Zpracování pojmenovaných entit v českých textech

Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza

Technická zpráva ÚFAL/CKL TR-2007-36
ISSN 1214-5521
Únor 2007

Abstrakt

Tato technická zpráva shrnuje výsledky práce na tématu pojmenovaných entit v Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy v Praze v letech 2005 a 2006. Obsahuje rešerši zahraničních přístupů k tomuto tématu, vlastní návrh klasifikace pojmenovaných entit v češtině, popis ruční anotace pojmenovaných entit na vzorcích z Českého národního korpusu, základní kvantitativní vlastnosti anotovaných dat a výsledky prvních experimentů s automatickým rozpoznáváním pojmenovaných entit v českých textech.

Vznik této zprávy byl podpořen z projektů 1ET101120503, GD201/05/H014 a MSM0021620838.

Obsah

Kapitola 1

Úvod

1.1 Termín pojmenované entity

Termín *pojmenované entity* používáme jako český ekvivalent anglického termínu *named entities*, který je v posledním desetiletí jedním z klíčových pojmů počítačového zpracování přirozeného jazyka (*natural language processing*, NLP), především v souvislosti s extrakcí informací (*information extraction*, IE) a zodpovídáním otázek (*question answering*, QA), ale také např. se strojovým překladem.

Za pojmenované entity jsou považována slova a slovní spojení, která v textu vystupují jako jména osob, geografické názvy, jména produktů, názvy organizací, ale také jako časové údaje apod. – tedy výrazy, které nemají apelativní význam, ale odkazují ke konkrétní osobě (např. příjmení *Brousek*), časovému úseku (např. číslo ve spojení *rok 1959* nemá běžný kvantifikační význam, ale odkazuje ke konkrétnímu časovému úseku v trvání jednoho roku) apod. Pro tyto výrazy je charakteristické, že tvoří více či méně složité konstrukce (srov. např. časové údaje, poštovní adresy), jejichž vnitřní struktura se ovšem neřídí syntaktickými pravidly daného jazyka, ale spíše odpovídá společenským konvencím. Pro popis struktury pojmenovaných entit se tedy zpravidla nehodí zavedené syntaktické pojmy. Pojmenované entity jsou však častou a podstatnou součástí textů – jejich identifikace a porozumění jejich struktuře má pro automatické zpracování přirozeného jazyka zásadní význam.

V technické zprávě jsme se rozhodli používat přímý ekvivalent anglického termínu hlavně proto, že pro jména lidí, geografické názvy, časové údaje apod., které jsou zpracovávány v souvislosti s počítačovým zpracováním jazyka, v češtině dosud žádný termín zaveden nebyl – užití tradičního ligvistického termínu 'vlastní jména' se zde nezdá být vhodné.

1.2 Pojmenované entity vs. vlastní jména

Vlastní jména, jejich původ a vlastnosti jsou předmětem onomastiky, jedné ze subdisciplín jazykovědy. Encyklopedický slovník češtiny ([Karlík et al., 2002], s. 205) definuje vlastní jméno jako 'speciální jazykový prostředek mající charakter substantiva nebo pojmenovacího spojení, jehož funkcí je označovat jedince, jednotlivinu nebo jako jednotlivinu chápané množství, odlišovat je od

ostatních jedinců nebo jednotlivin dané třídy a identifikovat je jako jedinečné objekty'. Jedná se o jména osob (*Bohuslav, Mácha, Petr Veliký*), příslušníků národů (*Američan*), názvy kontinentů, zemí, obcí atd. (*Severní Amerika, Německo, Karlovy Vary*), ulic (*Na Příkopě*), organizací a institucí (*Evropská unie*), jména artefaktů (*Vyšebrodský oltář, Rusalka*), označení významných událostí (*Velká francouzská revoluce*) apod. V češtině se vlastní jména typicky píšou s velkým počátečním písmenem (pravidla pro psaní velkých písmen u těchto slov jsou uvedena v Pravidlech českého pravopisu ([Hlavsa et al., 1998]).

Za pojmenované entity (míníme zde *named entities* v anglickém kontextu, český termín dosud nebyl užíván, srov. výše) jsou však zpravidla považována nejen vlastní jména, ale i další výrazy – souhrnně se často hovoří o výrazech 'jednoznačně' odkazujících např. k objektům, k časovým okamžikům. V internetové encyklopedii Wikipedia[1] je (s odkazem na teorii reference) k tomuto termínu uvedeno následující:[2]

> In the expression *named entity*, the word *named* restricts the task to those entities for which one or many rigid designators, as defined by Kripke, stands for the referent. For instance, the *automotive company created by Henry Ford in 1903* is referred to as *Ford* or *Ford Motor Company*. Rigid designators include proper names as well as certain natural kind terms like biological species and substances.
>
> There is a general agreement to include temporal expressions and some numerical expressions such as money and measures in named entities. While some instances of these types are good examples of rigid designators (e.g., the year 2001) there are also many invalid ones (e.g., I take my vacations in "June"). In the first case, the year *2001* refers to the *2001st year of the Gregorian calendar*. In the second case, the month *June* may refer to the month of an undefined year (*past June*, *next June*, *June 2020*, etc.). It is arguable that the named entity definition is loosened in such cases for practical reasons.

Termíny 'pojmenované entity' a 'vlastní jména' se však neliší pouze svým rozsahem.

Podstatný rozdíl mezi oběma termíny je v tom, v jakých oblastech jsou používány: o vlastní jména se mluví v lingvistické oblasti, o pojmenovaných entitách se hovoří v souvislosti s počítačovým zpracováním přirozeného jazyka.

V této technické zprávě budeme užívat oba termíny. o vlastních jménech budeme hovořit tak, jak to odpovídá lingvistickému pojetí tohoto termínu. Pojmenované entity pak budeme používat jako termín širší.

[1] `http://en.wikipedia.org/wiki/Named_Entity_Recognition`

[2] Uvedené pojetí termínu 'pojmenované entity' má blízko k významu termínu 'vlastní jména', který je v Encyklopedickém slovníku češtiny u tohoto termínu také uveden ([Karlík et al., 2002], s.206).

