Vybrané aktuální projekty Centra ZPJ

Vašek Němčík, Vojtěch Kovář

E-mail: xnemcik@fi.muni.cz,xkovar3@fi.muni.cz
http://nlp.fi.muni.cz/poc_lingv/

Obsah:

- Saara – systém na určování anafor
- SET – syntaktická analýza pomocí postupné segmentace věty

Gracie: Minulý týden byl můj bratr vyšetřovat vraždu a představ si, našel toho chlapa za hodinu.
George: Našel za hodinu vraha?
Gracie: Ne, *toho chlapa, co ho zabili.*
George: Tvůj bratr je nejen vysoký, ale i rychlý.
Gracie: No, před časem měli pan a paní Jonesovi manželskou krizi a můj bratr byl najat, aby sledoval paní Jonesovou.
George: No, je to nepochybně moc atraktivní žena.
Gracie: To je – a můj bratr ji sledoval ve dne v noci, půl roku.
George: A jak to skončilo?
Gracie: Požádala o rozvod.
George: Paní Jonesová?
Gracie: Ne, *žena mého bratra.*
(George Burns a Gracie Allen: "The Salesgirl")

Úvod

- text/**diskurs** – jednotka jazykové komunikace větší než:
- věta/**výpověď** – minimální obsahově úplná jednotka

věta	**výpověď**
langue	parole (de Saussure)
competence	performance (Chomsky)
produkt	*proces*
struktura	*chování*
nedůležité kdy/kde/jak	*podmínky/okolnosti/způsob*

- referenční výrazy
- reference (odkazování)
  jazykový výraz $\longmapsto$ mimojazyková entita

Reference

- **exofora** (vnější reference)
  výraz odkazuje k entitě ve světě přímo
  "Slunce", "Alpy", "Václav Havel", "ty schody před FI"
  - deixe – odkazování k entitám v rámci komunikační situace (gesta, "tady", "teď", "tamto", ...)
- **endofora** (vnitřní reference)
  entita je určena na základě vztahu k jinému výrazu v diskursu (nejen mimojazykový, ale i jazykový kontext ...)
  - anafora – výraz se vztahuje k výrazu dříve v textu
  - katafora – výraz se vztahuje k výrazu dále v textu
    méně častá; vysktuje se v beletrii (zvyšuje napětí):
    "Ranní světlo ho probudilo už v pět. Rychle se oblékl a nasnídal. Detektiv Jones věděl, že nemůže ztrácet čas."

Anafora

- **anafora** (anaphor) – anaforický výraz (× Chomsky)
  - zejména zájmena, ale i “ten muž”, ...
- **antecedent** – předcházející výraz, ke kterému se anafora vztahuje
- **anafora** (anaphora) – anaforická reference (jev)
- **anaphora resolution** – určování anaforických vztahů (hledání vztahů mezi anaforami a antecedenty)

Příklady:

- **[Petr]**$_i$ snědl **[koláč]**$_j$.
  [(on)]$_i$ Byl hladový a **[ten koláč]**$_j$ vypadal lahodně.
- **[Venus]**$_i$ rose at 0930, but I didn't see **[the thing]**$_i$.
- **[Jones]**$_i$ offered **[[his]**$_i$ **furniture]**$_j$ for sale, but nobody wanted **[the stuff]**$_j$.

Lze udělat úkrok stranou?

- Můžeme se tomu všemu vyhnout, třeba používáním jen přímé reference?
- Nemuseli bychom se zabývat kontextem ...

NE. Z mnoha vážných důvodů:

- Lidé jsou líní.
  - anafory jsou krátké a snadno se používají
  - patrně vlastní lidské komunikaci (ve všech jazycích!)
- diskurs není libovolná sekvence výpovědí
  - koherence – sémantická návaznost
  - kohese – gramatické a lexikální vztahy

  $\rightsquigarrow$ anaforické vztahy drží text pohromadě (umožňují nám se držet zamýšleného toku myšlenek)

Ilustrační příklad

[Jarda]$_i$ si koupil Porsche. **(On)**$_i$ Rád jezdí rychle.
[Jarda]$_i$ si koupil Porsche. **[Jarda]**$_{*i,j}$ rád jezdí rychle.

