WG4

Verbal Multiword Expressions in Treebanks

Veronika Vincze and István Nagy T.
University of Szeged, Hungary
{vinczev,nistvan}@inf.u-szeged.hu

Verbal Multiword Expressions

Verb-Particle Constructions (VPCs)

- Verb + particle: *show off*
- Particle + verb: *aufstehen* "to stand up'' (German)
- Spelt as one word/two words
- Compositional or not

Light Verb Constructions (LVCs)

- Verb + noun: *take care*
- Noun + verb: *Hilfe gewähren* "to grant aid'' (German)
- Semi-compositional: the sense of the noun is dominant
- Verb + object, verb + PP, verb + other arguments

Idioms

- Verb + argument(s):
  - *kick the bucket*
  - *see red*
  - *the cat is out of the bag*
- Non-compositional: none of the words preserve their meaning

English VPCs in the Penn Treebank

- The special relation of the verb and particle within a VPC is distinctively marked in the Penn Treebank (Marcus et al., 1993)
- The **particle** is assigned a specific part of speech tag (**RP**)
- It also has a specific syntactic label (**PRT**)
- Marked both at the level of morphological and syntactic annotations

Turn the light off.
VB DT NN RP
(S (NP-SBJ *) (VP turn (NP the light) (PRT off)))

Hungarian VPCs in the Szeged Treebank

- VPCs in Hungarian with specific orthographical rules:
  - **Spelt as one word** when the particle is right before the verb
  - They can be **separated** due to word order
- In the latter case, they are marked as the preverb of the verb both at the morphological (**Rp**) and syntactic levels (**PREVERB**)

***fel fog ébredni* "he will wake up"**
Rp Vaip3s---n Vmn
(VP (PREVERB fel) (V fog) (INF ébredni))

German VPCs in the TIGER Treebank

- VPCs in German with specific orthographical rules:
  - **Spelt as one word** when the particle is right before the verb
  - They can be **separated** due to word order
- In the latter case, they are marked as the particle of the verb both at the morphological (**PTKVZ**) and syntactic (**SVP**) levels

***Er nimmt es auf* "he picks it up"**
PPER VVFIN PPER PTKZV
(S (SB Er) (HD nimmt) (EP es) (SVP auf))

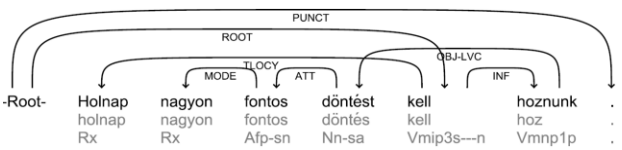
French MWVs in the French Treebank

- The French Treebank contains explicit morphological annotations for MWEs (Abeillé et al. 2003)
- Verbal MWEs are grouped in the treebank according to their POS patterns (like V N, V P N etc.)

***avoir lieu* "to take place"** VW+VN
***entrer en vigueur* "to enter into force"** VW+VPN
***avoir beau jeu* "to have an easy job"** VW+VAN

Hungarian LVCs in the Szeged Treebank

- The Szeged Constituency Treebank has been manually annotated for LVCs (Vincze and Csirik, 2010)
- No syntactic restrictions on the internal structure of LVCs (Vs. German)
- In the dependency version, dependency relations between the two members of LVCs were enhanced with **LVC-specific relations** (Vincze et al., 2013)
- The annotation found in the dependency treebank is more detailed
- In the constituency treebank, it was only the **syntactic boundaries of the phrases** that were marked
- In the dependency treebank the **inner syntactic relation** of the MWE is also encoded (see the sentence below "Tomorrow, we will have to make a very important decision.")



German LVCs in the TIGER Treebank

- In the TIGER corpus (Brants et al. 2004), LVCs that consist of a **verb and a prepositional phrase** are annotated
- The **PP** is marked with the relation **CVC** (collocational verb construction)
- **Verb-object pairs** are **excluded** from the annotation

***Abschied nehmen* "to take leave"** – not an LVC here
***zur Diskussion bringen* "to discuss"**
(zur Diskussion)$_{CVC}$ bringen

Comparisons

LVCs

- Annotated in 3 languages: German, Hungarian and French
- Level of annotation:
  - French: morphological level
  - Hungarian & German: syntactic level
- Coverage:
  - Hungarian & French: full coverage
  - German: restricted definition of LVC

VPCs

- Annotated in 3 languages: German, Hungarian and English
- Level of annotation:
  - morphological level
  - syntactic level
- Coverage:
  - English: full coverage
  - Hungarian & German: only if separated from the verb

References

Abeillé, Anne; Clément, Lionel; Toussenel, François 2003: Building a treebank for French. In Abeillé, Anne (ed.): *Treebanks: building and using parsed corpora*. Kluwer, Chapter 10.

Brants, Sabine; Dipper, Stefanie; Eisenberg, Peter; Hansen, Silvia; König, Esther; Lezius, Wolfgang; Rohrer, Christian; Smith, George; Uszkoreit, Hans 2004: TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation* 2, 597–620.

Marcus, Mitchell P.; Santorini, Beatrice; Marcinkiewicz, Mary Ann 1993: Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2): 313–330.

Vincze, Veronika; Csirik, János 2010: Hungarian Corpus of Light Verb Constructions. In: *Proceedings of COLING 2010*, Beijing, China, pp. 1110-1118.

Vincze, Veronika; Zsibrita, János; Nagy T., István 2013: Dependency Parsing for Identifying Hungarian Light Verb Constructions. In: *Proceedings of IJCNLP 2013*.