Finding Depictive Secondary Predicates in Large Web Corpora

Benjamin Burkhardt[1], Laura Kallmeyer[1] & Timm Lichte[1,2]

[1]CRC991: University of Düsseldorf
[2]University of Tübingen

Corpus Linguistics 2019, July 26th 2019

SFB 991

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Secondary Predicates

SECONDARY PREDICATE (**SP**): a typically sentence final, adjectival element that predicates one of the (main) verbal predicate's arguments; we call the predicated element the **TARGET**.

RESULTATIVES (**RSPs**) characterize states that are brought about by the event that is expressed by the main verb.

(1) Sean stomped the can$_i$ flat$_i$.

DEPICTIVES (**DSPs**) express properties that hold for at least some part of the event time, but do not immediately result from the verb event.

(2) Tom ate the pizza$_i$ cold$_i$.

- SP data in the literature is mostly introspectively constructed and relies on native speakers' grammaticality judgments.
- To our knowledge, a systematic corpus study of such phenomena has not been conducted so far.
- The goal of this contribution is to initiate such a study.

$\Rightarrow$ **How can SP data be detected in large web corpora?**

$\Rightarrow$ **What theoretically predicted SP data can we actually observe?**

Outline

Outline

Predicted Targets: Subject & Object

Based on their semantic compatibility, depictives can target the subject or the object.

(3) a. Kim ate the steak$_i$ raw$_i$.
    b. Kim$_i$ ate the steak hungry$_i$.

If both verbal arguments are semantically compatible with the depictive **TARGET AMBIGUITY** arises.

(4) Kim$_i$ ate the apple$_j$ unwashed$_{i/j}$.

DEPICTIVE STACKING is possible, but generally seems to decrease acceptability.

(5) a. ? Kim$_i$ ate the steak$_j$ raw$_j$ hungry$_i$.
    b.?? Kim$_i$ ate the steak$_j$ hungry$_i$ raw$_j$.
    c.?? Kim ate the steak$_j$ raw$_j$ salted$_j$.

Depictives may target unrealized agents, see (6-a), or theme arguments, see (6-b).

(6) a. The book$_j$ is to be read naked$_{i/*j}$.

    b. We$_i$ usually bake gluten-free$_{*i/j}$.

Predicted Impossible Targets

Oblique verbal arguments, i.e. non-direct objects and PP-objects, do not constitute viable targets.

(7) a. The cash machine$_i$ gave John$_j$ the money$_k$ hungry$_{*i/*j/*k}$.

   b. Peter crashed into him$_i$ tired$_{*i}$.

Depictives cannot target **modifying constituents** like PP-adjuncts.

(8) John drilled a hole with a power tool$_i$ new$_{*i}$.

Neither the **embedded** genitive noun in (9-a) nor the single conjuncts in (9-b) constitute viable targets.

(9) a. John met Maria's$_i$ father naked$_{*i}$.

   b. [John$_i$ and Paul$_j$]$_k$ met [Maria$_m$ and her boyfriend$_n$]$_o$ naked$_{*i/*j/k/*m/*n/o}$.

Theoretical predictions

possible targets of depictives

	subject	object	macroroles	all arguments	non-arguments	unrealised	stacked	embedded
Pylkkänen (2002)	X	X						
Geuder (2004)	X	X						
Müller (2008)*	X	X		X		X	X	
Burkhardt et al. (2017)	X	X	X			X	X	

- Most examples in the literature are either constructed or anecdotal.
- Lack of systematic empirical studies of the phenomenon (acceptability judgment studies and corpus studies)

Main Question: Which SP constructions are observable in real data?
k

Outline

Outline

Expectation: SPs are a relatively rare phenomenon

→ A large and stylistically diverse corpus of English is required.

- ENCOW16AX is a large corpus of English with ≈ **9,6 Billion tokens** (scrambled on sentence level).
- Since it is web-based, it covers **a wide variety of Englishes** and both formal and colloquial texts.
- The corpus creation pipeline is open source and the corpus itself can be used **free of charge**. (`https://corporafromtheweb.org/`)
- Annotation layers:
    - Lemmatization (TreeTagger)
    - Part-Of-Speech tags (Penn Treebank tag set, TreeTagger)
   - **Syntactic dependencies (Stanford dependencies, MaltParser)**

Selection of Depictives

- We conduct our corpus study by using a selection of 10 frequent adjectival stage level predicates, e.g. *naked*, *hot*, *cold*, and *happy*.
- A preliminary analysis indicates that **stage level predicates** are more likely to appear in SP constructions.
- The stage and individual level predicate distinction is not rigid, however the concept proves useful in our study.

