electronic corpora

For example, the verb *announce* in the Communication-Statement frame discussed in section 2 has a multitude of German translation equivalents including *bekanntgeben*, *bekanntmachen*, *ankündigen*, *anzeigen*, *ansagen*, and *durchsagen*. The choice of the German counterpart depends crucially on what part of

[13] Using semantic frames in combination with numerical indexing mechanisms is different from the Inter-Lingual-Index (ILI) employed by EuroWordNet that aims to create a minimalized and efficient list of sense-distinctions (Vossen & Bloksma, 1998; Peters et al., 1998). In contrast to ILI-records, GFN employs frame semantic descriptions to record lexicographically relevant corpus attestations of semantic and syntactic combinatorial properties of a lexical item without minimalizing sense distinctions.

Communication-Statement frame is highlighted by a given syntactic frame of *announce*. To illustrate, compare the following sentences in which different syntactic frames occur with *announce*.

1	[<speaker>They] *announced*[Tgt] [<message>the birth of their child].
2	[<medium>The document] *announced*[Tgt] [<message>that the war had begun].
3	[<speaker>The conductor] *announced*[Tgt] [<message>the train's departure] [<medium>over the intercom].

Table 17: Syntactic frames highlighting different parts of the Communication-Statement frame

All three usages of *announce* in Table 17 fall into the Communication-Statement frame. However, each sentence highlights different frame elements and their relationships to each other. That is, 17(1) highlights the Speaker and the Message, 17(2) the Medium and the Message, and 17(3) the Speaker, the Message, and the Medium. For each of the three sentences in Table 17, we find different German verbs expressing the semantic relationships between the frame elements evoked by *announce* in the Communication-Statement frame.

For example, cases in which *announce* occurs with only the Speaker and the Message frame elements as in 17(1) are typically translated into German by using *bekanntgeben*, *bekanntmachen*, *ankündigen*, and *anzeigen*.[14] In contrast, *ansagen* and *durchsagen* are typically not considered as translation equivalents for *announce* in 17(1). This is because the two verbs are primarily used in cases in which a Medium frame element represents some sort of (electronic) equipment used to communicate the Message to the Addressee such as in 17(3). This example shows that it will not be sufficient to simply generalize over senses of words that may be used as synonyms of each other. Instead, it will be necessary for GFN to capture the full range of possible German equivalents before arriving at decisions about which German verbs may serve as possible equivalents to a specific syntactic frame listed in an entry for an English lexical unit.[15] Only then will the GFN database serve as a reliable basis for machine translation by mapping English entries to their corresponding German counterparts and vice versa.

Because of space limitations, similar restrictions holding for other German translation equivalents of *announce* cannot be discussed here. Table 18 summarizes the distribution of the German verbs *bekanntgeben*, *bekanntmachen*, *ankündigen*, *anzeigen*, *ansagen*, and *durchsagen* as translation equivalents for the three syntactic frames associated with *announce* in Table 17 above. The difference in lexicalization patterns of frame elements between *announce* in the Communication Frame and its multiple German counterparts illustrated by Table 18 is an example of the types of problems that GFN will have to address.[16]

1	**speaker**	**TARGET**	**message**	
	NP.Ext	announce.v	NP.Obj	
	bekanntgeben, bekanntmachen, ankündigen, anzeigen			
2	**medium**	**TARGET**	**message**	
	NP.Ext	announce.v	Sfin_that.Comp	
	bekanntgeben, ankündigen, anzeigen			
3	**speaker**	**TARGET**	**message**	**medium**
	NP.Ext	announce.v	NP.Obj	PP_over.Comp
	ankündigen, ansagen, durchsagen			

Table 18: Different syntactic frames of *announce* and corresponding German verbs

5.2. Translation equivalents across multiple frames

GFN will also have to address cases in which multiple senses of a single English word are expressed by different words in German. The two senses of *walk* listed in Table 19 are an example. The first is the Self-motion sense in which a Self-mover moves towards a Goal (Johnson et al., 2001). The second sense of *walk* is the Motion-Cotheme sense which implies that the Self-mover and the Cotheme move together to a final destination, the Goal (Johnson et al., 2001).[17]

1	[<self-mover>Kim] *walked*[Tgt] [<goal>to the store].
2	[<self-mover>Kim] *walked*[Tgt] [<cotheme>Pat] [<goal>to the door].