$\rightsquigarrow$ delší/složitější věta zní divně (nutí k zamyšlení)

- **Kooperační princip** (Grice)
  Komunikační maximy:
  - kvality
  - relevance
  - kvantity
  - způsobu
- Posluchač předpokládá, že se jimi mluvčí řídí.
- Když ne, má to hlubší důvody.
- více o pragmatice v “IA091 Sémantika a komunikace”

Proč to učit počítače?

- zásadní úzké hrdlo mnoha NLP aplikací
- **Information Extraction**
  - **[Václav Havel]** is a Czech writer and dramatist.
    [He] was the ninth and last President of Czechoslovakia and the first President of the Czech Republic. *(Wikipedia)*
  - “the best doctor in Europe” $\longrightarrow$ Google

    Letters from Asia addressed loosely to The Best Doctor in Europe arrived on **[his]** doorstep.

    [His] own reputation as the best doctor in Europe couldn't save **[him]** from the tragedies of **[his]** life.
- Bez AR nenajdeme to, co hledáme.
  Pouze anaforické výrazy (které jsou samy o sobě prázdné).

Proč to učit počítače?

- **Strojový překlad**
- CZ $\mapsto$ EN
  [Sestřička] mu dala **[pilulku]**. Spolkl **[ji]** a do minuty usnul.
  [The nurse] gave him a pill. He swallowed **[her]** and fell asleep in a minute.
- DE $\mapsto$ EN
  Ich suche **[meine Uhr]**. Ich kann **[sie]** nirgendwo finden.
  I am looking for **[my watch]**. I can't find **[her]** anywhere.
- nelze překládat přímo (různé gramatické kategorie)
- navíc: různé vlastnosti anafor

Definice úlohy

- nalézt anaforické výrazy v textu
- určit k nim antecedenty
- určit typ vztahu
  - koreference
    (dva výrazy se odkazují ke stejnému promluvoému objektu)
  - bridging (asociativní/nepřímá anafora)
    (jakákoliv sémantická relace)
    - hyperonymie/hyponymie
      "Nábytek je drahý. Židle jsou nejdražší."
    - část/celek
      "Každý majitel bytu se snaží zabezpečit vchodové dveře."
    - entita/vlastnost
      "Pepa má nové auto. Barvu určitě vybírala jeho žena."
    - příčina/následek
      "Včera tu byl požár. Kouř je tu stále cítit."

Typy anafor

- **textová** vs. **gramatická**
  [Ben] takes a photo of **[himself]** every day.
- **pronominální** (pro NLP asi nejrelevantnější)
- **nominální**
  Od září bude do **[Brna]** létat nová letecká linka. Očekává se, že přinese **[druhému největšímu městu ČR]** nové turisty.
- **slovesná**
  John likes cats. So does Bill.
- **one-anaphora**
  John has a black Porsche. I would like one too.
- **nulová (zero) anafora**
  anafora není povrchově realisována
  v češtině (a ostatních pro-drop jazycích) nevyjádřené podměty

Typy pronominálních anafor

- osobní zájmena
  - silná: "jemu", "on", "ona"
  - slabá: "mu", "ho" (klitika)
  - nulová: $\emptyset$
- demonstrativní zájmena: "ten", "ta", "tomu"
- reflexivní zájmena: "se", "sebe", "svůj"
- posesivní zájmena: "jeho", "jejího"
- relativní zájmena: "který", "jenž"

ALE jsou i neanaforická zájmena:

- deixe: "to"
- expletivní/pleonastická zájmena:
  It's raining. / Es regnet.
  It is the first chapter, I enjoy the most.
  Zdá se, že tu někdo byl.

Znalosti potřebné pro AR

- **morfologie**
  - shoda v Φ-atributech (závislé na jazyce)
  - čeština: osoba, číslo, rod
  - angličina: pouze sémantický rod
    $\Rightarrow$ nutnost mít informaci jméno $\mapsto$ rod
- **syntax**
  - posice anafory/antecedentu v syntaktické struktuře věty
  - paralelismus
    tendence k zachování stejných syntaktických rolí:
    [Mary] met **[Lucy]** at the bus station.
    [She] asked **[her]** about the new neighbour.
- **pragmatika**
  - Griceův kooperační princip ...
  - komunikační situace + kontext
  - scénáře

Sémantika a znalosti o světě

- hraje při interpretaci anafor často rozhodující roli
- sémantická plausibilita zvyšuje/snižuje pravděpodobnost některé interpretace, některé lze zcela vyloučit