1.3 Pojmenované entity a počítačové zpracování přirozeného jazyka

Výrazy označované jako pojmenované entity jsou běžnou součástí textů psaných v přirozených jazycích. Čtenář znalý příslušného jazyka snadno porozumí, že slovo *křeček*, které je běžným obecným jménem (např. *Spižírna je pro křečka důležitá věc...*, SYN2000)[3], vystupuje v kontextu *JUDr. Václav Křeček* jako příjmení osoby, kdežto ve větě *Pokud můžete dát zvíře ke známým, ponechte jej, jak doporučuje Otto von Frisch, autor knihy Křeček – jak o něj pečovat a jak mu rozumět, raději v kleci, na kterou je zvyklé* (SYN2000) jako součást názvu knihy. Stejně tak je čtenáři jasné, že ve větě *V románě jsem na straně 485, zbývá asi třicet stran...* (SYN2000) odkazuje číslo *485* ke konkrétní straně, ale číslovka *třicet* vyjadřuje množství stránek (a uvedený zápis je rovnocenný zápisu *30*).

V souvislosti s rozvojem počítačového zpracování přirozeného jazyka se však ukázalo jako nezbytné, aby pojmenované entity byly v textech identifikovány a klasifikovány automaticky. I když v psaném českém textu jsou vlastní jména v mnoha případech signalizována velkým počátečním písmenem, pro automatickou identifikaci těchto slov to zdaleka není dostačující:[4] např. ve větě *Křečka reportéři vyzdvihli za jeho úsilí seznámit veřejnost s dopady zněčištění životního prostředí v oblasti takzvaného Černého trojúhelníku...* (SYN2005) nelze bez porozumění obsahu věty (a bez odpovídajících znalostí světa) a případně bez znalosti kontextu určit, zda se velké písmeno ve slově *křečka* vyskytuje pouze proto, že toto slovo stojí na začátku věty, nebo zda jde o vlastní jméno. Určit, k čemu daný výraz odkazuje (zda jde o příjmení nebo o název organizace apod., o číslo telefonu nebo množstevní údaj), je někdy na základě signálů obsažených v textu možné, v mnoha případech však takové signály nejsou spolehlivé nebo chybějí úplně. Např. slovo následující po akademickém titulu (*ing., JUDr.* apod.) bude v mnoha případech křestní jméno, ale stejně tak může jít o příjmení; číslo následující po řetězcích *tel.* nebo *telefon* (popř. *Tel.*, *Telefon*, *Tel./fax* apod.) bude pravděpodobně telefonní číslo, ovšem telefonní čísla, kterým takový řetězec nepředchází, podle tohoto kontextu nalézt nelze; navíc řada jiných pojmenovaných entit se v pevném lexikálním okolí nevyskytuje.

Aby bylo možné pracovat s pojmenovanými entitami obsaženými v textu pomocí automatických nástrojů, musejí být pojmenované entity v textu vyznačeny explicitně. Značky mohou být do textů umisťovány ručně, ovšem ruční anotace velkého množství textů je z řady důvodů (časových, finančních ad.) nezvladatelná, značkování textů tedy musí probíhat automaticky (značkovací nástroje jsou nazývány *named entity tagger*). Na základě těchto značek pak např. systém zodpovídání otázek vybere z textu informaci, která je správnou odpovědí na zadanou otázku. Např. položí-li uživatel systému otázku typu *Kdo získal v roce 1998 Nobelovu cenu za literaturu*, systém pro sestavení odpovědi bude hledat

[3] SYN2000 je lematizovaný a morfologicky anotovaný korpus vytvořený v Ústavu Českého národního korpusu na Filozofické fakultě Univerzity Karlovy v Praze, tento korpus má rozsah 100 milionů slovních jednotek.

[4] Myslíme zde běžné nestrukturované texty. Texty strukturované (např. *Autor: Ing. arch. Zdeněk Hanuš Generální projektant: ARKO, spol. s r. o., Hradec Králové Investor: město Hradec Králové*) tu nebudeme zvlášť rozebírat.

jazykový kontext, ve kterém se (v ideálním případě) vyskytují následující pojmenované entity: jméno osoby (to je pak samotnou odpovědí), rok (tato entita musí mít hodnotu *1998*) a název ocenění (*Nobelova cenu za literaturu*).

Přesné definice termínu pojmenované entity, hlavně co do jeho rozsahu, se v jednotlivých koncepcích liší – rozdílné jsou pak i sady značek užívané pro značkování textů (liší se co do počtu značek, jemnosti i hierarchického uspořádání). Sady značek definované v rámci koncepcí, které jsme se rozhodli probrat podrobněji, představíme v jednotlivých sekcích kapitoly 2. V sekci 2.8 pak nastíníme různé metody automatického zpracování pojmenovaných entit. Systémy, v nichž jsou značkované entity využívány (extrakce informací, zodpovídání otázek ad.), zde probírány nebudou.

1.4 Pojmenované entity v počítačovém zpracování češtiny

Ačkoli je automatické zpracování pojmenovaných entit předmětem zájmu počítačové lingvistiky už řadu let, pro češtinu zatím tato problematika nebyla soustavněji řešena.

Za první krok učiněný za účelem zpracování výrazů označovaných jako pojmenované entity v češtině lze považovat klasifikaci vlastních jmen, která je součástí Pražského závislostního korpusu a dalších korpusů češtiny. Tuto klasifikaci a další zdroje umožňující identifikaci pojmenovaných entit v čekých textech představíme v kapitole 3. Zpracováním jednoho typu pojmenovaných entit, bibliografických údajů (údajů o autorovi, názvu textu atd.), pro češtinu se ve své bakalářské práci s názvem *Získávání metainformací z textové podoby elektronicky dostupných článků* zabýval Marek Rychlý z Masarykovy univerzity v Brně ([Rychlý, 2003]).[5]

Rozpoznávání a klasifikace pojmenovaných entit v českých textech je jedním z cílů projektu Informační společnosti Grantové agentury Akademie věd České republiky nazvaného *Integrace jazykových zdrojů za účelem extrakce informací z přirozených textů* (1ET101120503). V této technické zprávě představíme jednotlivé úkoly zpracované v rámci tohoto projektu – od sestavení sady značek pro anotaci pojmenovaných entit v českých textech přes ruční anotaci trénovacích dat až k experimentům s automatickým zpracováním pojmenovaných entit.