Table 19: *Walk* in the Self-motion and Cotheme-Motion Frames

1	[<self-mover>Kim] *ging*[Tgt] [<goal>zum Geschäft].
2	[<self-mover>Kim] *begleitete*[Tgt] [<cotheme>Pat] [<goal> zum Geschäft].

Table 20: German translation equivalents of Table 19

The corresponding German examples listed in Table 20 illustrate that German *gehen* does not exhibit the same degree of polysemy as English *walk*. That is, instead of employing *gehen* to express the Motion-Cotheme semantics associated with English *walk* in 19(2), German requires the use of a different word, namely *begleiten*.[18]

[14] In reality, a much finer-grained distinction (including contextual background information) is needed to formally distinguish between the semantics of individual verbs. E.g., *anzeigen* is used in a much more formal sense than the other verbs. In contrast, *ankündigen* is primarily used to refer to an event that will occur in the future.

[15] Note that it will not suffice to only map a lexical unit's equivalents to German. Instead, GFN will have to map each syntactic frame of a German lexical unit back to a syntactic frame of an English lexical unit in order to ensure that the two are capable of expressing the same semantic space. Whenever there are discrepancies, a revision of mappings between lexical entries will be necessary.

[16] For an in-depth discussion of issues surrounding different polysemy networks across languages see Fillmore & Atkins (2000) on the polysemy of English *crawl* and French *ramper*.

[17] For an in-depth discussion of motion verbs occurring in multiple frames in English and German, see Boas (2001).

[18] Note that *accompany* in the Cotheme-Motion frame is used to express similar situations as those described by *walk* ([<self-mover>Kim] *accompanied*[Tgt] [<Cotheme>Pat] [<goal>to the door]).

GFN will deal with these cases of polysemy by referring to semantic frames as structuring devices and employing numerical indexing mechanisms to link the English lexical entries to their corresponding German counterparts as discussed in section 4.

6. Conclusions and Outlook

This paper has outlined a number of issues surrounding the planning and design of GermanFrameNet (GFN). Based on a discussion of the architecture of the English FrameNet database, it described the procedures by which GFN will create German lexicon fragments that are parallel to existing lexical entries in the FrameNet database. The result will be a corpus-based lexicon giving an exhaustive account of the syntactic and semantic combinatorial properties of several thousand German verbs, nouns, and adjectives taken from a variety of semantic domains.

Furthermore, this paper discussed how GFN will employ semantic frames and numerical indexing mechanisms to link the parallel lexicon fragments to each other. Based on a review of different polysemy patterns exhibited by a number of English and German verbs in the Communication-Statement frame, the paper finally suggested a fine-grained approach towards describing polysemy that crucially differs from the methods advocated by EuroWordNet (Peters et al., 1998; Vossen & Bloksma, 1998).

While GFN will primarily be a lexicographic project and will thus not be dedicated to NLP efforts as such, it will serve as a contrastive lexical database that we hope to be useful for research in machine translation. In particular, we expect information about the syntactic and semantic combinatorial properties of German and English lexical units (including information about their semantic frames) to be of great importance when it comes to automatically determining the closest translation equivalent for a given lexical unit.

7. References

Baker, C.F., C.J. Fillmore, and J.B. Lowe, 1998. The Berkeley FrameNet Project. In *COLING-ACL '98: Proceedings of the Conference, held at the University of Montréal*.

Boas, H.C., 2000. *Resultative Constructions in English and German*. Ph.D. thesis, University of North Carolina at Chapel Hill.

