Verbal Multiword Expressions in Treebanks

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

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 Vaijp3s—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 preverb of the verb both at the morphological (PTKVZ) and syntactic (SVP) levels
  
    Er nimmt es auf “he picks it up”
    
    PPER VVFIR PPER PTKVZ
    
    (S [SB Er] (HD nimmt) (EP es) (SVP auf))

French MWVs in the French Treebank
- The French Treebank contains explicit morphological annotations for MWEs (Abellé 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

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