[5] `http://is.muni.cz/th/51598/fi_b/txt2rim.rtf`

Kapitola 2

Zpracování pojmenovaných entit v zahraničních přístupech

2.1 Konference MUC (Message Understanding Conference)

Termín pojmenované entity byl zaveden v souvislosti s šestou konferencí Message Understanding Conference (MUC-6) pořádanou v roce 1995 ve Spojených státech. Abychom blíže objasnili, na jakém pozadí byl tento termín do zpracování přirozeného jazyka zaveden, v následujících odstavcích stručně probereme historii konferencí MUC. Uvedené informace jsou čerpány především z článku [Grishman and Sundheim, 1996b].

Hlavními cíli konferencí MUC, jichž se od konce 80. do konce 90. let konalo celkem sedm, byla podpora a vyhodnocování výzkumu v oblasti extrakce informací. V zásadě se však nejednalo o konference v obvyklém slova smyslu: týmy, které se chtěly konference účastnit, obdržely v předstihu zadání úkolu, tj. jaké informace mají být v textu vyhledány, a texty, na nichž byly vyhledávací systémy vyvíjeny (trénovací data). Těmito systémy pak byl zpracován soubor testovacích dat, dosažené výsledky byly porovnány s ručně připravenými odpověďmi a systémy vyhodnoceny. Na samotné konferenci pak byly jednotlivé systémy prezentovány a porovnávány.

První z těchto konferencí (MUC-1, 1987) byla iniciována střediskem NOSC (Naval Ocean Systems Center) společně s DARPA (Defense Advanced Research Projects Agency) a úkolem účastníků bylo vyvinout systém pro automatickou analýzu vojenských zpráv. Zatímco na této konferenci navrhl každý tým pro výsledky analýzy svůj vlastní formát a systémy nebyly formálně vyhodnoceny, v rámci druhé konference MUC (MUC-2, 1989) byla pro každou událost popisovanou v analyzovaném textu sestavena tzv. šablona (*template*) obsahující deset míst (*slots*; např. typ události, o niž v textu jde, činitel, čas a místo události atd.). Úkolem vyvíjených systémů bylo do jednotlivých míst těchto šablon doplnit požadované informace.

Systémy vyvinuté pro MUC-2 byly vyhodnoceny pomocí parametrů *recall* (úplnost výsledku) a *precision* (přesnost). Parametr *recall* vyjadřoval, jaký je

poměr počtu míst, které daný systém vyplnil správně (za správně vyplněná místa jsou považována ta, která odpovídají ručně vyplněné šabloně), k počtu míst, které měly být podle ručně vypracovaných odpovědí vyplněny. Parametr *precision* byl definován jako poměr počtu míst, které systém vyplnil správně, k počtu míst, které systém celkem vyplnil (správně i nesprávně).

Na dalších konferencích (MUC-3, 1991; MUC-4, 1992; MUC-5, 1993) byl počet míst v šablonách dále rozšiřován. Příklad zprávy analyzované v rámci konference MUC-3 a příslušnou vyplněnou šablonu lze nalézt například v článku [Grishman and Sundheim, 1996a]. V rámci MUC-5, která byla organizována jako součást amerického vládního programu Tipster pro výzkum a vývoj v oblasti vyhledávání a extrakce informací, se navíc od jednoduchého přiřazení jediné šablony jedné události přešlo ke komplexnější struktuře šablon pro jedinou událost.

V rámci první a druhé konference MUC byly analyzovány vojenské zprávy týkající se námořních operací. Třetí a čtvrtá konference se zabývala zprávami informační služby Foreign Broadcast Information Service (FBIS) o teroristických útocích ve Střední a Jižní Americe. Tématem zpráv pro MUC-5 byly podniky se zahraniční účastí (*joint ventures*) a výroba elektronických obvodů (*electronic circuit fabrication*), a to poprvé ve dvou jazycích – kromě dosavadní angličtiny také v japonštině.

2.2 Konference MUC-6

Zatímco v rámci prvních pěti konferencí MUC řešily zúčastněné týmy po řadu měsíců vždy jeden úkol, který byl pro všechny společný, bylo pro šestou konferenci definováno úkolů více. Jedním z cílů této konference, tak jak byly vymezeny na společné schůzce agentury DARPA, účastníků programu Tipster a zástupců americké vlády v prosinci 1993, bylo vyvinout pro extrakci informací takové dílčí technologie, které budou nezávislé na konkrétním úkolu a které bude možno bezprostředně využívat pro různé nástroje. V rámci této konference tedy měly být vyvinuty systémy, které by v textech identifikovaly jména lidí, názvy organizací, geografické názvy apod. – pro tyto výrazy byl zaveden termín pojmenované entity. Podrobnosti týkající se definice tohoto úkolu (*named entity recognition task*) a formát anotace vyvinutý pro tento úkol uvedeme v 2.3.

Dále byl v rámci MUC-6 řešen tentýž úkol jako na předchozích konferencích: vyhledat v textech požadované informace a umístit je do šablon. Struktura šablon byla však oproti MUC-5 zjednodušena: každé události odpovídala jedna šablona (*event-level template*), která prostřednictvím další šablony odkazovala k šablonám reprezentujícím prvky zapojené do dané události (lidi, organizace, produkty apod.) – tyto šablony nejnižší úrovně (*low-level templates*) byly nazývány prvky šablon (*template elements*). Prvek šablony např. pro organizace obsahoval šest míst (název, město, země atd.) a vyplňována byla pouze ta místa, k nimž byla odpovídající informace v textu uvedena explicitně (v případě země mohla být informace odvozena z uvedeného města). Sestavování jednotlivých prvků šablon bylo vyhlášeno jako další z úkolů pro MUC-6 (označovaný jako *template element task* oproti výše popsanému úkolu vyplňování šablon nazvanému *scenario template task*).

Posledním úkolem, který byl řešen v rámci MUC-6 a který byl definován s cílem směřovat k hlubšímu porozumění analyzovaným textům, bylo identifikovat v textu vztahy koreference, tedy výrazy odkazující k témuž prvku (*coreference task*).

Tématem textů, s nimiž se pracovalo v rámci MUC-6, byly změny ve výkonném managementu společností. Vyhodnocení všech čtyř úkolů se konalo v září 1995, samotná konference pak v listopadu téhož roku v Marylandu.

