

Úvod do korpusové lingvistiky 4



**AUTOMATICKÁ MORFOLOGICKÁ ANALÝZA
(TOKENIZACE, ANALÝZA A DESAMBIGUACE,
TAGGING A LEMMATIZACE)**

Značkování pomocí nástrojů automatické morfológické analýzy



- Morfológický slovník
- Jednotky + interpretace
- Word – lemma – tag
- Identifikace
- Obecně nejednoznačné přiřazení
- disambiguace

Rozdělení textu na tokeny



- Token – jednotka pro další analýzu
- Jednotky typu word
- Textová slova
- Interpunkce
- problémy

Lemmatizace a značkování proloženě psaného textu



• pool

Výskytů: 3 i.p.m.: 0.0 (vztaženo k celému korpusu) | ARF: 1

| | | | | | | |
|--------------------------------|---------|--------------|-----------------|-----------------|----------------|-----------------------------------|
| Mladá fronta DNES, 5. 4. 1993 | Liver p | /p/NN-----8- | o /o/RR--6----- | o /o/RR--6----- | l /l/NN-----8- | - Překážkový dostih Velká národní |
| Mladá fronta DNES, 8. 10. 1993 | Black p | /p/NN-----8- | o /o/RR--6----- | o /o/RR--6----- | l /l/NN-----8- | - V Británii bylo loni |
| Mladá fronta DNES, 11. 4. 1994 | Liver p | /p/NN-----8- | o /o/RR--6----- | o /o/RR--6----- | l /l/NN-----8- | (Od našeho zvláštního zpravodaje |

Tvary s volným morfémem -s



- Lemma a slovní druh (pos)

Výskytů: 9 i.p.m.: 0,01 (vztaženo k celému korpusu) | ARF: 1

| | |
|---------------------------------------------------------------|------------------------------------------------------------------------|
| Abyste neumřel smíchy: tajemství Donalda Lama a Berty Coolové | řekl jsem . " O čems /co/PQ--6--2----- to tedy k sakru mluvil |
| Zavírací hodina | " " Všechno , o čems /co/PQ--6--2----- mluvil , způsob , jak |
| Kukaččí vejce | " Takže objevuješ , o čems /co/PQ--6--2----- neměl ani tušení , Thorne |
| Idiot - 4. část | " Nevytáčeš se ! O čems /co/PQ--6--2----- jí psal ? " " |
| Kdo zabil Tomáše Wellse | náš souhlas . " O čems /co/PQ--6--2----- mluvil s hrobníkem ? " |
| Básně, eseje, překlady | každý nápor časem se v čems /co/PQ--6--2----- rozpustí . Na mladých je |
| Zapadlo slunce za dnem, který nebyl | ruší klid A srdce po čems /co/PQ--6--2----- ' minulém práhne , po |
| Zapadlo slunce za dnem, který nebyl | ' minulém práhne , po čems /co/PQ--6--2----- ' , co nelze vyslovit |
| Zapadlo slunce za dnem, který nebyl | oko modré , stále o čems /co/PQ--6--2----- snil . Hráč druhý , |

Tvary se spojovníkem



- První část kompozit

Výskytů: 6 i.p.m.: 0,05 (vztaženo k celému korpusu) | ARF: 4 | Výsledek je promíchán

| | | | | |
|-----------------------------|-----------------------------------------------|----------|-------------------------|------------------------------|
| D Test, č. 5/2005 | připomíná obdobnou vzájemnou impertinenci v | anglicko | /anglicko/A2-----A----- | - skotské rozmluvě : " |
| Shakespeare | a bídná usedlost ! To | Anglicko | /anglicko/A2-----A----- | , jež zvyklé bývalo být |
| Uvnitř velryby | co se namane , když | Anglicko | /anglicko/A2-----A----- | si věrno zůstane . Je |
| Lidé a země, č. 12/2004 | dictionary FRAUS ILUSTROVANÝ STUDIJNÍ SLOVNÍK | ANGLICKO | /ANGLICKO/X@----- | - ČESKÝ / ČESKO - |
| Nenasytná | jejich pevné ruce vládl na | anglicko | /anglicko/A2-----A----- | - skotské hranici klid a |
| Deníky Bohemia, 12. 4. 2007 | půl šesté přijde na řadu | anglicko | /anglicko/A2-----A----- | - bengálský fil Antarmahal , |

Automatická morfoloická analýza



- Obecně víceznačná
- Tvarová homonymie
- Homonymie na úrovni slovního druhu
- Slovnědruhové přechody a přesahy

Tvarová homonymie

Word: který



který:

