# Multiword Expressions

## How to recognize and annotate better?

Zuzana Nevěřilová

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

RASLAN 2015, Karlova Studánka, 04 Dec 2015

Partially supported by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

# Multiword Expressions

## Definition

"idiosyncratic interpretations that cross word boundaries (or spaces)"

[Sag et al., 2002]

# Multiword Expressions: Classification

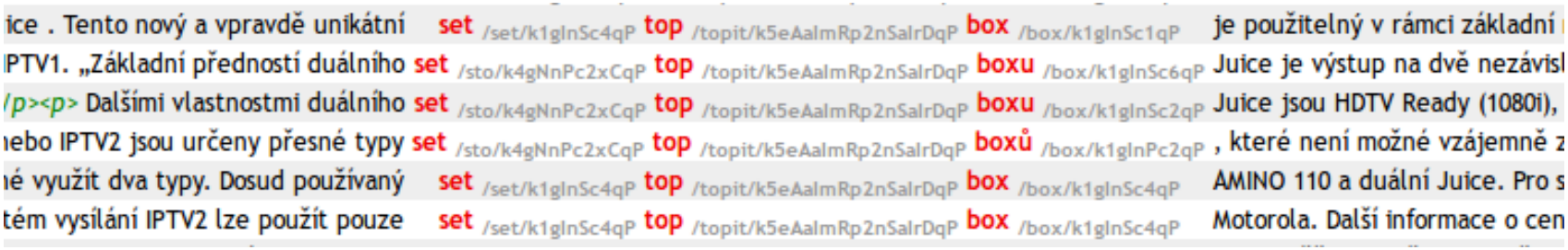
- frozen MWEs (no changes in lemmata nor word order)
  *San Francisco*, *a priori*
- fixed lemmata, possibly non-fixed word order, possible gaps
  *cry over spilt milk*, *natáhnout bačkory*
- open slots
  *to hear X straight from the horse's mouth*, *mít X v paži*

# Annotation of MWEs: Current State

- set top box
- hala bala
- křížem krážem
- a priori
- Karel IV.
- TOP 09
- per se
- hot dog
- po o

# Annotation of MWEs: Current State

set top box



# Annotation of MWEs: Current State

hala bala

| | | | | |
|---|---|---|---|---|
| veliká zkušenost a dávat tam | hala /hala/k1gFnSc1qP | bala /bal/k2eAgFnSc1d1qG | prcnout tam nějaké studenty, to je prostě |
| tím do přebalení věcí jen tak | hala /halo/k1gNnSc2qP | bala /bal/k1gInSc2qG | naházených do kufrů.  Kapitola jed |
| narychlo poutírala co se dalo, | hala /halo/k1gNnSc2qP | bala /bal/k2eAgMnSc4d1qG | Montyho přečísla, nasadila předváděčku |
| h), které se tam dají stavět " | hala /hala/k1gFnSc1qP | bala /bal/k1gInSc2qG | ". A co víme? Čím víc věží, tím menší šance |
| eumím nebo jsem líná, takže | hala /hala/k1gFnSc1qP | bala /bala/k6eAd1qG | menší stehy kolem toho otvoru. Tak teď |
| i, dobré. Už nemáme obrázky | hala /halo/k1gNnPc4qP | bala /bal/k1gInSc2qG | a texty šup sem - šup tam . Už se zde nese |
| e designér, nemůžete jen tak | hala /hala/k1gFnSc1qP | bala /bal/k1gInSc2qG | třeba poskládat různé typy nábytku v nové |
| o všude) a on si na tu čtcrtku | hala /hala/k1gFnSc1qP | bala /bala/k1gFnSc1qG | nalepuje vystřižené obrázky - a tvoříte |

# Annotation of MWEs: Current State

křížem krážem

# Annotation of MWEs: Current State

a priori

| | | | | |
|---|---|---|---|---|
| že je to jinak, oni nejsou | a /a/k8xCqP | priori /priori/k6eAd1 | | špatní, dosáhnou nějakého stupně pozn |
| , otcovství a mateřství se | a /a/k8xCqP | priori /priori/k6eAd1 | | liší: Pak jsou tam ještě ty mužský a žen |
| straně žen-matek a jsou | a /a/k8xCqP | priori /priori/k6eAd1 | | podezíravé vůči otcům. Nejsou podle je |
| k vlastní obrodě a nikoliv | a /a/k8xCqP | priori /priori/k6eAd1 | | beznaději z neodvratitelné hrozby záni |
| zážitky, nemusí se ovšem | a /a/k8xCqP | priori /priori/k6eAd1 | | jednat o zážitky negativní. V jednu chv |
| tento koncesní parametr | a /a/k8xCqP | priori /priori/k6eAd1 | | naplňovat každá nájemní smlouva. </p> |
| rozpor, že by rozum víru | a /a/k8xCqP | priori /priori/k6eAd1 | | negoval. Rozumové verifikace víry se ob |

# Annotation of MWEs: Current State

Karel IV.

| | | |
|---|---|---|
| Slovanech (Emauzy) založený | Karlem /Karel/k1gMnSc7 IV /IV/kA | . r. 1347. Je patrné, že v staroslově |
| něstnanosti, který používal již | Karel /Karel/k1gMnSc1 IV /Iva/k1gFnPc2qP | , ale poněkud nákladný," komentov |
| Přemysl Otakar II. Otec vlasti | Karel /Karel/k1gMnSc1 IV /IV/kA | . jej ve čtrnáctém století zařadil m |
| adaci při něm zřídil roku 1363 | Karel /Karel/k1gMnSc1 IV /IV/kA | . Přilehlá kapitulní budova byla pop |
| a pivo se zde vařilo již za časů | Karla /Karel/k1gMnSc2 IV /IV/kA | . Později byl v komplexu vybudován |
| ><p> A co peněz mohl ušetřit | Karel /Karel/k1gMnSc1 IV /IV/kA | ., kdyby ten Karlštejn a pražský mo |