2.3 Formát vyvinutý pro konferenci MUC-6

Termín pojmenované entity byl v rámci MUC-6 zaveden jako označení pro výrazy 'jednoznačně identifikující' osoby, organizace a místa (*entities*), časové údaje (*times*) a dva typy množstevních údajů (*quantities*; finanční částky a množství procent). Úkolem zúčastněných týmů bylo vyvinout takový nástroj, který nalezne všechny výskyty pojmenovaných entit (tedy výrazů uvedených tří druhů: *entities, times, quantities*) v každém textu, který je součástí testovacích dat, a určí jejich typ. Správným výstupem je jediná jednoznačná značka pro každou entitu. Pro vymezení pojmenované entity a určení jejího typu byl používán značkovací jazyk SGML. Pro tři druhy pojmenovaných entit byly zavedeny tři SGML tagy:

- `ENAMEX` (entities)
- `TIMEX` (times)
- `NUMEX` (quantities)

Rozsah entity byl vymezen vždy dvojicí tagů (počátečním a koncovým tagem stejného druhu). Typ takto vymezené entity (jako SGML elementu) pak byl zachycen hodnotou atributu `TYPE`. Pro každý z tagů byla definována jiná množina hodnot atributu `TYPE`.

V případě tagu `ENAMEX` mohl atribut `TYPE` nabývat tří hodnot:

- `ORGANIZATION` (název firmy či státní nebo jiné organizace)
- `PERSON` (jméno osoby / rodiny)
- `LOCATION` (název politicky nebo geograficky vymezeného místa, tedy např. jméno města, názvy pohoří apod.)

Př. *pracuje v Deloitte and Touche*:
`<ENAMEX TYPE="ORGANIZATION">Deloitte and Touche</ENAMEX>`

Z časových údajů měly být jako pojmenované entity identifikovány pouze 'absolutní' údaje (např. *10. října* nebo *v roce 2005*, nikoli indexické výrazy typu *dnes* nebo *loni* apod.). Pro tag `TIMEX` byly definovány následující dvě hodnoty atributu `TYPE`:

- `DATE` (datum, rok, období apod.)
- `TIME` (čas, např. ve 12 hodin)

Př. *v roce 1999*: <TIMEX TYPE="DATE">roce 1999</TIMEX>

Rozpoznávání množstevních údajů bylo omezeno pouze na finanční částky a procenta – jako pojmenované entity přitom měly být identifikovány číselné výrazy psané slovy i číslicemi. V případě tagu NUMEX byly tedy pro atribut TYPE definovány dvě hodnoty:

- MONEY,
- PERCENT.

Př. *180 milionů dolarů*:
<NUMEX TYPE="MONEY">180 milionů dolarů</NUMEX>

Z uvedeného seznamu typů pojmenovaných entit je zřejmé, že pojmenované entity byly definovány relativně úzce. Za pojmenované entity nebyly v pojetí definovaném pro MUC-6 považovány např. názvy politických stran, názvy cen (pokud byly pojmenovány po lidech, nebylo jako entita vyznačeno ani obsažené jméno, např. *Nobelova cena*) atd. V tomto pojetí se nepočítalo ani se zanořováním pojmenovaných entit – např. pokud název společnosti obsahoval geografické jméno, byl celý název vyznačen jako organizace, geografické jméno však vyznačeno nebylo.[1]

Př. *Hyundai of Korea, Inc.*:
<ENAMEX TYPE="ORGANIZATION">Hyundai of Korea, Inc.</ENAMEX>)

V dokumentu [NED, 1995] vydaném pro MUC-6, z něhož pocházejí výše uvedené informace a příklady (popř. anglické předlohy příkladů), jsou podrobně řešeny i speciální konstrukce, jako jsou konstrukce s elidovanými členy (př. *Severní a Jižní Amerika*), koordinační struktury apod. – tyto případy zde nebudeme popisovat. Související pravidla pro tokenizaci analyzovaných textů (např. ve kterých případech je tečka součástí entity a ve kterých nikoli) byla uvedena ve zvláštním dokumentu (viz [Tok, 1995]). Koncepce pojmenovaných entit vyvinutá pro MUC-6 byla přijata jako obecný formát.

Zde uvádíme příklad věty s vymezenými a klasifikovanými pojmenovanými entitami podle pokynů vydaných pro konferenci MUC-6 (ukázka převzata ze článku [Grishman and Sundheim, 1996b]):

```
Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with
<ENAMEX TYPE="PERSON">Martin Puris</ENAMEX>, president
and chief executive officer of <ENAMEX TYPE="ORGANIZATION">
Ammirati & Puris</ENAMEX>, about <ENAMEX TYPE="ORGANIZATION">
McCann </ENAMEX>'s acquiring the agency with billings of
<NUMEX TYPE="MONEY">$400 million</NUMEX>, but nothing has
materialized.
```

[1] Výjimkou jsou místní údaje obsažené v časových údajích, zde je zanoření místní entity do časové vyžadováno - př. *1:30 p.m. Chicago time*: <TIMEX TYPE="TIME">1:30 p.m. <ENAMEX TYPE="LOCATION">Chicago</ENAMEX> time</TIMEX>.

V rámci MUC-6 řešilo úkol rozpoznávání pojmenovaných entit celkem 15 týmů, které vyvinuly 20 systémů. Testovací data označkovaná vyvinutými systémy byla porovnána s ruční anotací těchto dat.

Na rozdíl od automatického značkování měli anotátoři možnost vyplnit v atributu `TYPE` více alternativních hodnot, pokud nebyli ani na základě kontextu či znalosti světa schopni rozhodnout, která z nich je správná (jednotlivé hodnoty byly oddělovány svislou čarou). Pro případy, kdy si anotátoři nebyli jisti, zda má být daný výraz vyznačen jako pojmenovaná entita, byl zaveden zvláštní atribut `STATUS`: jeho přítomnost znamenala, že vyznačení výrazu jako entity je fakultativní. Další atribut (`ALT`) mohli anotátoři použít v případě, kdy si nebyli jisti rozsahem pojmenované entity: jako entita byl vyznačen výraz většího rozsahu, v atributu `ALT` pak byla uvedena část tohoto výrazu – odpověď systému pak byla považována za správnou, ať už odpovídala celému výrazu, nebo části uvedené v atributu (český příklad ekvivalentní anglickému příkladu uvedenému v [NED, 1995]: jako pojmenovaná entita vyznačen výraz *celý rok 1987*, v atributu `ALT` pak pouze *rok 1987*).