  After the [bartender] served [the patron], [he] got a big tip. After the [bartender] served [the patron], [he] left a big tip.
- iniciální interpretace (hned)
- pokud pozdější informace vedou ke sporu: reinterpretace (backtracking)
- **garden-path effect**
- význam slov
- znalosti o světě
- inference

Sémantika a znalosti o světě

- If the baby does not thrive on raw milk, boil it.
- The FBI's role is to ensure our country's freedom and be ever watchful of those who threaten it.
- Stehlíková ustoupila od sbírky. Romové o ni nestojí.
- Klaus dostal dopis podepsaný Aničkou. Má ho policie.
- A: I ve Veselé vačici by mohla být volná místa.
  B: Jé, tam jsem ještě nebyla. Slyšela jsem, že tam chodí studenti. A že prý dobře vaří.
- 'I said disarm only!' Lockhart shouted in alarm over the heads of the battling crowd, as Malfoy sank to his knees; Harry had hit him with a Tickling Charm, and he could barely move for laughing.
  (J. Rowling: Harry Potter and the Chamber of Secrets)

Sémantika a znalosti o světě

- Genau so sei es ihm vorgekommen, sagte Gauss, schlief ein und wachte bis zum abendlichen Pferdewechsel an der Grenzstation nicht mehr auf. Während die alten Pferde ab- und neue angeschirrt wurden, assen sie Kartoffelsuppe in einer Gastwirtschaft.
  (Daniel Kehlmann: "Die Vermessung der Welt: Die Reise")

- všechny tyto znalosti je obtížné shromáždit
- i kdyby byly k disposici, bylo by obtížné v nich hledat

- AR je považováno za **"AI-úplný problém"**
  AR je stejně obtížný problém jako naučit počítače myslet.

  $\Rightarrow$ nutno si úkol zúžit

Teoretické problémy

- John loves his wife. So does Bill.
- The man who gave his **[paycheque]** to his wife was wiser than the man who gave **[it]** to his mistress.
- If any man owns **[a donkey]**, he beats **[it]**.
- **[No one]** will be admitted to the examination, unless **[he]** has registered four weeks in advance.
- [The man who shows he deserves **[it]**] will get [the prize **[he]** desires].

AR algoritmy

- heuristická pravidla (70. léta)
  - SHRDLU – "block world" Terryho Windograda
  - Hobbsovo syntaktické hledání
  - jednoduchá pravidla, vzory, časté instance
- sématické teorie
  - centering, focusing – modelování lokální koherence
  - BFP algoritmus
  - výpočetně problematické
- knowledge-poor (90. léta)
  - kacířství motivované praktickými potřebami
  - založené na datech, která lze dostatečně úspěšně spočítat (morfologie, povrchová syntax, jednoduché sémantické třídy)
  - RAP – váhování
  - CoGNIAC (pouze 6 pravidel – vysoká přesnost, malé pokrytí)
  - MARS – váhování

AR a strojové učení

- statistika a strojové učení dnes v NLP převažují
- AR není klasifikační problém

předefinování umožňující použití std. ML metod:

- **1 instance**: dvojice anafora-antecedent
- **atributy**: knowledge-poor informace
- **cílový atribut**: 1 pro koreferentní dvojici, jinak 0
- velký nepoměr negativních a positivních instancí
- nutno část negativních instancí odstranit z trénovacích dat

AR a čeština

- mnoho teoretických prací (FGP: Sgall, Hajičová)
- PDT 2.0 – velký ručně anotovaný korpus, 3 roviny
- anotace pronominání koreference
- implementace:
  Zdeněk Žabokrtský, Nguy Giang Linh
- pouze v rámci formalismu PDT 2.0

- **Saara**
- lze aplikovat na volný text
- různé algoritmy, zdroje dat, pre-processing
- možnost férového porovnání algoritmů

