Annotation Issues

Kübra Adali, Hiwa Asadpour, Natalia Klyueva, Ivana Matas Ivankovic, Carlos Ramisch, Veronika Vincze

Victoria Rosen

Variability

Challenge 1

How to automatically identify inflected forms of MWEs in a corpus of an agglutinative language while the inflected forms can not be a real MWE?

While only their base forms are present in the lexicon, i.e. how to match a lexicon against a corpus?

Agglutinative Languages

In agglutinative languages, the inflection can not be stopped!!!!!

❖ Sağlıklılaştıramadıklarımızdansınız .

(You are one of the person that we could not cure).

Sağlık-lı-laştır-amadık - larımızdan - sınız

Lit : (health) (y) (make) (that we couldn’t) (one of the persons) (you are)

MWEs in agglutinative languages : Hungarian

A dékán újabb előadást
the dean new-COMP presentation-ACC
tartott szükségesnek
hold-PAST-3SG necessary-DAT
“The dean thought that another presentation was necessary.”

előadást tart
presentation-ACC hold
“to have a presentation”

valamilyennek tart valamit
somewhat-DAT hold something-ACC
“to regard something as something”

szépnek tartja a lányt
beautiful-DAT hold-3SG-OBJ the girl-ACC
“he thinks that the girl is beautiful”

MWEs in agglutinative languages : Turkish

“kafası bozul-” *(get angry)*

- "Kafam bozuldu." *(I got angry)*. The lemma form :"kafa- boz-".
- "Kafamız bozulur." *(We get angry)* The lemma form: "kafa- boz- ".
- "Kafası bozulmuş." *(He got angry)* The lemma form is "kafa- boz- ".

❖ So, we can not use surface forms of the words of MWE in a lexicon.

❖ We have to use the lemma forms in the lexicons to collect the MWEs.

❖ By this way, we can find the MWEs that matches in the lexicon.

The Problemmatic Case of Lemma Form Usage

- In some cases, the lemma-matching comparison results in accepting word groups that are not really MWEs.

(1) « Ara - dı- ğımı bul- du -m ». (it is not a MWE).

Lt: (search) (p.tense) (what I am) (find) (p.tense) (I)

(I found what I searched for.) The lemma form is "ara- bul-" :

(2) « Ara - ları - nı bul-du -m » (it is MWE)

Lt: (relationship) (their) (Acc.) (find) (p.tense) (I)

(I made them agree and come together.) The lemma form is "ara- bul-" :

- How can we identify this kind of MWEs in agglutinative languages???

Conclusions

- Homonymy is responsible for most ambiguous cases

- We can store the homonymy information of the words in the MWEs in lexicon.

- POS tagging can help while deciding the MWE.

