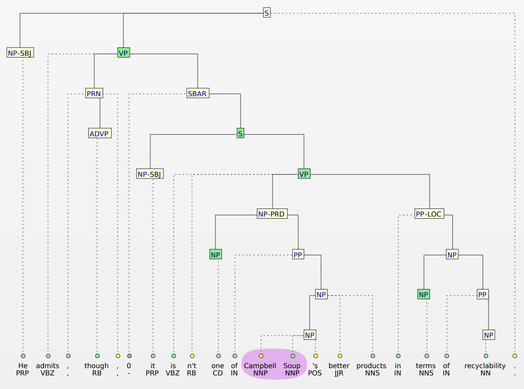
Identification of Multiword BBN Named Entities in Dependency Annotation of Wall Street Journal

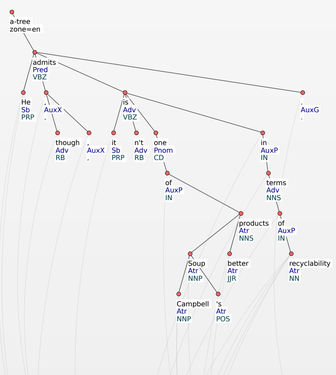
Eduard Bejček, Pavel Straňák
Charles University in Prague, MFF,
Institute of Formal and Applied Linguistics
{bejcek,stranak}@ufal.mff.cuni.cz

WORK IN PROGRESS

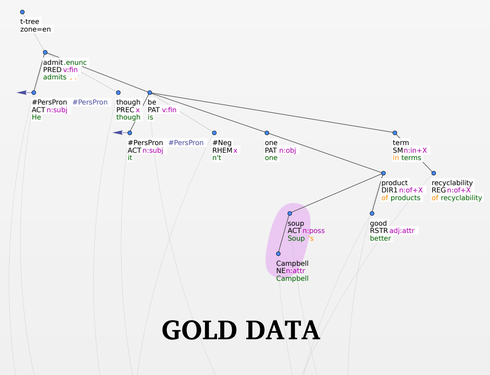
☞ Data — Wall Streeet Journal + BBN entities + Prague Czech-English Dependency Treebank



automatically transformed to dependency syntax



semiautomatically transformed to deep syntax



BBE entities inserted into the deep layer by Toman

GOLD DATA

WSJ + Vadas and Curran's NP structures
+ BBN named entities

PCEDT is a parallel treebank,
each node is linked to Czech node.
That will be used in the future.

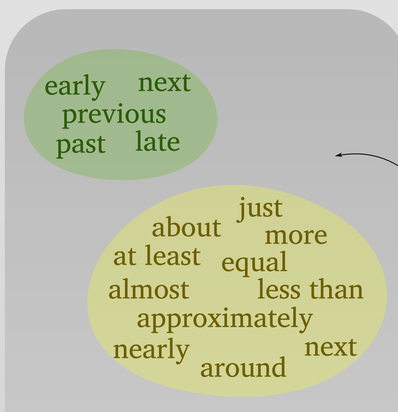
☞ Method — syntactic-lexicon based

☞ similar to method presented in Athens

☞ create a **lexicon** out of all multiword BBN named entites (MWNEs) in training data
- i.e. more than one node; each MWNE has to be connected graph
- add deep syntactic information with lemmas to each MWNE
- types of nodes (=functors) not used yet
- search for the structures from the lexicon

☞ **data sparseness**
- substitute values from a group (numbers, decades, decimals, months etc.) by a single value
- two sets of "universal attributes", that can complement numbers or time expressions
- there are many other improvement possible
  (e.g. any kombination composing a CARDINAL beneath a percent sign is a PERCENT),
  but it starts to be rule based

☞ if **more than one tag** was associated with the tree structure in the lexicon,
we use the most frequent for the time being