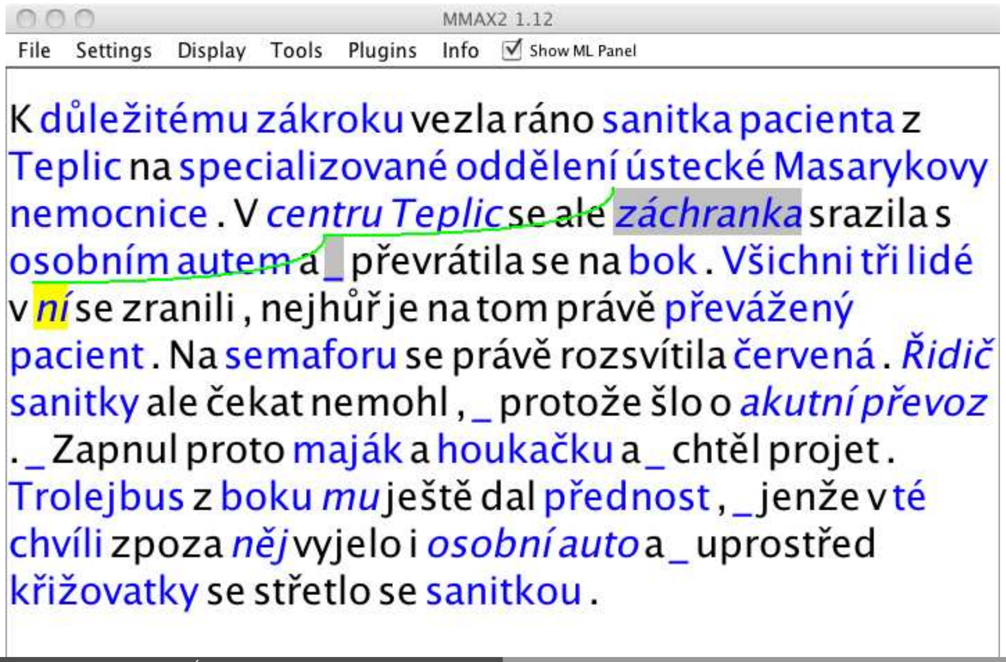
Saara

- roviny abstrakce:
  - technická rovina
    různé formalismy/formáty dat $\mapsto$ vertikál
  - "markable" rovina
    "markables" + jejich vlastnosti a vztahy nad ní se definují AR algoritmy
  - "supervisor"
    definuje, který pre-processing a algoritmus se použije
- "markable"
  - jakákoliv jednotka složená z jednodušších jednotek
  - možno definovat různé roviny
    referenční výrazy – klause – věty
  - atributy
  - vztahy mezi markables $\rightsquigarrow$ koreferenční třídy
  - MMAX 2

Saara

- import dokumentu (vertikál)
- pre-processing
  - rozdělení vět do klausí
  - detekce nevyjádřených subjektů
  - model diskursu – detekce markables
- AR $\rightsquigarrow$ koreferenční třídy
- výstup
  - vertikál
  - MMAX2 XML pro visualisaci

Saara



Hobbs syntactic search

- jako syntaktickou strukturu předpokládá frázové stromy
- X-bar theory (Chomsky, Jackendoff)
  X – complement – X' – adjunct – X' – specifier – XP
- algoritmus je definován jako procházení stromu
- začíná se v listu dané anafory
- podle kategorie aktuálního uzlu se volí další cesta
- prominentnější posice jsou procházeny dříve
- lze adaptovat na jiné formalismy
- jednoduché, ale nefunguje špatně

BFP algoritmus

- každá výpověď:
  - forward-looking centers (setříděné)
  - preferred center (ten nejvýše postavený)
  - backward-looking center
- formulována 2 jednoduchá pravidla, neformálně:
- preferováno je odkazování zájmeny
- preferováno je zachovávání backward-looking center
- počítají se různé kombinace a filtrují se ty, které nevyhovují pravidlům
- kombinace, která představuje nejplynulejší přechod center

RAP

- identifikace NP, filtrování nereferenčních, reflexiva atd.
- přidělí se iniciální váhy kandidátům (součet)
- při hledání antecedentu ke konkrétní anafoře se pro danou kombinaci váhy dále upravují (katafora, paralelismus, ...)
- antecedentem je kandidát s nejvyšší vahou
- při zpracovávání nové věty se všechny váhy podělí dvěma

Factor type	*Initial weight*
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object and oblique complement emphasis	40
Head noun emphasis	80
Non-adverbial emphasis	50

Pražské algoritmy

- hned několik algoritmů
- formulovány "na papíře"
- vyhodnocovány ručně
- jako RAP také váhovací princip
- modeluje aktivaci objektu v mysli posluchače
- zohledňuje se informace o AČV
- teoreticky logické, ale prakticky nepotvrzené