P₄MS₁-----

P₄IS₁-----

P₄IS₄-----

P₄MS₅-----

P₄IS₅-----

Tvarová homonymie

Word: který



který:

k2gMnSc1

k2gInSc1

k2gInSc4

k2gMnSc5

k2gInSc5

k2gMnPc1wH k2gMnPc4wH k2gMnPc5wH

k2gInPc1wH k2gInPc4wH k2gInPc5wH

k2gFnSc2wH k2gFnSc3wH k2gFnSc6wH

k2gFnPc1wH k2gFnPc4wH k2gFnPc5wH

k2gNnSc1wH k2gNnSc4wH k2gNnSc5wH

k2gNnPc1wH k2gNnPc4wH k2gNnPc5wH

Homonymie tvarů od různých lemmat



- ženu /žena
- ženu/hnát
- 1.ř. ... je **<ženu/hnát/VB-S---1P-AA.*>** i muže vyšetřit
- 3.ř. ... cílem je **<ženu/hnát/VB-S---1P-AA.*>** ženu zaujmout

Mladá fronta DNES, 29. 6. 2005 obavách z poruchy plodnosti je **ženu** /hnát/VB-S---1P-AA---| i muže vyšetřit a dodat

Mladá fronta DNES, 15. 8. 2005 . " Přes den je **ženu** /hnát/VB-S---1P-AA---| , ale v noci to

Reflex, č. 31/2009 minut a jejím cílem je **ženu** /hnát/VB-S---1P-AA---| zaujmout . Používá se k

Korespondence že nezahlám a že to **ženu** /hnát/VB-S---1P-AA---| svínským krokem , incident neincident

Třísky do masa a nelitostný . Dopředu se **ženu** /hnát/VB-S---1P-AA---| . Víím , cítím ,

Homonymie tvarů od různých lemmat pila/pila x pila/pít



- 5. ř. <Pila/pít/VpFS---3R-AA.*> je přímo propojena s počítačem

| | |
|---------------------------------------------|----------------------------------------------------------------------------------------------|
| Zahradní ploty a zídky | pily v přímém směru . Pila /pila/NNFS1-----A----- může být i na baterii |
| Pevné pouto | s kudrnatými vlasy , která pila /pít/VpFS---3R-AA---I čaj z termosky a mluvila |
| Sobota | motoroky a ječí jako motorová pila /pila/NNFS1-----A----- . V tuto dobu se |
| Maminka, č. 1/2005 | , návštěvy u psychotronika , pila /pít/VpFS---3R-AA---I různé bylinné čaje . S |
| Mladá fronta DNES, 30. 3. 2006 | firmy - například JAWA , pila /pila/NNFS1-----A----- , dodavatel uhlí . Středočeský |
| Stavitel, č. 4/2007 | pila a zkracovací pila . Pila /pít/VpFS---3R-AA---I je přímo propojena s počítačem |
| Osudný zvrát | mohla ho pít stejně jako pila /pít/VpFS---3R-AA---I mléko . Všechno se změnilo |
| Tichá žena | ve kterém jsem jako dívka pila /pít/VpFS---2R-AA---I kakao usazená na chaise-longue postavím |
| Cizinec přichází | chléb života a jeho ústa pila /pít/VpNP---3R-AA---I víno nesmrtelnosti , a to |
| Stálo to za hovno..., ale aspoň byla sranda | whisky , natož aby ji pila /pít/VpNP---3R-AA---I . Lkaní , mládím tak |
| Ztracená | toho vypila ? Proč jsi pila /pít/VpFS---2R-AA---I ? Ale nahlas položila jedinou |
| Poslední útočiště | . " Co myslíte , pila /pít/VpFS---3R-AA---I hodně ? " " Ne |
| Otrok | začala zuby do nader a pila /pít/VpFS---3R-AA---I až do svítání . Ráno |
| Naplnění | jako prs , z nějž pila /pít/VpFS---3R-AA---I . Vypadalo to , že |
| Tabatěrka z Bagomba | vy , Harry : ' Pila /pít/VpNP---3R-AA---I ! Siba tu beng-beng . |
| Rozený svůdce | vysála , když jste mu pila /pít/VpFS---2R-AA---I krev . " Nita vyšpulila |
| K znamená Kennedy | patrně mezitím svlékla . Jak pila /pít/VpNP---3R-AA---I , nechal spadnout na zem |
| Právo, 5. 2. 2009 | pod Hostýnem na Kroměřížsku . Pila /pít/VpFS---3R-AA---I Javořice , a. s. má |
| MIčení jehňátek | čaje ? " Při studiu pila /pít/VpFS---2R-AA---I Mappová nápoj , který si |
| Geraldova hra | mezi návštěvníkovými nohama byla motorová pila /pila/NNFS1-----A----- . Jessie si tím byla |