Filter 1: Adjectives with a verbal head

SPs appear in positions where usually adverbial modifiers could appear.

(10) a.   Kim left$_i$ the room angry$_i$.   (Adverbial)
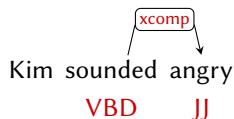   b.   Kim$_i$ left the room angry$_i$.   (Depictive)

Kim   left   the room   angry / angrily
    VBD   JJ   RB
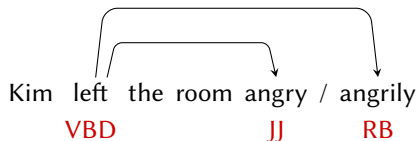
⇒ **We are looking for adjectives (JJ) with a verbal head (VB, VBD, VBG, VBN, VBP, VBZ).**

SPs have a modifying dependency relation to their head.

(11) a. Kim left the room angry.   (Depictive)
     b. Kim sounded angry.   (Argument)
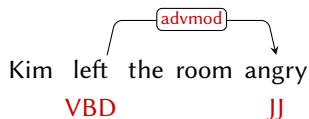
⇒ **We are looking for adjectives (JJ) with an advmod dependency relation to their verbal head (VB, VBD, VBG, VBN, VBP, VBZ).**

Filter 3: No copula verbs

We observe a lot of wrong parses of the MaltParser that include a copula construction.

(12) a. Kim got very drunk.   (Copula)
    b. Kim$_i$ left very drunk$_i$.   (Depictive)

Kim left / got very drunk
     VBD   VBD   JJ
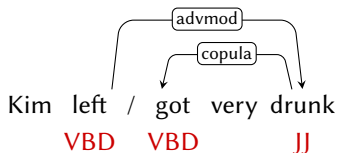
⇒ **We exclude sentences with copula and "copula-like" lemmata.**
   be, get, make, become, how, keep, stay, feel, look, seem, remain, appear

Why we don't use NoSketch Engine

ENCOW only includes the following annotation information for each token:

`[token-index, token, lemma, POS tag, head index, dependency]`

However: For our filter, we need access to all annotation information of the adjectives head token. The `head index` does not suffice!

⇒ To query/filter for our data via NoSketch Engine, **ENCOW would have to be reformated** to include all the head information.

⇒ Therefore, we used our own Python scripts instead.

Outline

The following three steps are applied to extract samples for each DSP candidate item (implemented in Python) :

Step	Filter	Sentence Count
0.	None (complete ENCOW16AX)	$\approx$ 421 Million
1.	POS: JJ, DepRel: advmod	$\approx$ 4 Million
2.	No copula constructions	$\approx$ 2.4 Million
3.	Sampling	200 per adjective

To receive samples of 200 sentences we query the subcorpus without any Copula and apply some additional filters based on the adjective at hand.

MaltParser Annotation Results

Construction	naked	hot	cold	happy
Depictives	137/68,5%	39/19,5%	20/10%	32/16%
Actor oriented	120/60%	8/4%	9/4,5%	29/14,5%
Undergoer oriented	13/6,5%	21/10,5%	11/5,5%	3/1,5%
Unrealized target	4/2%	11/5,5%	–	–
Sentence initial	5/2,5%	–	–	–
Resultatives	9/4,5%	7/3,5%	–	1/0,5%
Copula(-like)	–	41/20,5%	23/11,5%	61/30,5%
Adverbial Uses	–	10/5%	5/2,5%	2/1%
Nominalizations	–	7/3,5%	29/14,5%	8/4%
Adnominal Uses	–	30/15%	55/27,5%	19/9,5%
Multi Word Expression	–	21/10,5%	–	–
Other	41/20,5%	34/17%	53/26,5%	60/30%
Out	13/6,5%	11/5,5%	15/7,5%	17/8,5%

■ The annotation was done manually by the speaker.