- Lexicon: store the POS for MWEs and their parts too

Idiomatic MWEs in treebanks

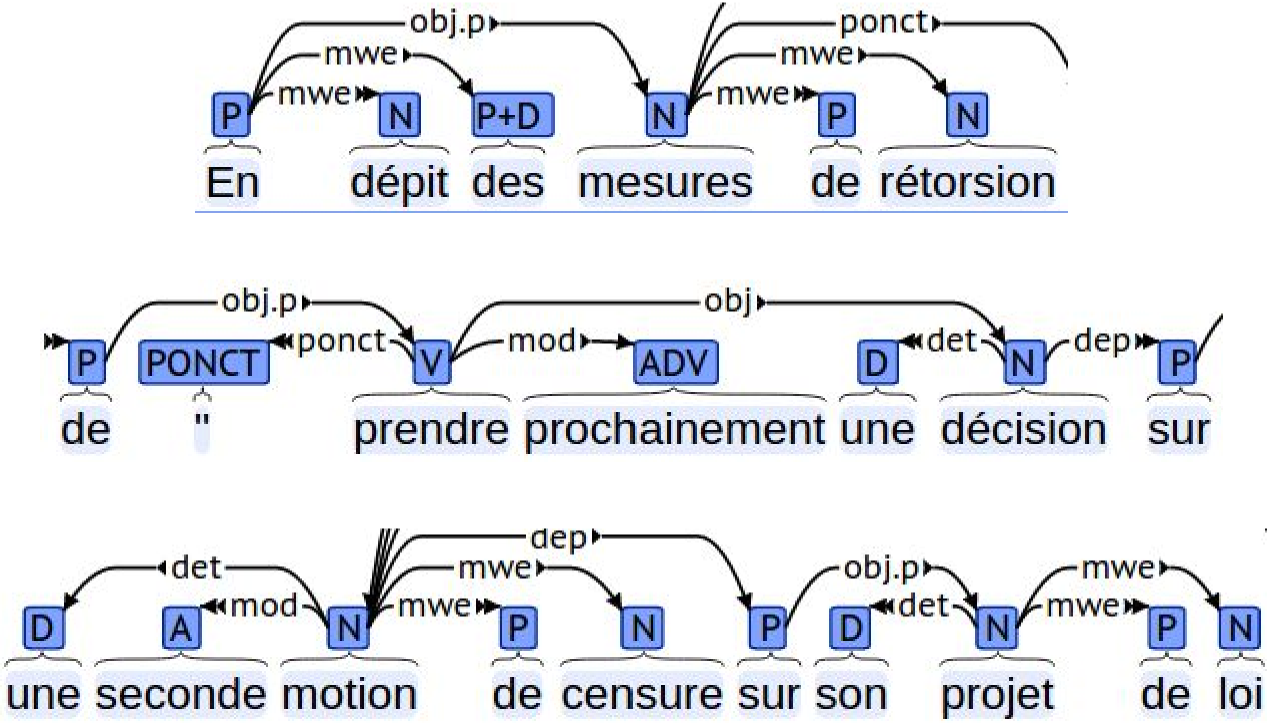
Challenge

Is it possible to represent semantically idiosyncratic MWEs in (syntactic) treebanks? How?