Gracie: Last week my brother went out on a murder case, and you know, he found that man in an hour.
George: He found the murderer in an hour?
Gracie: No, *the man who was killed*.
George: Not only is your brother tall, but he's fast.
Gracie: And then Mr. & Mrs. Jones were having matrimonial trouble, and my brother was hired to watch Mrs. Jones.
George: Well, I imagine she was a very attractive woman.
Gracie: She was, and my brother watched her day and night for six months.
George: Well, what happened?
Gracie: She finally got a divorce.
George: Mrs. Jones?
Gracie: No, *my brother's wife*.
(George Burns and Gracie Allen in "The Salesgirl")

Syntaktická analýza přirozeného jazyka

Syntaktická analýza:

- odhalení povrchové struktury věty
- základ pro analýzu jazyka na vyšších úrovních

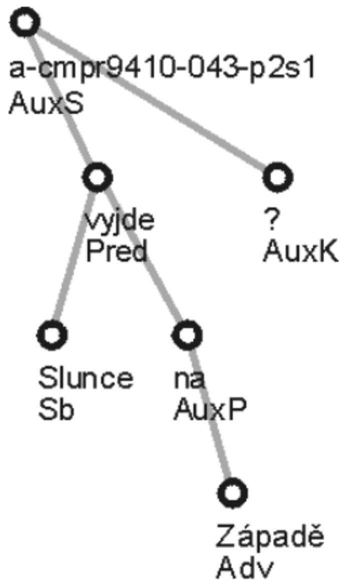
Závislostní formalismus:

- strukturní vztahy kódovány závislostmi mezi slovy na vstupu
- pražský korpus závislostních stromů PDT

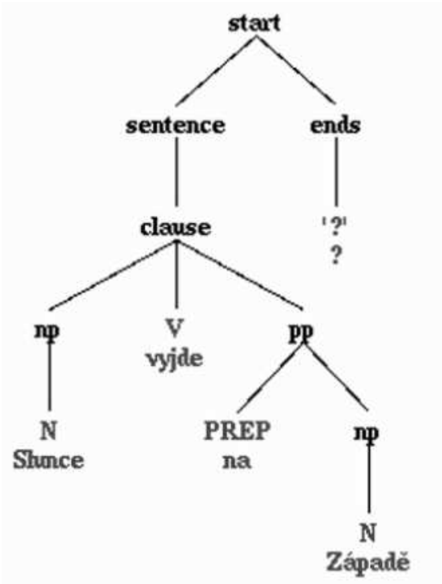
Složkový formalismus:

- strukturální vztahy popisovány stromem odvození z gramatiky
- brněnský analyzátor `synt`

Závislostní strom – příklad



Složkový strom – příklad



Syntaktická analýza přirozeného jazyka

Parciální syntaktická analýza:

- nezajímá nás kompletní strom, jen některé vztahy
- např. systém `VaDis`, **Word Sketches**

Použití syntaktické analýzy:

- jakékoli pokročilejší zpracování jazyka
- např. vztahy mezi slovy → logické konstrukce
- identifikace frází v textu
- ...

Metoda postupné segmentace věty

Základní myšlenky:

- některé syntaktické jevy jsou lépe rozpoznatelné než jiné
- nejprve určíme snadnější vztahy, dále pokračujeme složitějšími
- z každé úrovně dostaneme parciální syntaktickou informaci

Principy:

- využití principů parciální analýzy pro analýzu úplnou
- rozdělení procesu analýzy do několika vrstev
- pravidlový systém – množina vzorků
- **pattern matching** – vyhledávání vzorků v textu

Jazyk pro definici pravidel

Každé pravidlo obsahuje dvě části – **šablonu** a **akce**

- šablona určuje, co se v textu má hledat
- akce určují, jaké syntaktické vztahy mají být vyznačeny
- a morfologické shody
- pravděpodobnostní ohodnocení nalezených vzorků – délka, pravděpodobnost pravidla

Příklady pravidel:

```
prep ... noun       AGREE 0 2 c MARK 2 DEP 0
noun ... noun2      MARK 2 DEP 0

[tag k1] ... [tag k1c2]       MARK 2 DEP 0

verb ... comma conj ... verb ... bound      MARK 2 7 <relclause>
```