Discussion: MaltParser Annotation Results

- No instance of target stacking, but some quirky data.
  - **Embedded target?**

    (13) a. Images of women$_i$ swirl naked$_i$ on the ceiling [...].

    b. Naked$_i$, you can see her$_i$ ribs through the dusty white of her back.

  - **Unrealizable implicit target?**

    (14) It feels so much better naked$_i$.

- Still there is a lot of noise due to errors in the syntactic annotation.
- Therefore, we tried the same filter but with dependencies from the Standford Parser.

Outline

Stanford Parser Annotation Results

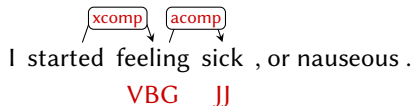| Construction	naked	hot	cold	happy
Depictives	148/74%	27/18,5%	25/10%	38/19%
Actor oriented	122/61%	1/0,5%	9/4,5%	29/14,5%
Undergoer oriented	23/11,5%	22/11%	16/8%	9/5,5%
Unrealized target	–	4/2%	–	–
Sentence initial	–	–	–	–
Resultatives	12/6%	1/0,5%	–	–
Copula(-like)	9/4,5%	57/28,5%	41/20,5%	94/47%
Adverbial Uses	–	34/17%	19/9,5%	1/0,5%
Nominalizations	2/1%	1/0,5%	30/15%	3/1,5%
Adnominal Uses	1/0,5%	6/3%	13/7,5%	5/2,5%
Multi Word Expression	–	11/5,5%	–	2/1%
Other	17/8,5%	48/24%	55/22,5%	46/23%
Out	11/5,5%	15/7,5%	17/8,5%	11/5,5%

■ The annotation was done manually by the speaker.

Quirky Dependency Parses

- Adverbial Modifier Dependency

... patients who did n't feel sick to begin with ...

rcmod ... advmod

VB   JJ
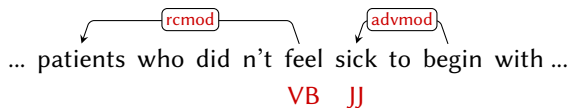
- Adjectival Complement Dependency

I started feeling sick , or nauseous .

xcomp   acomp

VBG   JJ

- Sentential Complement and Copula Dependency

It actually makes me feel sick ...

xcomp   cop

VB   JJ

Outline

Implications & Future Work

- We presented an ongoing empirical study on the usage of DSPs in web corpora.
- Filters use POS information and syntactic dependencies
- Unfortunately, the performance of the two used parsers (MaltParser, Stanford Parser) is poor wrt. secondary predicates.
- Still interesting data that seem unexpected from point view of current literature.
- Contrast between findings for *naked* and the other adjectives

Future work:

- Adjust filters? Use another parser (Berkley)? More principled adjective selection?
- More fine grained semantic analysis of the relation between the depictive and its target. –> semantic conditions on DSPs
- Analysis of semantic properties of "depictive friendly" verbs

Thanks for your kind attention!

Pssst, ask me about the implemented grammar resource and corpus filter scripts on Github.

[1] Burkhardt, Benjamin, Timm Lichte & Laura Kallmeyer. 2017. Depictives in English: An LTAG approach. In *Proceedings of the 13th international workshop on tree adjoining grammars and related formalisms*, 21–30. Umeå, Sweden: Association for Computational Linguistics. `http://www.aclweb.org/anthology/W17-6203`.

[2] Geuder, Wilhelm. 2004. Depictives and transparent adverbs. In Jennifer R. Austin, Stefan Engelberg & Gisa Rauh (eds.), *Adverbials: The interplay between meaning, context, and syntactic structure*, 131–166. Amsterdam: John Benjamins Publishing Company.

[3] Müller, Stefan. 2008. Depictive secondary predicates in German and English. In Christoph Schroeder, Gerd Hentschel & Winfried Boeder (eds.), *Secondary predicates in Eastern European languages and beyond* (Studia Slavica Oldenburgensia 16), 255–273. Oldenburg: BIS-Verlag.

[4] Pylkkänen, Liina. 2002. *Introducing arguments*. Cambridge, MA: Massachusetts Institute of Technology dissertation.