# Identifying Multi-Word Expressions with Recurring Tree Fragments


Federico Sangati
FBK, Trento & Edinburgh Univ.
sangati@fbk.eu

Andreas van Cranenburgh
Huygens ING, Royal Netherlands Academy of Arts & Sciences; ILLC, Univ. of Amsterdam. andreas.van.cranenburgh@huygens.knaw.nl

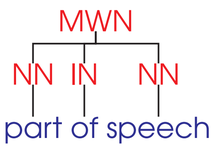
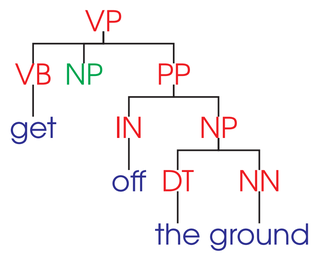
Institute for Logic, Language and Computation
University of Amsterdam


## Abstract


We investigate ways of automatically detecting MWEs in large treebanks:

- Arbitrarily large syntactic constructions extracted from a treebank; i.e., tree fragments, as in TSGs, cf. Green et al. (2013).
- Fragments may include any number of lexical units (L) and possible intervening gaps (X)
- Association measures over words select MWEs from candidate tree fragments


## Related Work

| | Ramisch et al. (2010) | Green et al. (2013) | This work |
|---|---|---|---|
| Unsupervised | YES | No | YES |
| Association measures | YES | No | YES |
| Syntax | POS tags | flat rules | hierarchical |
| Gaps | No | No | YES |
| Representation | ⟨ JJ_mountain, NN_bike ⟩ |  |  |





PARSEME Working Groups:

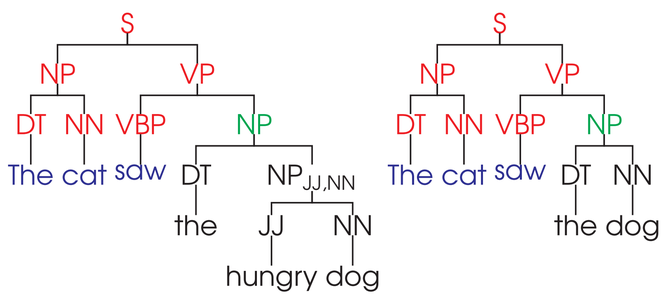
**WG3** - Statistical, Hybrid and Multilingual Processing of MWEs
Recurring fragments can be used for MWE-informed statistical parsing approach.

**WG4** - Annotating MWEs in Treebanks
Automatically derived MWEs, enriched with their syntactic structures, can be employed to automatically label existing treebank with MWE-informed tags, and can lead to the creation of resources such as MWE lexicons and valence dictionaries.

## Fragment Extraction

Using Tree Kernel Technique:

- Given a pair of trees, we can extract their *overlapping fragments*.
- When applied to a treebank, this yields a *set of recurring patterns*.
- Fragments can be seen as *building blocks* of the treebank.
- Can be extracted efficiently (Sangati et al., 2010; van Cranenburgh, 2014).

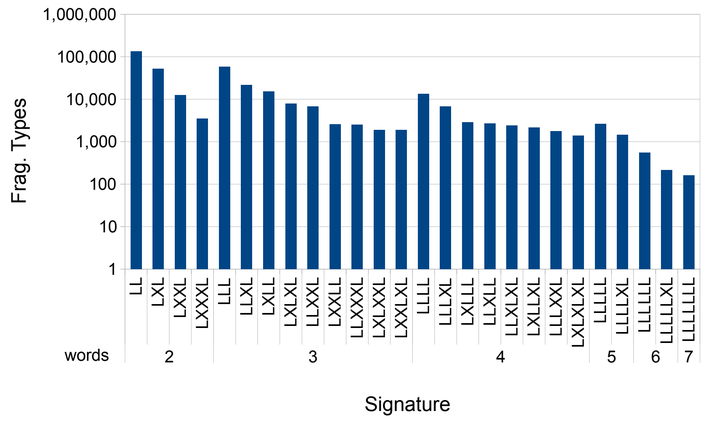



Source Code:
http://github.com/andreasvc/disco-dop

## Data

| Treebank | |
|---|---|
| Corpus | Automatically Annotated English Gigaword |
| Section | NYT |
| Sampling | Every 150 sentences |
| Size | 500K sentences |

| Fragment Counts | |
|---|---|
| Total Recurring Fragments | 4.3M |
| ≥ 1 content + 1 non-punct. word | 2.8M |
| freq. ≥ 5 | 400K |