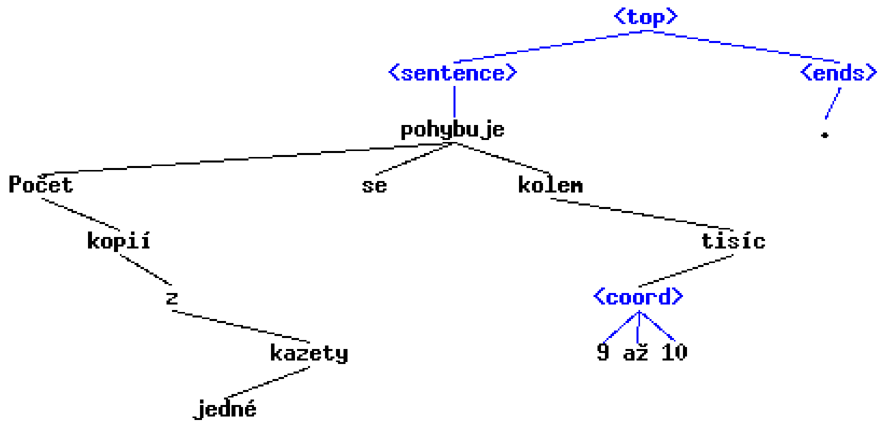

Výstup analýzy

Tzv. **hybridní stromy** – kombinují závislostní a složkové prvky

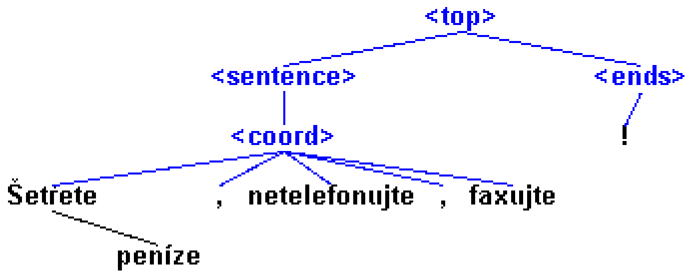
- čitelnější pro člověka
- rozlišování složkových a závislostních jevů je výhodou při analýze
- možnost převodu do čistě závislostního i čistě složkového formátu

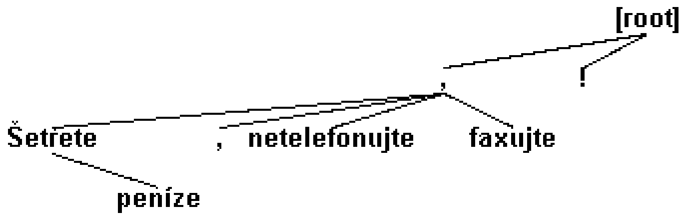
Na výstupu analýzy je vždy **jediný strom**, na stderr se vypisují **všechny nalezené vzorky** – zachycení možné víceznačnosti

Hybridní strom – příklad



Hybridní a závislostní strom





Implementace – systém SET

„Syntax in Elements of Text"

- implementace v jazyce Python
- objektový model věty, pravidel a syntaktických vztahů
- ucelený soubor pravidel pro analýzu syntaxe češtiny
- 3000 řádků kódu, 50 pravidel

Funkce:

- analýza morfologicky označkovaného textu
- výstup ve formě různých typů stromů, frází a kolokací
- reprezentace víceznačnosti ve formě výpisů na `stderr`
- grafická vizualizace výstupu

Přesnost a rychlost

Přesnost závislostního výstupu (vzhledem k datům z PDT):

Testovací sada	Přesnost – průměr	Přesnost – medián
PDT e-test	76,14 %	78,26 %
BPT2000	83,02 %	87,50 %
PDT50	92,68 %	94,99 %

Rychlost:

- asymptoticky $O(R\,N\,log(R\,N))$
- v praxi 0.14 sekundy na větu

Shrnutí

Syntaktická analýza metodou postupné segmentace věty:

- postupně vyhledáváme vzorky v textu (**pattern matching**)
- vybíráme a vyznačujeme nejpravděpodobnější z nich

Výhody navrženého přístupu:

- jednoduchost a průhlednost ve srovnání s formálními přístupy
- čitelnost kódu (Python vs. C)
- čitelnost množiny pravidel
- nezávislost na anotovaných datech

`http://nlp.fi.muni.cz/projects/set`