Úspěšnost systémů vyvinutých pro MUC-6 byla hodnocena pomocí parametrů pokrytí a přesnost. Většina z nich dosáhla pokrytí i přesnosti vyšší než 90 %. Nejlepší systém ([Grishman and Sundheim, 1996b]) měl pokrytí 96 % a přesnost 97 %; podrobný přehled výsledků viz [Sundheim, 1995].

2.4 Rozšířené klasifikace pojmenovaných entit založené na formátu MUC-6

Projekty navazující na konference MUC sadu kategorií, která byla navržena v rámci této konference (tj. tři druhy entit, v jejichž rámci bylo rozlišeno celkem sedm typů, srov. sekce 2.3), mírně rozšiřovaly, popř. modifikovaly. Např. v projektu IREX (sekce 2.6) byla množina typů rozšířena o typ `ARTIFACT`; konference CoNLL (sekce 2.6) se soustředila na čtyři typy pojmenovaných entit, jména osob, geografické názvy, názvy organizací a ostatní názvy.

Obecně však lze říci, že s rozvojem automatického zpracování přirozeného jazyka (stále nových aplikací v oblasti extrakce informací a zodpovídání otázek) je třeba stále jemnějšího třídění pojmenovaných entit a dochází tedy k prudkému rozšiřování sady užívaných značek. Hierarchie pojmenovaných entit sestávající z 29 typů a 64 podtypů (např. názvy uměleckých děl jako typ, v jehož rámci je rozlišováno, zda jde o název knihy, hry, písně, obrazu, sochy nebo o jiný název) byla sestavena firmou BBN pro systém zodpovídání otázek ([Brunstein, 2002]). Ještě podrobnější je Sekinova 'rozšířená hierarchie pojmenovaných entit' ([Sekine, 2003]). Cílem, který byl při sestavování této hierarchie sledován, bylo pokrýt hlavní typy pojmenovaných entit vyskytující se v novinových článcích jakéhokoli zaměření. Sekine ([Sekine, 2004]) uvádí dva zásadní problémy, s nimiž je nutno se při radikálním rozšířením počtu typů pojmenovaných entit vyrovnat:

- jak kategorie definovat, tzn. co má být vymezeno jako samostatná kategorie ('problém kategorizace světa do sémantických kategorií a určení správně kategorie pro každé slovo (každý jeho výskyt)');

- jakými metodami text definovanými značkami označkovat, především jak zajistit, aby bylo značkování konzistentní (některé typy se v textech vyskytují příliš zřídka).

Sekine navrhuje možné řešení uvedených problémů: opustit metody řízeného učení a přejít k metodám využívajícím částečně řízené učení (bootstrapping) nebo neřízeného učení (clustering nebo extrakce informací na základě lingvistických znalostí).

2.5 Konference MET (Multilingual Entity Task)

Poslední, sedmá konference MUC se uskutečnila v roce 1998. Současně s ní byla uspořádána v pořadí již druhá konference s názvem Multilingual Entity Task Evaluation (MET). Konference MET, podporované programem Tipster, se omezují na rozpoznávání pojmenovaných entit a řeší tento úkol pro jiné, typologicky odlišné jazyky než angličtinu. V rámci MET-1, která se konala v roce 1996, byly vyhodnoceny systémy na rozpoznávání pojmenovaných entit v japonských, čínských a španělských novinových článcích. Na MET-2 byly analyzovány rovněž japonské a čínské texty, experimentálně také texty v thajštině. Systémy vyvíjené pro tyto konference dávají výstup ve formátu definovaném pro konferenci MUC-6 (srov. sekce 2.3).[2]

2.6 Projekty navazující na MUC a MET

Zatímco úkolem systémů vyvíjených v rámci prvních ročníků konference MUC bylo identifikovat v konkrétním typu textů požadované informace a umístit je do předem definovaných šablon, šestá konference definovala identifikaci takových informací (souhrnně zde označených jako pojmenované entity) a určení jejich typu jako obecný úkol nezávislý na konkrétní aplikaci, jehož řešení má pro automatické zpracování textů psaných v přirozeném jazyce podstatný význam. Na konference MUC, které se soustředily na anglické texty, navázaly konference MET, které tento úkol rozšířily na další jazyky.

Po konferencích MET následovaly obdobně zaměřené projekty i mimo USA. Rozpoznávání pojmenovaných entit bylo např. jedním z úkolů projektu IREX (Information Retrieval and Extraction Exercise), který probíhal od roku 1998 do roku 1999 v Japonsku,[3] v letech 2002 a 2003 pak bylo vyhlášeno jako společný úkol konferencí CoNLL (Conference on Computational Natural Language Learning). Úkolem zde bylo vyvinout systém pro rozpoznávání pojmenovaných entit s využitím metod strojového učení. Tento systém měl pracovat nezávisle na konkrétním jazyce.[4,5] Konference CoNLL tak navázala na předchozí experimenty, v jejichž rámci byl jeden systém rozpoznávání pojmenovaných entit aplikován na texty v různých jazycích, např. výsledky rozpoznávání pojmenovaných entit v čínských, anglických, francouzských, japonských, portugalských

[2] `http://www-nlpir.nist.gov/related_projects/tipster/met.htm`
[3] `http://nlp.cs.nyu.edu/irex/index-e.html`
[4] `http://www.cnts.ua.ac.be/conll2002/ner/`
[5] `http://www.cnts.ua.ac.be/conll2003/ner/`

a španělských zpravodajských textech s použitím statistických metod publikovali Palmer a Day ([Palmer and Day, 1997]), rozpoznáváním entit v angličtině, řečtině, hindštině, rumunštině a turečtině na základě morfologických a kontextových informací se zabývali autoři článku [Cucerzan and Yarowsky, 1999].