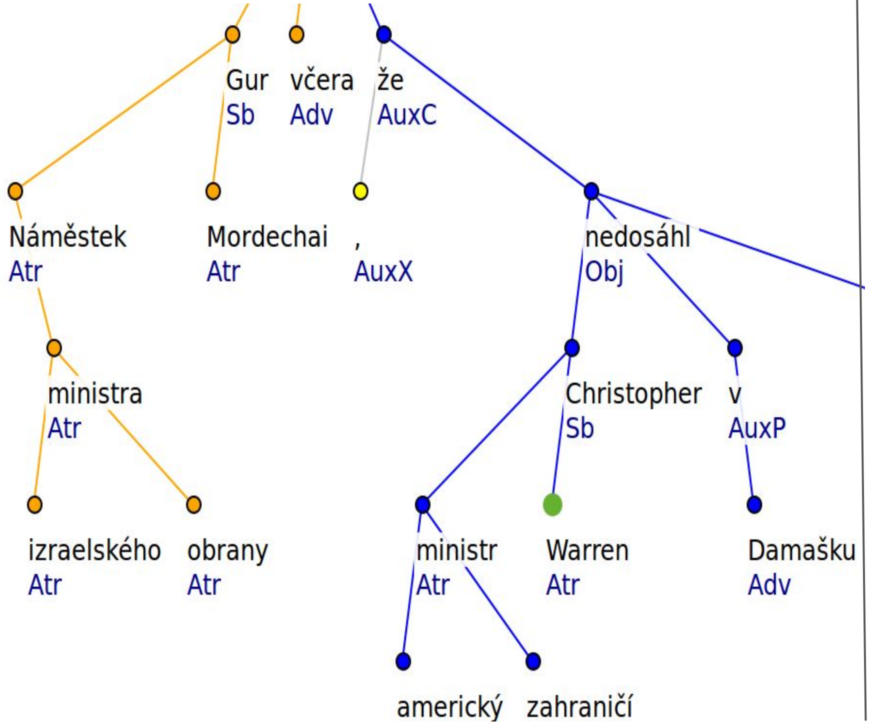
Examples from our languages

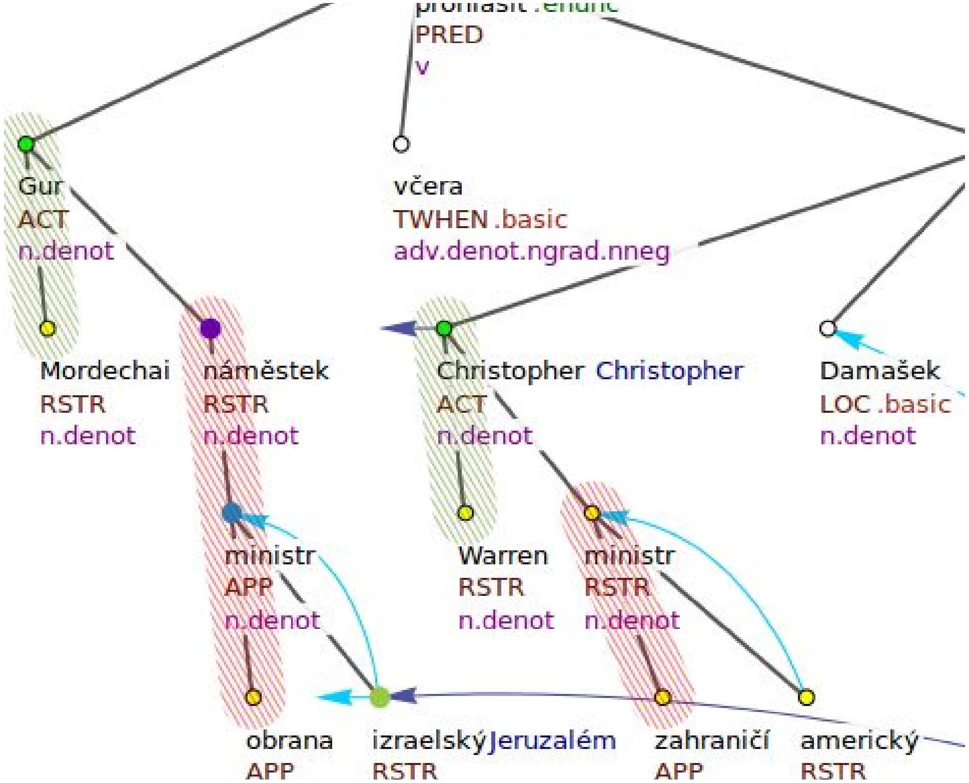
Name	**LVC**	**Fixed preposition**	**Nominal compound**
Carlos *French*	*Prendre une décision* (take a decision)	*En dépit de* (in spite of)	*Projet de loi* (law project)
Hiwa *Farsi*	(dast daadan) دست دادن (shake a hand) "help"	(be jaye inke) به جای اینکه (in spite of)	(tah diig) ته دیگ (the bottom of pot) "leftovers"
Hiwa *Kurdish*	(fishar dan) فشار دان (give pressure)	(la bar way) له به ر وه ی (because of this)	(ru rash)رو ره ش (black face) "sin"
Ivana *Croatian*	*doći do zaključka* (to conclude)	*s obzirom na* (with respect to)	*krevet na kat* (bunk bed)
Kübra *Turkish*	*Hasta etmek* (make sb ill or sick	(no preps in TR)	*Çoban salatası* (sheperd salad) "a kind of salad"
Natalia *Czech*	*projevit zájem* - to express interest	*v zájmu* - in the interest	*ministr ekonomiky* (minister of economics)
Veronika *Hungarian*	*hasznot húz* (advantage-ACC take) “take advantage of”	ADV: *kerek perec* (round pretzel) “plainly, directly”	*fekete doboz* (black box) “black box”
Victoria *Norwegian*	*Ta en avgjørelse* (take a decision) “make a decision”	*I tilfelle* (in case) “in case of”	*Gjøren og laden* (doing and not doing) “behavior”