Boas, H.C., 2001. Frame Semantics as a framework for describing polysemy and syntactic structures of English and German motion verbs in contrastive computational lexicography. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja (eds.), *Proceedings of Corpus Linguistics 2001*.

Dorr, B., 1990. Solving Thematic Divergences in Machine Translation. In *Proceedings of the 28th Annual Conference of the Association for Computational Linguistics*, University of Pittsburgh.

Emele, M. 1994. TFS – The Typed Feature Structure Representation Formalism. In *Proceedings of the International Workshop on Sharable Natural Language Resources (SNLR)*.

Corresponding indexing links will thus link 20(1) with 19(2) as well as the corresponding entry for *accompany*.

Fillmore, C.J., 1982. Frame Semantics. In The Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm*. Seoul: Hanshin.

Fillmore, C.J., and B.T.S. Atkins, 1992. Toward a frame-based lexicon: The semantics of RISK and its neighbours. In A. Lehrer and E. Kittay (eds.), *Frame, fields, and contrasts: New essays in semantic and lexical organization*. Hillsdale: Erlbaum.

Fillmore, C.J., and B.T.S. Atkins, 1994. Starting where the dictionaries stop: the challenge for computational lexicography. In B.T.S. Atkins and A. Zampolli (eds.), *Computational Approaches to the Lexicon*. Oxford: Oxford University Press.

Fillmore, C.J., and B.T.S. Atkins, 1998. FrameNet and Lexicographic Relevance. In *Proceedings of the First International Conference on Language Resources and Evaluation*.

Fillmore, C. J., and B.T.S. Atkins, 2000. Describing Polysemy: The Case of 'Crawl'. In Y. Ravin and C. Laecock (eds.), *Polysemy*. Oxford: Oxford University Press.

Fontenelle, T., 2000. A Bilingual Database for Frame Semantics. *International Journal of Lexicography: 232-248*.

Kameyama, M., R. Ochitani, and S. Peters, 1991. Resolving Translation Mismatches with Information Flow. In *ACL-29*.

Lowe, J.B., C.F. Baker, and C.J. Fillmore, 1997. A frame-semantic approach to semantic annotation. In M. Light (ed.), *Tagging Text with Lexical Semantics: Why, What and How?* Special Interest Group on the Lexicon, Association for Computational Linguistics.

Heid, U.,1994. Contrastive Classes – Relating Mono-lingual Dictionaries to Build an MT Dictionary. In F. Kiefer, G. Kiss, and J. Pajzs (eds.), *Papers in Computational Lexicography – COMPLEX 1993*.

Heid, U., and K. Krüger, 1996. A Multilingual Lexicon based on Frame Semantics. In *Proceedings of the AISB Workshop on Multilinguality in the Lexicon, Brighton*.

Johnson, C., C.J. Fillmore, E. Wood, J. Ruppenhofer, M. Urban, M. Petruck, C. Baker, 2001. *The FrameNet Project: Tools for Lexicon Building*. Manuscript. Berkeley, CA, International Computer Science Institute.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, 1990. *Five Papers on WordNet*. CSL Report 43. Cognitive Science Laboratory. Princeton University.

Peters, W., I. Peters, and P. Vossen, 1998. Automatic Sense Clustering in EuroWordNet. In *Proceedings of the First International Conference on Language Resources and Evaluation*.

Petruck, M. 1996. Frame Semantics. In J.-O. Östman, J. Blommaert, and C. Bulcaen (eds.), *Handbook of Pragmatics*. Amsterdam/Philadelphia: Benjamins.

Snell-Hornby, M. 1983. *Verb descriptivity in English and German: A contrastive study in semantic fields*. Heidelberg: Carl Winter.

Vossen, P., and L. Bloksma, 1998. Categories and Classifications in EuroWordNet. In *Proceedings of the First International Conference on Language Resources and Evaluation.*

Vossen, P., W. Peters, and J. Gonzalo, 1999. Towards a Universal Index of Meaning. In *Proceedings of the ACL-99 SIGLEX workshop*, University of Maryland.