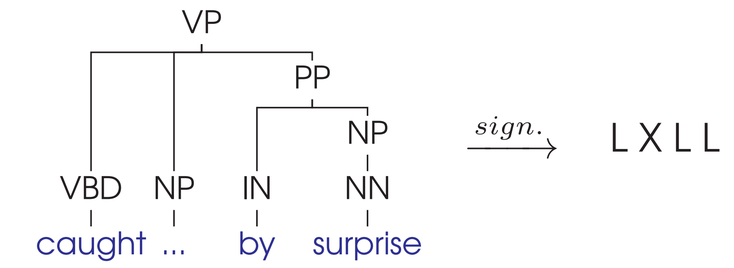



## MWE Selection

Per-Signature Multivariate Generalization of Pointwise Mutual Information (PMI):

$$\mathrm{PMI}(L_1, L_2, \ldots, L_n) = \log \frac{p(L_1, L_2, \ldots, L_n)}{\prod_{i-1}^{n} p(L_i)}$$

where $p(L_1, L_2, \ldots, L_n)$ is computed within the set of fragments sharing the same signature (e.g., L X L L).

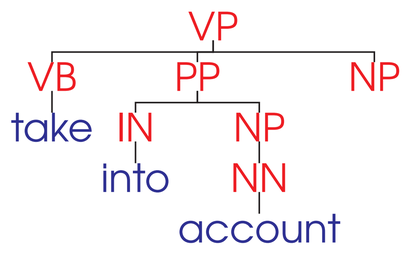



$$\mathrm{PMI}(\text{caught,by,surprise}) = \log \frac{p(\text{caught,by,surprise})}{p(\text{caught}) \cdot p(\text{by}) \cdot p(\text{surprise})}$$

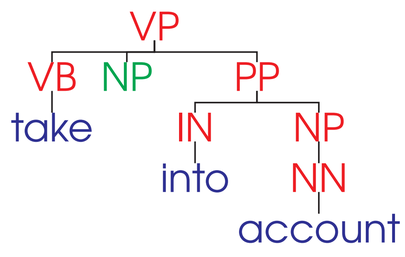
## Open Issues

- Signatures
  - differences: words, PoS tags, syntactic categories
  - outer categories (before/after lex. span)
- PMI for $> 2$ tokens
- Overlapping with sub/supersets of fragments
- Other association measures for syntactic trees
- Larger Treebank

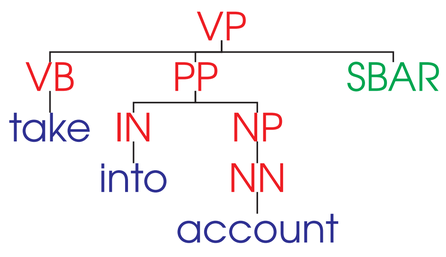
## Results




Freq. = 8




Freq. = 7




Freq. = 6

3 words (VB_take X L L)

| PMI | Freq. | Signature Pattern |
|---|---|---|
| 18.0 | 6 | VB_take NP IN_into NN_account |
| 14.6 | 6 | VB_take NP IN_for VBN_granted |
| 13.6 | 7 | VB_take DT NN_look IN_at |
| 12.9 | 6 | VB_take NP TO_to NN_court |
| 12.5 | 6 | VB_take NN RB_away IN_from |
| 12.4 | 17 | VB_take NP RB_away IN_from |
| 12.0 | 6 | VB_take JJ NN_action TO_to |
| 11.2 | 5 | VB_take NP RB_away IN_from |
| 10.5 | 6 | VB_take QP NNS_years TO_to |
| 8.3 | 10 | VB_take DT NN_time TO_to |

3 words (VB_take L L)

| PMI | Freq. | Signature Pattern |
|---|---|---|
| 15.3 | 13 | VB_take IN_into NN_account |
| 9.8 | 5 | VB_take NN_responsibility IN_for |
| 9.7 | 8 | VB_take NN_credit IN_for |
| 9.3 | 12 | VB_take DT_a NN_look |
| 8.4 | 88 | VB_take NN_advantage IN_of |
| 8.4 | 7 | VB_take NN_place IN_on |
| 8.3 | 6 | VB_take NN_effect IN_in |
| 8.1 | 14 | VB_take NNS_steps TO_to |
| 8.0 | 6 | VB_take DT_a NN_chance |
| 7.9 | 16 | VB_take NN_place IN_in |

## References


Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a framework for multiword expression identification. In *Proc. of LREC'10*.

Federico Sangati, Willem Zuidema, and Rens Bod. 2010. Efficiently Extract Recurring Tree Fragments from Large Treebanks. In *Proc. of LREC'10*.

Andreas van Cranenburgh. 2014. Linear average time extraction of phrase-structure fragments. Presented at CLIN 2014, Leiden, The Netherlands.