French Treebank


obj.p
mwe
mwe
ponct
mwe
mwe
P
N
P+D
N
P
N
En dépit des mesures de rétorsion
obj.p
ponct
obj
mod
det
dep
P
PONCT
V
ADV
D
N
P
de " prendre prochainement une décision sur
dep
det
mwe
mod
mwe
obj.p
det
mwe
mwe
D
A
N
P
N
P
D
N
P
N
une seconde motion de censure sur son projet de loi

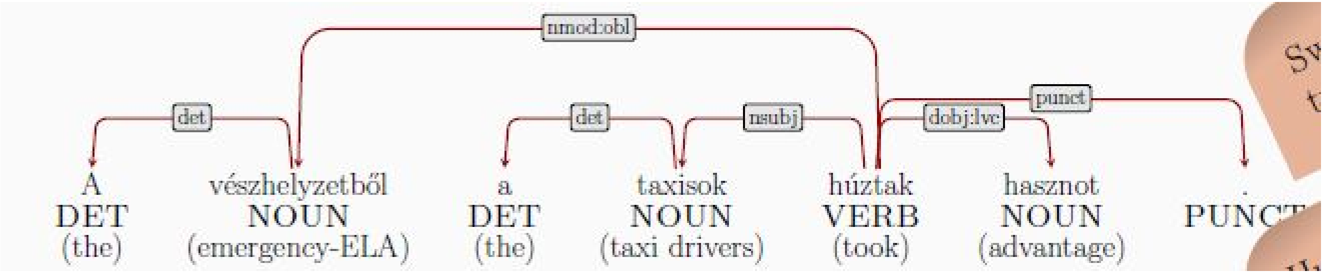
Czech Treebank: syntactic vs. tectogrammatical tree





Hungarian Treebank

LVC:



Other MWEs are not annotated:
kerek perec
round pretzel
“plainly, directly”

ADJ + NOUN combination used as an ADV
no specific label at any level of annotation

Conclusions

Differences across languages:

- some MWEs are marked at the syntactic level (Hungarian LVCs, French contiguous MWEs, Norwegian prepositions)
- some are marked at a different layer
    - Czech, PDT - tectogrammatical layer,
    - Norwegian - F-structure,
    - UD - enhanced layer
- some are not annotated at all
    - Hungarian *kerek perec (round pretzel)*
    - French and Norwegian LVCs

2 levels of annotation can be a solution (syntactic/semantic layer)

Metaphors, collocations, MWEs

Challenge 3

What is the status of metaphors?
As they are not totally compositional (i.e. their meaning cannot be calculated from the original meaning of the words), should they be considered as multiword expressions, e.g. idioms?
Or should they be treated differently from both compositional phrases and MWEs?

My heart was broken.
His temper was boiling.
Waves of spam emails inundated his inbox.

Data collection

- What can be broken? (e.g. verb + object combinations with *break*)
- Searching in corpora in several languages
- Grouping data:

Idiomatic:

Ceviz kırmak - break the walnut -- “go on the loose” (Turkish)

Mil shkaan - breaking the neck, also means shame on you (Kurdish)
Zlom vaz - should you break your neck (Good luck!) (Czech)
Metaphorical:

Break his heart/soul (many languages)

Sense groups across languages

- Physical breaking (leg, arm, table…)
- Emotional/abstract breaking (heart, promise...)
- Idiomatic meaning (language/tongue...)

Conclusions

- No general answer - decision on each case (cognitive category not linguistic one)
- Clearcut examples: “broken heart” - just collocations (some are “more universal” as the very same expressions might occur in many languages (“breaking a language” - CZ, CR, HU), others are more “language-specific”)
- Questionable cases: apply tests (cf. Shared Task guidelines)
- Not to mix up with idioms/LVCs