# Annotation of MWEs: Current State

TOP 09

| | | | |
|---|---|---|---|
| členem a místopředsedou | TOP | /topit/k5eAalmRp2nSalrDqP | 09.  V říjnových volbách bude jed |
| iticky naivní se spoléhat na | TOP | /topit/k5eAalmRp2nSalrDqP | 09, když je vede v Praze bývalý občanský |
| r Schwarzenberg ze strany | TOP | /topit/k5eAalmRp2nSalrDqP | 09, který 19. dubna 2011 vyřešil vládní |
| idkovými bytostmi. Jestliže | Top | /topit/k5eAalmRp2nSalrDqP | 10 byl vpádem antisupermanské ironie do |
| se rozhodl, že vstoupím do | TOP | /topit/k5eAalmRp2nSalrDqP | 09, začal jsem se k tomuto spolku chovat |

# Annotation of MWEs: Current State

per se

| | | | | |
|---|---|---|---|---|
| )5 - Sestavy citlivostí existují | per | /prát/k5eAalmRp2nSalrDqP | se | /se/k3c4xPyFqP , nejsou nijak připraveny na nějakou sumár |
| oli jako vítězství demokracie | per | /prát/k5eAalmRp2nSalrDqP | se | /se/k3c4xPyFqP - především demokracie v její současné |
| y staví řadu překážek, které | per | /pero/k1gNnPc2qP | se | /se/k3c4xPyFqP účinně brání šikanóznímu výkonu práva ako |
| :nosti jsou takové povahy, že | per | /pero/k1gNnPc2qP | se | /se/k3c4xPyFqP znemožňují uplatnění náhrady škody ze str |
| otivním změnám, není odpor | per | /pero/k1gNnPc2qP | se | /se/k3c4xPyFqP ; je to odpor proti ztrátě něčeho.  |
| . Navíc je to totální nesmysl | per | /pero/k1gNnPc2qP | se | /s/k7c7qP .  "takže mluvit v souvislosti se sna |
| nelze bez dalšího definovat | per | /pero/k1gNnPc2qP | se | /se/k3c4xPyFqP jako divadelní.  Z těchto premis se |

# Annotation of MWEs: Current State

hot dog

| | | | | | |
|---|---|---|---|---|---|
| ɔych vyloženě věci typu | hot | /hot/k0 | dog | /doga/k1gFnPc2qP | , hamburger, svíčková se sedmi / |
| za výrazy air condition, | hot | /hot/k0 | dog | /doga/k1gFnPc2qP | , toaster doporučuje vhodnější p |
| konzumní společnosti / | hot | /hot/k0 | dog | /doga/k1gFnPc2qP | <g />. , hranolky, cola, pivo |
| ımburger v housce nebo | hot | /hot/k0 | dog | /doga/k1gFnPc2qP | , tzv. „Completo“. Za tři čtvrtě d |
| e si sami můžete udělat | hot | /hot/k0 | dog | /doga/k1gFnPc2qP | , párek v rohlíku....nebo vám uděl |
| ıli k obědu jen hranolky, | hot | /hot/k0 | dog | /doga/k1gFnPc2qP | , chvíli šišky z ml.masa. Určitě to |

# Annotation of MWEs: Current State

po o

| | | | | |
|---|---|---|---|---|
| mladej zvracel, tak jsem pro něj | po | /po/k7c6qP | o /o/k7c4qP | jela a páč jsem nestíhala, tak ho nedali |
| din denně - některé jdou domů „ | po | /po/k7c6qP | o /o/k7c4qP | “, jiné čekají na odchod do pozdního odpol |
| ıstu, tak to prý neeee , ale aspoň | po | /po/k7c6qP | o /o/k7c4qP | musíme čistit(po úrazu má přední zoubky |
| :o chtěl domu vždycky nejpozději | po | /po/k7c6qP | o /o/k7c4qP | Ke kočárku, právě deltim jsem měla na jiři |
| cinka zatím chodí na půl dne, tak | po | /po/k7c6qP | O /O/kA | jde domů, tento měsíc mám ještě rodičák |
| -) Fíla byl na WC naposledy včera | po | /po/k7c6qP | O /O/kA | , od té doby nic, ani teď po snídani, tak |

# Method for Discovering Fixed MWEs

observation: many people are unsure how to write frozen MWEs

*a priori*, *apriori*, *a-priori*
*hot dog*, *hotdog*, *hot-dog*
...

all variants found in web corpora

# Result

26,700 MWE candidates, 5,500 contain a one-letter word

# Result

26,700 MWE candidates, 5,500 contain a one-letter word

not all of them really are MWEs with problematic annotation:

*český jazyk*, *ne ne*, *A B C D*

# Result

26,700 MWE candidates, 5,500 contain a one-letter word

not all of them really are MWEs with problematic annotation:

*český jazyk*, *ne ne*, *A B C D*

not all of them are MWEs in every context

*pro aktivní*, *po o*, *to do*, *dál nic*

# Conclusion and Future Work

- the method is simple
- the method needs messy (web corpus) data
- the method seems to be relatively language independent
- how to discover incorrect annotation
- how to discover the right concepts
- what are the correct tags


Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002).
Multiword expressions: A pain in the neck for nlp.
In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.