V posledních letech byla realizována celá řada projektů zaměřených na rozpoznávání pojmenovaných entit – pozornost se přitom ubírá různými směry. Jsou vyvíjeny systémy pro zpracování pojmenovaných entit v konkrétních jazycích s cílem dosáhnout dalšího zlepšení výsledků, rovněž jsou vyvíjeny systémy využívající nových metod. Kromě toho vzniká řada systémů zabývajících se rozpoznáváním pojmenovaných entit v textech z omezené domény, např. systémy rozpoznávající biochemické termíny. Ve snaze opustit dosavadní poměrně úzké pojetí úkolu rozpoznávání pojmenovaných entit a zaměřit sémantickou anotaci korpusů šíře, byl v rámci konference LREC v roce 2004 uspořádán workshop s názvem Beyond Named Entity Recognition.

2.7 Značkování pojmenovaných entit podle směrnic TEI

Ve směrnicích *Guidelines for Electronic Text Encoding and Interchange*, které vydává Text Encoding Initiative (TEI), jsou definovány prostředky pro značkování jakéhokoli elektronického textu v přirozeném jazyce: struktura textu (např. rozsah odstavců, přímé řeči apod.) a informace, které jsou v textu obsaženy, se po vyznačení definovanými značkami stávají strojově čitelnými, a mohou být tedy zpracovávány počítačovými programy. Směrnice TEI používají značkovací jazyk SGML, od verze P4 ([TEI, 2003]), která byla vydána v roce 2003, pak jazyk XML.

Záběr směrnic TEI je velmi široký – zde se zaměříme pouze na značky pro explicitní vyznačení výrazů a frází, které v textu vystupují jako pojmenování osob, organizací a časových údajů. Se značkováním takových výrazů počítají směrnice TEI od své první verze zvané P1, která byla vydána v roce 1990. Ve verzi P4 ([TEI, 2003]) je jim věnována zvláštní kapitola, kap. 2.3.4. *Names, Numbers, Dates, Abbreviations, and Adresses* (ve II. části směrnic s názvem *Core Tags and General Rules*). Na začátku této kapitoly je potřeba vyznačovat takové výrazy[6] zdůvodněna následovně ([TEI, 2003], kap. II.2.3.4):

> This section describes a number of textual features which it is often convenient to distinguish from their surrounding text. Names, dates, and numbers are likely to be of particular importance to the scholar treating a text as source for a database; distinguishing such items from the surrounding text is however equally important to the scholar primarily interested in lexis.

V uvedené kapitole je upraveno značkování jmen a názvů, čísel a měr, časových údajů, zkratek a adres.

[6] Dále budeme i zde hovořit o pojmenovaných entitách, i když v kontextu TEI není tento termín používán.

Pro značkování jmen a názvů (hovoří se zde obecně o *reffering strings* – řetězcích referujících k určité osobě, místu, objektu atd.) jsou zavedeny dva tagy:

- `rs`
- `name`

Pokud je to pro řešení konkrétní úlohy užitečné, může být typ pojmenované entity vyznačený jedním z těchto tagů určen hodnotou atributu `TYPE`. Hodnotami tohoto atributu pak mohou být např. `person`, `place` nebo `ship` – jejich množina ovšem není pevně dána, hodnoty mohou být 'vhodně' dodefinovány podle konkrétního úkolu.

Touto otevřenou povahou se značkovací pravidla uvedená ve směrnicích TEI zásadně liší od koncepce vyvinuté pro konferenci MUC, jejíž striktní definice značkovacího jazyka i tematické vymezení textů byly nutným předpokladem pro vývoj vzájemně si konkurujících systémů: směrnice TEI jsou koncipovány jako obecná pravidla a doporučení pro značkování elektronických textů jakéhokoli literárního žánru, jakékoli formy a jakéhokoli stáří.

Zatímco tag `name` má být přiřazován pouze vlastním jménům, tedy v souladu s tím, co je za pojmenované entity považováno v koncepci MUC), v případě tagu `rs` počítají směrnice TEI s tím, že v textech mohou být kromě vlastních jmen značkovány např. i obecné výrazy referující ke konkrétní osobě, organizaci apod. – srov. označkování výrazu *his lady* v následující větě (tag `q` vyznačuje přímou řeč; [TEI, 2003], kap. II.2.3.4.1):

```
<q>My dear <rs type="person">Mr. Bennet</rs></q>, said
<rs type="person">his lady</rs> to him one day ...
```

Vedle atributu `type` jsou v souvislosti s tagy `rs` a `name` uvedeny také atributy `key` a `reg`. První z nich, atribut `key`, slouží jako prostředek pro provázání pojmenovaných entit v textu, které odkazují k téže osobě, organizaci apod.: za jeho hodnotu se zvolí (libovolný) řetězec, který se pak vyskytne u všech entit referujících k téže jednotce.[7] V atributu `reg` se uvádí 'normalizovaná' podoba entity, která se vyskytla v textu – u jmen osob např. nejdříve příjmení, potom jméno (takový prostředek je velmi užitečný např. při různých formátech dat: normalizované datum, např. `1997-05-12`, může být zapsáno řadou způsobů jako *twelfth day of May...*, *12th May 1997*, *May 12, 1997* atd.).

Tagy a příslušné atributy pro další entity řešené ve směrnicích TEI v kapitole 2.3.4. *Names, Numbers, Dates, Abbreviations, and Adresses* zde nebudeme podrobně rozebírat – principy jejich značkování jsou obdobné jako u jmen a názvů, vždy ovšem se zohledněním specifik jednotlivých druhů entit (více srov. [TEI, 2003], kap. II.2.3.4]).

Podrobnější sada tagů a příslušných atributů umožňující detailnější značkování pojmenovaných entit v textech je ve směrnicích TEI ([TEI, 2003]) uvedena zvláště v kap. 4.7. *Names and Dates* (ve IV. části směrnic s názvem *Additional Tag Sets*). I z této kapitoly přiblížíme pouze pravidla pro značkování jmen.

[7] Jde vlastně o prostředek k zachycování koreferenčních vztahů, zvl. typu zvaného 'bridging anaphora'. K pojetí koreference a jejímu zachycování v Pražském závislostním korpusu srov. [Mikulová et al., 2005].

