A Framenet and Frame Annotator for German Social Media

Eckhard Bick

University of Southern Denmark
eckhard.bick@mail.dk

Lexical semantic resource types

- WordNet (Fellbaum 1998)
  - extensive ontology, but no inherent method for disambiguation or annotation, better for nouns than verbs
- FrameNet (Baker et al. 1998, Johnson & Fillmore 2000, Ruppenhofer et al. 2010)
  - systematic classification of abstracted verb senses with semantically restricted "slot-filler" arguments: potential for disambiguation, but coverage problem at the token level
- PropBank (Palmer et al. 2005)
  - consecutive argument structure annotation of corpus verbs (propositions): better coverage and statistical balance, but less generalization than FrameNet
- VerbNet (Kipper et al. 2006)
  - less granularity, more limited set of roles and predicate classes

Frames: From resource to annotation

- **corpus-driven:** German SALSA framenet (Burchardt et al. 2006) and SHALMANESER parser (Burchardt et al. 2009)
  - 600 different frames, half specific to German, but coverage limited to the hand-annotated corpus
  - heuristic frame assignment for lexicon gaps, based on Word Net synsets
  - ML-based corpus annotation
- **parser-driven:** PFN-DE ("parser framenet", this paper)
  - unabridged lexicon, simple parsing-oriented framenet scheme
  - matches the valency lexicon and noun ontology of an existing morphosyntactic parser (GerGram)
  - supports a rule-based frame annotator that directly exploits GerGram's tags and dependency links
  - matches a similar system for Danish (Bick 2011), allowing comparable corpus annotation for our bilingual Social Media Corpus (XPEROHS, Baumgarten et al. 2019)

Cross-lingual framenet bootstrapping

- 1. step: **Danish-German verb sense matches** based on an MT dictionary (GramTrans, Bick 2007)
  - MT: verb polysemy resolution by listing arguments with semantic slot filler information
  - **valency patterns as an anchor for frame transfer:** harvest a Danish frame (including its selection restrictions) by matching the MT dictionary's argument list, choosing - for each translation - the frame with the same valency pattern
- 2. step: **manual checking** of the harvested frames
  - german valency patterns used to identify gaps in existing entries and as skeletons for verbs without an MT entry
- 3. step: **identifying frame lexicon gaps** in a preliminary annotation of the XPEROHS corpus
  - frequency-based manual frame additions
  - systematic check of construction verbs for idiomatic senses/constructions

PFN-DE: Lexicon size and granularity

- Frame lexicon size:
  - 11,333 verb lemmas
  - 14,695 different lemma+frame combinations
  - 1.297 frames / lemma (1.237 semantic types / lemma), Zipfian distribution
  - coverage: all entries in the parser lexicon have at least 1 frame), corpus: 1-2% lexical frame failure rate
- Frame types
  - 483 types (almost all Danish frames also used for German)
  - 1,700 different combinations of "atomic" frames, to capture additional lexical information (aspect, directionality, urgency), often triggered by prefixes:
    - ***weiter**laufen (run on)* - fn:run&continue
    - ***los**laufen (start running)* - fn:run&start
    - ***ver**glimmen (stop burning)* - fn:burn&stop
  - 7,316 distinct role/complement-specified "syntactic" frames
- Non-verbal predicates
  - 1,400 nouns, 400 adjectives
  - systematic frame transfer from verbs to deverbal nouns and participle adjectives
    - *erkranken --> Erkrank**ung** (falling ill):* inherits 'sick' frame, preposition trigger *(Erkrankung **an**)* and §CAU argument role

Lexical support for the frame annotator:
syntactic and semantic slot restrictions

- e.g. *bestehen:* 5 meanings
  - *'pass' [an exam]* (accusative-monotransitive): <FN:**succeed**/S**§AG**'H/O**§TH**'occ>
  - *'consist of'* (PP-monotransitive: *b. aus*): <FN:**consist**/S**§HOL**'cc/P-aus**§PART**'cc|H>
  - *'insist on'* (PP-monotransitive *b. auf*): <FN:**demand**/S**§SP**'H/P-auf**§TH**'cc|act>
  - *'be'* (PP-monotransitive: *b. in*): <FN:**be_copula**/S**§TH**'ac|act/P-in**§ATR**'ac|act>
  - *'persist'* (intransitive): <FN:**persist**/S**§PAT**'conv|build|inst>
- for complements other than np's and pp's, syntactic form or POS can be specified instead of semantic type:
  - 'fcl' – finite clause, 'icl' – non-finite clause, 'num' – numeral
- only 834 valency patterns were sense-ambiguous
  - --> 92-93% of verbs could in theory be sense-disambiguated using syntactic clues alone

Semantic roles

- 44 atomic semantic roles
- 88 combinations, e.g. §AG-EXP subj. of *zuhören (listen)*

