Multiword Expression Identification with Recurring Tree Fragments and Association Measures

Proceedings of the 11th Workshop on Multiword Expressions

WG3

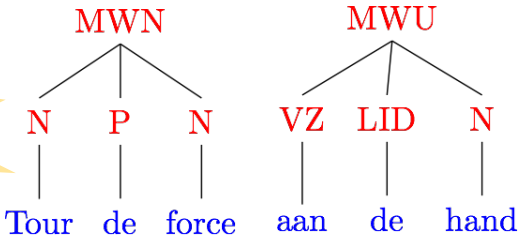
Parseme 5th GM
Iași, 23-24 September 2015

Federico Sangati
federico.sangati@gmail.com

Andreas van Cranenburgh
andreas.van.cranenburgh@huygens.knaw.nl

Investigate ways of automatically detecting **MWEs in treebanks** by searching for **recurring patterns**.

Only contiguous and flat MWEs



French TB

Dutch TB

PARSING with 2-DOP (TSG)
(implicit identification of MWEs as in Green et. al 2011)

SUPERVISED

SMALL MANUAL TREEBANK **with** ANNOTATED MWE

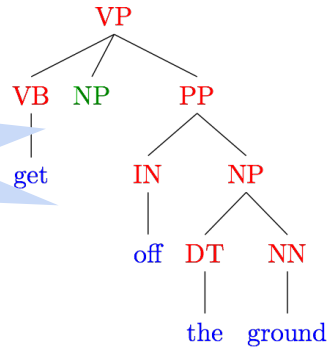
UNSUPERVISED

LARGE AUTOMATIC TREEBANK **without** ANNOTATED MWE

RECURRING FRAGMENTS SEEKER

Syntax compositionality index
(high when a fragment is often seen as a single block, low when it is typically generated by smaller units)

Also discontiguous and hierarchical MWEs



Annotated English Gigaword

Ranking of fragments to identify MWEs

Association measures defined on tree fragments (e.g., PMI, LLR)