Jako základní tag pro značkování výrazů odkazujících k osobě je zde zaveden tag `persName`. Hodnoty atributu `type`, který se může v kombinaci s tímto tagem objevovat, mohou být dodefinovány (př. možných hodnot: `married name`, `religious name` apod.). Pro jména osob se v uvedené kapitole dále počítá s tagy `surname`, `foreName`, `roleName`, `addName`, `nameLink` a `genName`, které mohou být rovněž doplněny vhodnými hodnotami atributu `type`. Pro uvedené tagy se dále počítá také s atributy `key` a `reg` (a dalšími, viz [TEI, 2003], kap. IV.4.7.1). Uvádíme zde dva příklady značkování jmen osob:

```
<persName key="MRT1"> <foreName type="given">Margaret
</foreName> <foreName type="abbrev">Maggie</foreName>
<foreName type="unused"> Hilda</foreName> <surname
type="maiden">Roberts</surname> <surname type="married">
Thatcher</surname> </persName>
<persName key="MUAL1" type="religious"> <foreName>Muhammad
</foreName> <surname>Ali</surname> </persName>
```

Tag `persName` je synonymní s výše uvedenou kombinací `<name type="person">`, což ovšem vede k tomu, že stejný text je možno značkovat několikerým způsobem. To je další podstatný rozdíl oproti značkování podle pravidel sestavených pro konference MUC. Pro ilustraci zde uvádíme čtyři rovnocenné způsoby označkování jména osoby ve větě *That silly man David Paul Brown has suffered the furniture of his office to be seized the third time for rent* ([TEI, 2003], kap. IV.4.7.1):

```
That silly man <rs key="DPB1" reg="Brown, David Paul"
type="person"> David Paul Brown</rs> has suffered the
furniture of his office to be seized the third time for rent.
That silly man <rs key="DPB1" reg="Brown, David Paul"
type="person"> <name>David Paul Brown</name> </rs> has
suffered ...
```

```
That silly man <name key="DPB1" reg="Brown, David Paul"
type="person"> David Paul Brown</name> has suffered ...
That silly man <persName key="DPB1" reg="Brown, David Paul">
David Paul Brown</persName> has suffered ...
```

Směrnice TEI byly použity pro značkování textů v řadě projektů. Seznam těchto projektů je zveřejněn na internetových stránkách Text Encoding Initiative Consortium na adrese `http://www.tei-c.org/Applications/`

2.8 Vybrané přístupy k automatickému rozpoznávání pojmenovaných entit

V následujících odstavcích odkazujeme na některé přístupy k rozpoznávání pojmenovaných entit. Pochopitelně zdaleka nejde o soustavný a vyčerpávájí přehled.

V článku *Unsupervised Models for Named Entity Classification* autoři využívají redundance, díky které je v textu v mnoha případech možné rozpoznat pojmenovanou entitu ze jména samotného i z kontextu, ve kterém je užito ([Collins and Singer, 1999]). Ukazují, že díky této redundanci je možné výrazně snížit množství označkovaných dat a k natrénování klasifikátoru lze použít neznačkovaná data. Pro inicializaci postačilo vložit pouze sedm příkladů pojmenovaných entit (*seeds*): *New York*, *California* a *U.S.* jako příklady názvů míst, *I.B.M.* a *Microsoft* jako příklady názvů organizací a informaci o tom, že výrazy obsahující *Mr.* jsou jména osob a výrazy obsahující *Incorporated* jsou jména organizací. Funkci systému lze zjednodušeně popsat takto: v neznačkovaných datech systém narazí např. na výraz *Mr. Cooper*, kde z už dostupných informací dokáže vyvodit, že *Cooper* je jméno osoby; v dalších kontextech pak zjistí, že slovu *Cooper* často předchází slovo *president*, proto lze očekávat, že slova následující za slovem *president* budou také jména osob.

V obvyklých postupech při rozpoznávání pojmenovaných entit jsou za základní jednotky považována slova, nanejvýš je pracováno s prefixy nebo sufixy o pevné délce. V článku *Named Entity Recognition with Character-Level Models* ([Klein et al., 2003]) autoři argumentují, že významnou informaci lze získat při zpracování textu po znacích (*word-internal substring features*). Tuto myšlenku realizují s využitím HMM a ve svém experimentu dosahují redukce množství chyb o 30 %.

V článku *Named Entity Discovery Using Comparable News Articles* je pro rozpoznávání pojmenovaných entit využita překvapivá myšlenka: autoři vycházejí jednak z předpokladu, že rozložení užití vlastních jmen v novinových textech podél časové osy se od rozložení užití obecných jmen liší (vlastní jména mají výraznější peaky), a jednak z toho, že vlastní jména je obtížné parafrázovat, proto lze očekávat, že jejich rozložení v paralelních řadách článků ze dvou nebo více novinových zdrojů budou více synchronizované ([Shinyama and Sekine, 2004]). Pomocí kosínové podobnosti na vektorech odpovídajících časovým řadám určí nejvíce synchronizované výrazy a ukazují, že u velké části z nich jde o pojmenované entity.

V článku *Named Entity Recognition through Classifier Combination* autoři ukazují, že pomocí kombinace několika klasifikátorů pojmenovaných entit založených na různých principech lze dosáhnout výrazného zlepšení úspěšnosti ([Florian et al., 2003]). Klasifikátory, které používají, jsou založeny na následujících metodách: Robust Risk Minimization Classifier, Maximum Entropy Classifier, Transformation-Based Learning Classifier, HMM Classifier.

V práci *Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition* ([Sassano and Utsuro, 2000]) je poukázáno na skutečnost, že v jazycích s psanou formou bez zřetelných separátorů slov (např. japonština) je významným problémem to, že hranice pojmenovaných entit často neodpovídají hranicím jednotek na výstupu morfologického analyzátoru. K rozpoznávání pojmenovaných entit používají řízené učení, konkrétně metodu *decision list learning*.

Autoři článku *Fine Grained Clasification of Named Entities* tvrdí, že v komplexních aplikacích jako *question answering* se obvyklé hrubé klasifikace (osoba, organizace, místo) jeví jako nedostatečné, a navrhují řešení, jak rozpoznávat jemnější typy jmen osob ([Fleischman and Hovy, 2002]). Osm podtříd, které

autoři zavádějí, je navrženo tak, aby pokrývaly osobní jména, která se často vyskytují v jejich korpusu. Jde o jména osob z oblastí (1) sport, (2) vláda a politika, (3) policie, (4) duchovenstvo, (5) obchod, (6) umění a zábava, (7) o právníky, (8) lékaře a vědce. Trénovací data si vytvářejí pomocí bootstrappingu: na začátku pro každou z osmi skupin založí počáteční množinu sta instancí (*seed*), na nich natrénují klasifikátor C4.5, tento klasifikátor spustí na velký korpus a z něj vyberou další instance, u kterých klasifikátor rozhodl s vyšší než 90% spolehlivostí.