	Semantic role	**surface verb args %**	**secondary v- args %**	**all surface args %**
§TH	Theme	18.88	20.67	**36.17**
§ATR	Attribute	8.13	1.32	8.91
§LOC-TMP	Point in time	8.64	0	6.33
§MNR	Manner	8.13	0	5.49
§LOC	Location	6.24	2.90	5.30
§AG	Agent	**9.29**	**38.63**	5.72
§EXT	Extension	1.89	0.05	3.04
§META	Meta adverbial	3.95	0	2.59
§COG	Cognizer	4.01	**8.21**	2.27
§DES	Destination	2.13	0.76	1.82
§BEN	Beneficiary	2.56	1.32	1.81
§PAT	Patient	2.44	**4.33**	1.61
§REFL	Reflexive	2.48	0	1.44
§ID	Identity	0.01	0	1.21
§SP	Speaker	1.95	**6.57**	1.17
§CAU	Cause	1.49	1.24	1.02
§ACT	Action	1.34	1.40	2.19
§REC	Recipient	0.94	0.94	1.75
§EV	Event	1.18	1.61	1.56
§EXP	Experiencer	1.32	2.63	1.31
§DON	Donor	0.12	0.31	0.07

Light / non-role complements

- verb particles -- syntactic dummy tag (MV<), no role
  - *sie machte das Licht **aus** (she turned **off** the light)*
    lemma: "**aus**machen" (turn off), fn:deactivate
- support verbs: complement-based semantics and dependents -- full syntactic tag, dummy role (§INC)
  - *jmd. **Hilfe** leisten (help sb.),* fn:help, cp. nominal frame: *Hilfe für (help for)*
- PP incorporates (§INC on the noun, blocks other roles)
  - ***auf der Strecke** bleiben* (be lost, 'stay on the road'), fn:disappear
  - ***in Kraft** treten* (come into effect, 'step into power'), fn:activate

The frame annotator

- run as an additional module after GerGram morphosyntactic annotation
- uses the same formalism as GerGram and DanGram (Constraint Grammar), with full structural and tag compatibility with both parsers
- frame choice triggered by syntactic and semantic clues (GerGram tags) in iterative disambiguation and mapping steps
  - e.g. <FN:**tell**/S**§SP**'H/D**§REC**'H/O**§MES**'fact|sem-s|fcl> *(e.g. melden, zutragen)*
  - presence of a finite clause object (O:fcl) triggers this frame, if there is no other frame with O:fcl for the same lemma
  - field-based assignment of roles:
    - subject (S) --> §SP (speaker)
    - fcl object (O) --> §MES (message)
    - dative object (D) --> §REC (receiver)

Frame mapping and disambiguation

- 1. Frame template mapping (disambiguation through lemmatization)
  - *er nahm den Bus* (he took the bus), lemma: nehmen <FN:take/...>
  - *er nahm 5 kg zu* (he put on 5 kg), lemma: zu|nehmen <FN:increase/...>
  - *er nahm ihr die Aufgabe ab* (he relieved her of the task) lemma: ab|nehmen <FN:rid/...>
- 2. Frame template selection
- 3. Frame template removal
- 4. Role instantiation
- 5. Mapping of free roles

Selection and removal rules

- removal is simpler, safer and more robust than selection, because a single mismatch can trigger the former
- lexical matches are safest, e.g *Wert* (worth) in:
  - *legen* <FN:**mind**/S**§COG**'H/O-Wert**§INC**/P-auf**§TH**'all>
- syntactic functions are relatively safe, but not always expressed (check for competing lower-valency frames)
- most important are semantic slot fillers, to disambiguate frames with identical valency skeletons
  - shallow noun ontology with 200 categories, e.g. <Hprof> (profession), <Hfam> (family member), <Hideo> (ideological), <sem-r> (readable), <sem-c> (concept), <sem-s> (sayable)
  - to allow for fuzzy matches, the grammar lumps tags into umbrella categories, e.g. 'HUMAN', 'THING', 'PLACE'
  - progressive relaxation of the matching algorithm:
    - precise match --> umbrella match
    - all slots match --> some matches --> one or no slot matches
    - highest number of syntactic matches

Exploiting (secondary) dependencies

- in order to constitute semantic rather than syntactic links, dependencies need to be raised for prepositions and transparent nouns
- dependency trees can only be used directly if roles manifest as surface constituents, and these need to be nouns to allow semantic matches
  - in 45% of cases, there is no, or only pronominal, surface representation
- improvement: assign secondary/additional dependency links for relatives, infinitive subjects, coordination etc.

	filled slots (incl. secondary dep.)	**filled slots (primary dep. only)**
SUBJ	74.5 %	72.7 %
ACC	73.1 %	72.9 %
DAT	60.3 %	60.3 %
SC	97.7 %	97.7 %