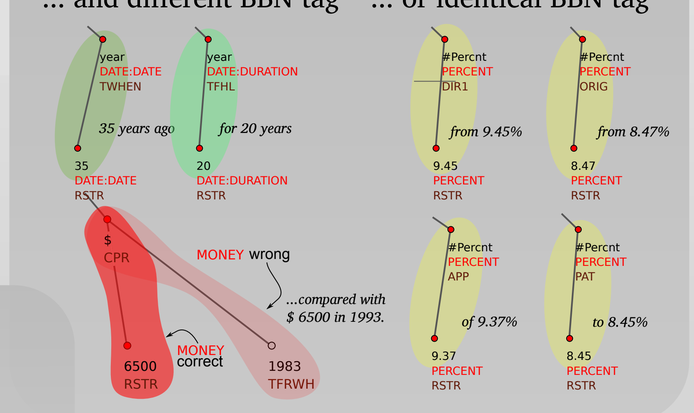


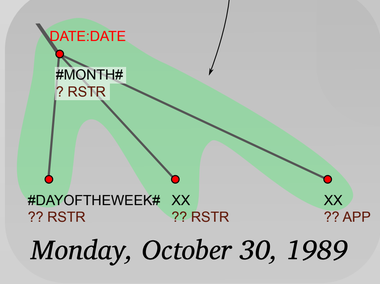
extracted from CzEng corpora

It's not clear which additional information to use.
E.g. functors and/or auxiliary words.
In some cases it can help,
in others it bears no information at all.

Different auxiliaries,
different functors...

... and different BBN tag ... or identical BBN tag

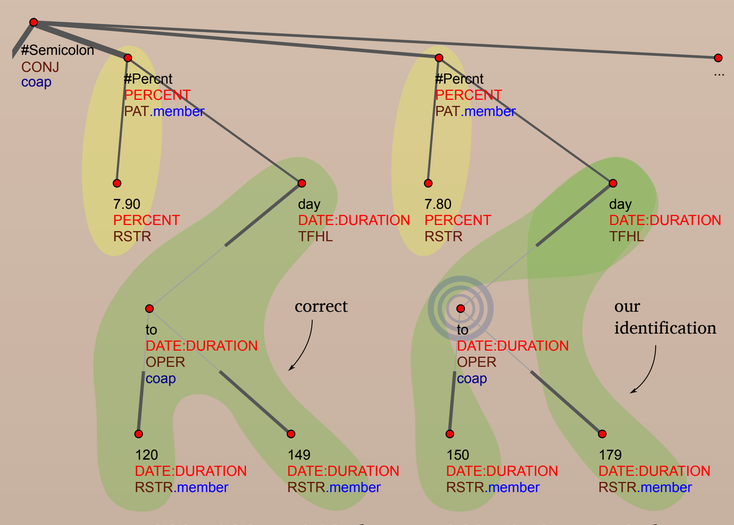




Monday, October 30, 1989

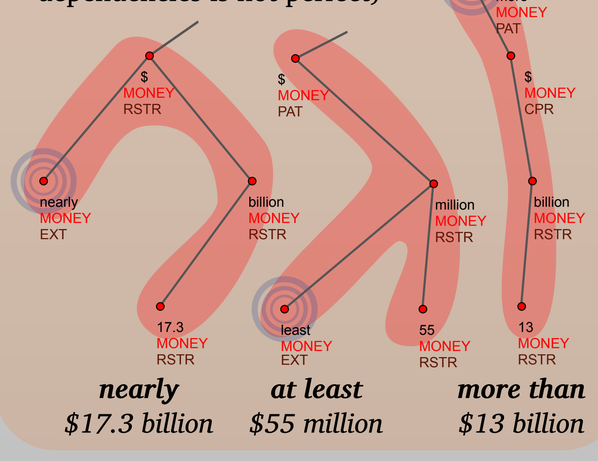
☞ Evaluation & Examples

Basically, we can only find, what we have seen.
No unexpectedly changed structures,
no **unseen named entities** (e.g. name of person mentioned several times, but in a single document).



