Towards a Dutch Parseme Corpus

Gosse Bouma and Jan Odijk and Carole Tiberius

University of Groningen, University Utrecht, Dutch Language Institute

university of groningen

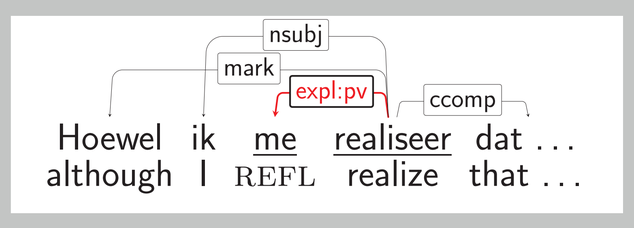
Utrecht University

/instituut voor de Nederlandse taal/

Conversion from UD-Dutch Lassysmall and Alpino

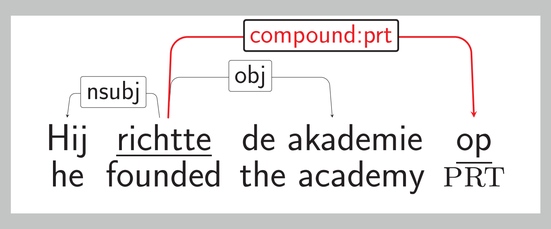
- Parseme corpora can be constructed by adding an annotation layer to Universal Dependencies corpora (Savary et al, 2023)
- We follow this approach by automatically adding MWE annotation to UD Dutch LassySmall and Alpino
- The converted data still requires manual verification in FLAT

Inherent Reflexive Verbs (IRV)



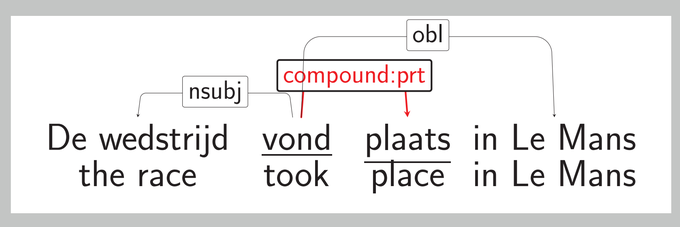
- Mark verbs with an expl:pv dependent as IRVs

Verb-Particle Constructions (VPC)

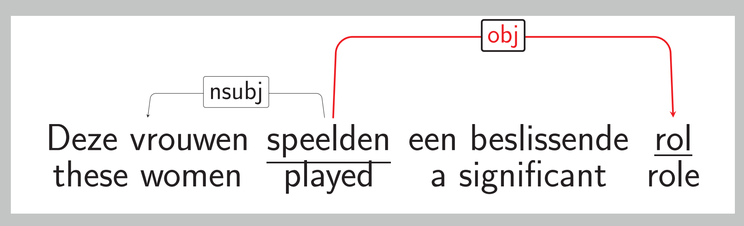


- Verbs with a ADP, ADJ, or ADV compound:prt dependent are VPC
- When written as a single word (*oprichtte, founded*), the underscore in the lemma identifies the particle (*op_richten, to found*)

Light Verb Constructions (LVC)

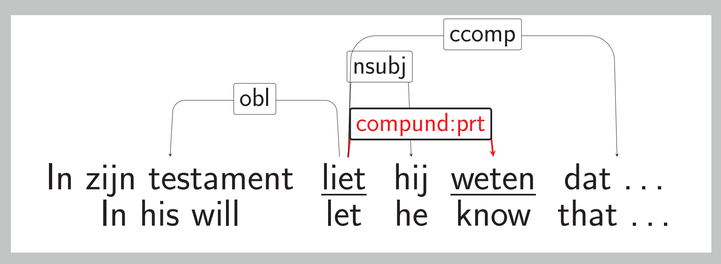


- Verbs with a NOUN compound:prt dependent are LVC



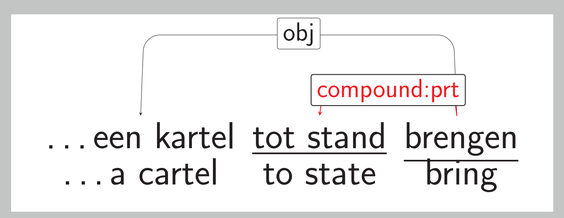
- Verbs with an obj dependent that is listed explicitly in the Alpino lexicon are LVC

Multi-Verb Constructions (MVC)

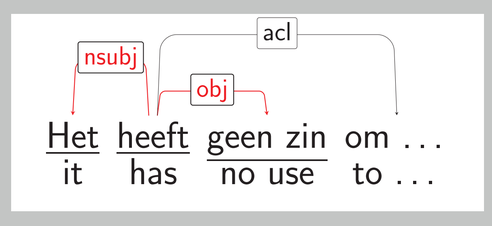


- Verbs with a verbal `compound:prt` dependent are MVC

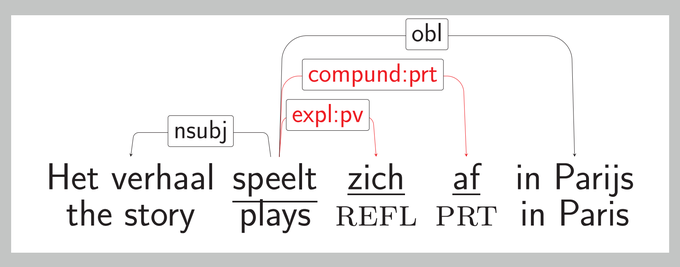
Verbal idioms (VID))



- Verbs with a compound:prt dependent that is phrasal are VID

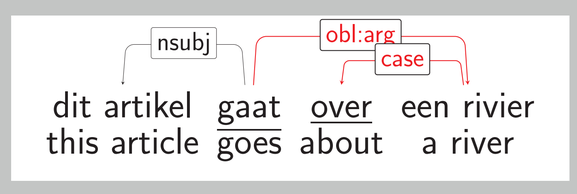


- Verbal idiomatic expressions involving a subject, (non-predicative) object, or oblique dependent of a verb that is listed with this pattern in the Alpino-lexicon, e.g. het heeft geen zin, 'it is useless', are VID



- Cases where two or more phenomena co-occur, i.e. when a verb has both a verbal particle and a predicative noun as dependents, are VID

Inherent Adpositional Verbs (IAV)



- IAV are an optional extension of the Parseme verbal MWE guidelines.
- The next release of the UD Dutch corpora identifies these as verbs with an obl:arg dependent, following UD guidelines

Validation and Results

- All automatic annotation decisions were validated by creating Grew-match queries that search for the various types of MWE in Dutch UD Dutch corpora (see abstract for corresponding links)
- We also plan to validate results against the DUCAME MWE finder (Odijk et al, to appear)

Class	Alpino	LassySmall
VPC	2937	1025
IAV	1372	570
VID	1347	419
LVC	354	98
IRV	188	90
MVC	41	4
Total	6239	2206

- Statistics for various MWE classes after automatic annotation of the Dutch UD corpora
- The distinction between VPC.full and VPC.semi and LVC.full and LVC.cause needs to be made manually

g.bouma@rug.nl