Annotation example

Word	Lemma	Secondary tag, Frame	POS, morphology	Syntactic function	Semantic role	Dep. link
Ich *(I)*	ich		PERS	@SUBJ>	**§COG**	#1->2
verstehe *(understand)*	verstehen	<mv>**<FN:comprehend>**	V PR 1S FIN	@FS-STA		#2->0
nicht *(not)*	nicht		ADV	@ADVL>		#3->2
warum *(why)*	warum	<clb><interr>	ADV	@ADVL>	**§CAU**	#4->7
es *(there)*	es		PERS	@S-SUBJ>	**§TH-NIL**	#5->7
Eltern *(parents)*	Eltern	<HH>	N nG P ACC	@ACC>	**§TH**	#6->7
				R:c-subj:17	R:sd-**COG**:17	
gibt *(are)*	geben	<mv>**<FN:exist>**	V PR 3S FIN	@FS-<ACC	**§TH**	#7->2
,	,		PU	@PU		#8->0
die *(that)*	die	<clb>**<rel>**	INDP nG P NOM	@SUBJ>		#9->17
die *(the)*	die	<def>	ART F S ACC	@>N		#10->11
Erziehung *(education)*	Erziehung	**<FN:teach>**	N F S ACC	@ACC>	**§ACT**	#11->17
ihrer *(their)*	sie	<poss>	DET nG P GEN	@>N		#12->13
Kinder *(children)*	Kind	<H>	N NEU P GEN	@N<	**§BEN**	#13->11
möglichst *(as possible)*	möglich	<jcan>	ADV SUP	@>A		#14->15
früh *(early)*	früh	<atemp>	ADV	@ADVL>	**§LOC-TMP**	#15->17
Fremden *(strangers)*	Fremder	<ADJ:jsoc><Q-><nadj>	N nG P ACC	@DAT>	**§REC**	#16->17
überlassen *(leave)*	überlassen	<mv>**<FN:allow>**	V INF	@FS-N<	**§ATR**	#17->6
				R:p-subj:6		
wollen *(want)*	wollen	<aux>**<FN:wish>**	V PR 3P FIN	@AUX		#18->17

Evaluation: Data

- Corpus: 2 years of Twitter (~ 2 billion words)
  - extraction of all main verb-lemmas and their semantic class frame (f >= 1000 for noise reduction)
  - 8894 lemma-frame combinations (= 202.4 million tokens)
  - Manual check for non-German words and POS errors: 7,726 real German verb frames, representing 6,127 lemmas and 193.4 million tokens
    - = half the German verb lexicon (= 99.9% token coverage according to Zipf's law)
  - 1.245 frame classes / verb lemma, ca. = lexicon distribution and therefore likely to be representative in spite of the frequency cut-off
  - ambiguity higher at the token-level: 3.126 frame senses / verb

Evaluation: Ambiguity and coverage

- coverage failures
  - token level: 1.11% no frame + 0.25% no surviving frame
  - type level: 5.88% (impact of very rare verbs)
- frame ambiguity
  - higher at the token-level: 3.126 frame senses / verb
  - unevenly distributed: 78.6% monosemous verbs, 10 most frequent verbs (10.36% of all verb tokens) are very ambiguous:

verb lemma	token count	frame senses
lassen	2824239	11
geben	2455458	10
machen	2124256	34
spielen	1457122	4
nehmen	1416502	24
sehen	1414451	5
kommen	1251055	13
bleiben	1250034	3
haben	1237781	8
halten	1226771	17

Evaluation: Performance

- random sample of tweets (9,054 parser tokens) annotated and manually evaluated
  - 884 main verb tags
  - 20 wrong POS, 1 wrong lemma, 1 aux/mv error), often due to spelling errors in the word or its context
  - 8 verbs not recognized as such
- frame tagger performance
  - coverage: 99% (1 verb OOV, 8 cases where the correct frame was not among the ones listed in the lexicon)
  - recall / precision:

	R	P	F-score
total incl. POS errors	90.7%	96.5%	93.6
ignoring POS errors	93.0%	97.4%	95.2

- comparison
  - English Twitter out-of-domain (Hartmann et al. 2017): 62.17% full frame identification
  - German SHALMANESER (Burchardt et al. 2009): 79% correct WSD
  - in-domain German SRL test data (CoNLL 2009):
    - without linguistic features (Do et al. 2018): F=73.5
    - with syntax-aware neural networks (Cai & Lapata 2019): F=82.7%

Conclusions and outlook

- New resource: a German framenet intended for direct integration into a parser pipeline
  - valency-based, "framenet light" approach
  - bilingual compatibility Danish-German
  - coverage on par with morphosyntactic parsing
  - robust frame sense annotation (F=93.6 for social media data)
- Future work
  - add missing senses to existing verb entries (now: precision better than recall)
  - reduce underlying tagging errors for POS and dependency in the face of non-standard orthography
  - test the assumption that other domains *without* orthographical problems should work as well, given the general nature of the underlying morphosyntactic parser
- Cross-lingual aspects
  - It might be possible to generalize the Danish-German parser interoperability and dictionary-based bootstrapping to further (related?) languages. Thus, work is ongoing for a compatible Portuguese framenet and annotator.

Info: **framenet.dk**
Demo: **visl.sdu.dk/de**

contact: *eckhard.bick@mail.dk*