...; 7.90% 120 to 149 days; 7.80% 150 to 179 days; ...
...because no range was seen in the train data

Inconsistencies in PCEDT
(head finding algorithm for dependencies is not perfect)



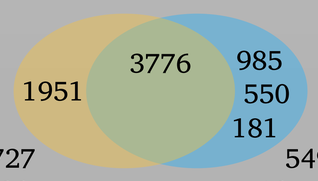
nearly *$17.3 billion* ***at least*** *$55 million* ***more than*** *$13 billion*

Ten-fold cross-validation:

Found in average: 5491 MWNEs
- 3776 correct
- 985 completely wrong
- 550 with an intersection
- 181 with a wrong tag

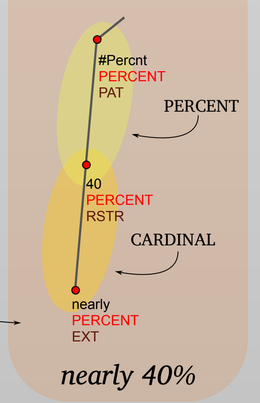
Not identified at all: 1951

... according to gold data with errors



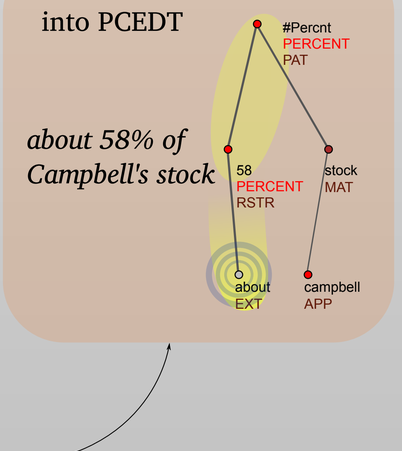
5727 gold MWNEs — 5491 MWNEs identified by us

"Per partes"



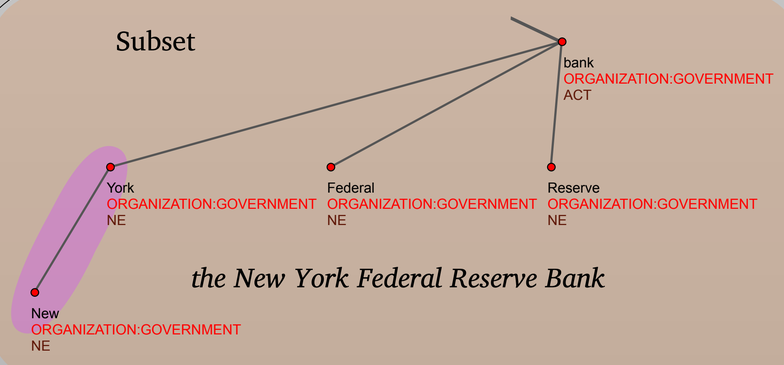
nearly 40%

Errors in transformation into PCEDT



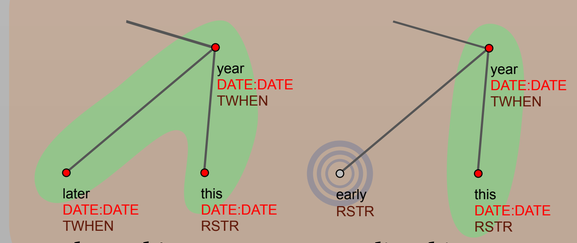
about 58% of Campbell's stock

Subset



the New York Federal Reserve Bank

Inconsistencies in BBN annotation



later this year *earlier this year*

LINDAT CLARIN

This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

This research has been supported by grant COST CZ LD14117 of the Ministry of Education, Youth and Sports of the Czech Republic.

PARSEME
4th General Meeting,
Valletta, Malta,
19 – 20 March, 2015. **WG4**