Slovnědruhové přechody a přesahy



| | | |
|-----------------------------------------------------------|-----------------------------------------------------------------------------|----------------------------------------|
| Rose Madder | dělá , měla živůtek obtočený kolem /kolem/RR--2----- | krku a jazyk jí visel |
| Utrpení oddaného Všiváka | smutný . Očima ztěžka bloudí kolem /kolem/Db----- | a chvěje se třasem , |
| Lidé a země, č. 4/2002 | první linii a svou divokostí kolem /kolem/RR--2----- | sebe šířily hrůzu a děs |
| Kdo je Mallory? | Ann . 2 . Teprve kolem /kolem/RR--2----- | šesté večer si Corridon všiml |
| Nezbytné věci | " Opustil záchodek - prošel kolem /kolem/RR--2----- | Keetona , aniž by o |
| Osobní korespondence | s tím kamarádem na kolech kolem /kolem/RR--2----- | našich a když jeli nazpátek |
| Dům | truhly a konečně je rozhodila kolem /kolem/Db----- | . Byl to krásný přiběh |
| Lidé a země, č. 4/2009 | , Kominici , Kobyly drží kolem /kolem/RR--2----- | ramena a zpívají My jsme |
| DVDMAG, č. 2/2007 | . Napínavý děj se točí kolem /kolem/RR--2----- | dvou mužů : Prvního důstojníka |
| Svatba | manželství . Tanec se závojem Kolem /kolem/RR--2----- | půlnoci si nevěsta sundá závoj |
| Světlo dne | svištění dopravy cákající vodou . Kolem /kolem/Db----- | projížděly patrové autobusy s hořejším |
| D Test, č. 10/2005 | 4,0 mmol/l u mužů a kolem /kolem/RR--2----- | 4,5 mmol/l u žen) |
| Němci v Praze 1861-1914 | z pražských německých židovských spisovatelů kolem /kolem/RR--2----- | první světové války , že |
| Staří Egyptané | k vytvoření jednotného egyptského státu kolem /kolem/RR--2----- | roku 3000 př. n. l. |
| Stanice Bazilišek | se usadila , začaly se kolem /kolem/RR--2----- | ní vysouvat displeje a monitory |
| Věřil jsem, že mám řůu času, ale teď si nejsem tak jistej | 1917 ; dvojnásobný vítěz závodu kolem /kolem/RR--2----- | bloku v roce 1918 ; |
| Smrtící zrada | než šest měsíců . Měla kolem /kolem/RR--2----- | sebe okruh blízkých přátel , |
| Paměť esejisty | jsem vůbec zkusil psát . Kolem /kolem/RR--2----- | osmnácti let jsem se v |
| Devět miliard božích jmen | postrčí na stabilní oběžnou dráhu kolem /kolem/RR--2----- | Měsíce . Zapálíš je , |
| V nemilosti | řítit se s polodivokými koňmi kolem /kolem/RR--2----- | padesáti kilometrů za hodinu přes |

Praktická ukázka:
Sním je místo něho.



| | | |
|--------------|--------------|---------|
| Sním | sníst | VpS--1d |
| | snít | VpS--1n |
| je | být | VpS--3n |
| | ono | PPSN4-- |
| | oni | PPPM4-- |
| | ony | PPPI4-- |
| | ony | PPPF4-- |
| | ona | PPPN4-- |
| | místo | místo |
| místo | | R---2-- |
| něho | něha | NNSF5-- |
| | on | PPSM2-- |
| | on | PPSM4-- |
| | on | PPSI2-- |
| | ono | PPSN2-- |
| . | . | Z----- |

Lemma a tag



- Výsledkem automatické morfologické analýzy a desambiguace
- Závisí na rozsahu a obsahu slovníku, nad nímž pracuje AMA
- Závisí na použité desambiguace

Rozsah a obsah slovníku



- Pouze interpretace uložené ve slovníku
- Pouze jednotky uložené ve slovníku
- Tvary nerozpoznané AMA
- Tag=X.*

Desambiguace



- Stochastické metody
- Pravidlové metody
- Hybridní metody
- Guessery

Desambiguace



- Zjednoznačnění = volba konkrétní (kontextově správné) interpretace z nabízených možností
- Problémy: nepřítomnost správné interpretace, nemožnost jednoznačně určit, která interpretace je správná

Stochastické metody



- Ruční analýza dat – trénovací data
- Metody strojového učení
- Automatické nástroje založené na matematické pravděpodobnosti

Pravidlové metody



- Implementace pravidel, která v jazyce platí (kognitivně plausibilní přístup).
- Pozitivně formulovaná pravidla
- Negativně formulovaná pravidla

Hybridní metody



- Kombinují různé přístupy statistické a pravidlové

Guesser



- Hadač – program, který pracuje na různých základech (statistika/pravidla) a bez znalostní databáze (slovníku) se snaží „uhádnout“ příslušnou interpretaci.
- Guessery byly testovány na korpusu SYN2005.