Kapitola 3

Pojmenované entity v českých textech

Jak jsme již zmínili výše, v rámci automatického zpracování češtiny nebyla dosud problematice pojmenovaných entit věnována soustavnější pozornost. V první části této kapitoly představíme nejprve základní klasifikaci vlastních jmen, která je součástí morfologické anotace Pražského závislostního korpusu a dalších českých korpusů. Ve druhé sekci této kapitoly pak uvedeme některé elektronické zdroje, které by mohly být při automatickém zpracování pojmenovaných entit využity (např. seznamy českých příjmení).

3.1 Zpracování pojmenovaných entit v korpusech češtiny

Pražský závislostní korpus (Prague Dependency Treebank, PDT) je soubor českých novinových textů, které byly anotovány na několika rovinách. Anotační schéma PDT vychází z funkčního generativního popisu, závislostního popisu češtiny, který je od 60. let 20. století rozpracováván na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze (srov. zvl. [Sgall, 1967], [Panevová, 1980], [Sgall et al., 1986]). v první verzi PDT 1.0 ([Hajič et al., 2001]) jsou texty anotovány na morfologické rovině a na rovině povrchové syntaxe (tzv. analytické rovině). Ve druhé verzi PDT 2.0 ([Hajič et al., 2006]) byly texty anotovány navíc na rovině hloubkové syntaxe (tzv. tektogramatické rovině) – ke komplexní anotaci na této rovině srov. především [Mikulová et al., 2005].

Součástí morfologické anotace Pražského závislostního korpusu je základní klasifikace vlastních jmen. Při morfologické anotaci v rámci PDT byla každému tokenu – tj. slovnímu tvaru, číselnému výrazu a interpunkčnímu znaménku – přiřazena morfologická značka (tag) a morfologické lema. V morfologickém tagu, který má podobu řetězce o 15 pozicích, je udána informace o slovním druhu a gramatických kategoriích analyzovaného slovního tvaru. Morfologické lema má dvě části, vlastní lema a sadu technických přípon, má tedy následující strukturu:

`vlastnílema_:P1_;P2_,P3_^(P4)`

Morfologické lema vždy obsahuje první část, tedy vlastní lema, druhá část morfologického lematu tvořená technickými příponami `P1` až `P4` je nepovinná

– součástí morfologického lematu nemusí být žádná z přípon nebo zde může vystupovat pouze některá z nich, přípona jednoho typu se navíc v lematu může vyskytnout i víckrát. Pravidla pro morfologickou anotaci PDT včetně seznamů všech možných hodnot jednotlivých pozic morfologického tagu i hodnot přípon, které mohou být součástí morfologického lematu, jsou uvedeny v technické zprávě [Hana et al., 2002], v revidované podobě potom v [Zeman et al., 2005].

Vlastní lema funguje jako jednoznačný identifikátor analyzovaného tvaru, k 'základnímu' tvaru slova (tj. infinitivu pro slovesné tvary, tvaru nominativu singuláru pro substantiva atd.) je proto v některých případech přidána číslice odlišující jednotlivé významy, př. *vazba-1* jako lema s významem 'uvěznění' a *vazba-2* s významem 'spojení'. Druhá část morfologického lematu se skládá ze čtyř přípon (`P1, P2, P3` a `P4`), z nichž každá je k vlastnímu lematu připojena specifickou kombinací znaků: přípona `P1` je uvozena znaky `_:`, přípona `P2` znaky `_;`, přípona `P3` znaky `_,` a přípona `P4` znaky `_^`. Přípona `P1` uvádí informaci o vidu slovesa nebo informaci, zda se jedná o zkratku. Hodnota přípony `P2` určuje typ vlastního jména nebo obdobných výrazů a lze ji tak chápat jako prostředek klasifikace některých pojmenovaných entit – této příponě se dále věnujeme podrobněji. Hodnota přípony `P3` určuje stylový příznak (zda daný výraz patří do hovorové češtiny, zda jde o slangový výraz nebo cizí slovo atd.). V příponě `P4` je uváděn komentář různého typu, např. derivační informace nebo informace k rozlišení jednotlivých významů daného slova. Jako příklad morfologického lematu obsahujícího všechny čtyři popsané přípony (příponu typu P2 dvakrát) zde uvádíme lema odpovídající zkratce slovenské politické strany *HZDS*, jak se objevuje v anotaci na morfologické rovině PDT 2.0 (struktura tohoto lematu je popsána v tabulce 3.1):

`HZDS-1_:B_;K_;p_,t_^(Hnutie_za_demokratické_Slovensko)`

`HZDS-1`	vlastní lema	
`_:B`	přípona `P1` (vid / zkratka)	`B`: zkratka
`_;K`	přípona `P2` (typ vlastního jména)	`K`: společnost, organizace, instituce
`_;p`	přípona `P2` (typ vlastního jména)	`p`: politika, vláda, armáda
`_,t`	přípona `P3` (stylový příznak)	`t`: cizí slovo
`_^(Hnutie_za_demokratické_Slovensko)`	přípona `P4` (komentář)	vysvětlení zkratky

Tabulka 3.1: Struktura morfologického lematu odpovídajícího zkratce *HZDS* na morfologické rovině PDT 2.0. V prvním sloupci jsou uvedeny jednotlivé části lematu; ve druhém sloupci, jakou funkci tyto části plní; ve třetím sloupci je pak vysvětlen význam hodnot přípon, které se v lematu vyskytují.

Přípona `P2` je přiřazována morfologickým lematům vlastních jmen a dalších

Hodnota	Popis	Příklad
G	geografické jméno	*Praha, Ústí nad Labem*
Y	křestní jméno, dříve jako defaultní hodnota	*Petr, John*
S	příjmení	*Dvořák, Zelený, Agassi, Bush*
E	označení příslušníka národa	*Čech, Kolumbijec*
R	název produktu	*Tatra* (auto)
K	název společnosti	*Tatra* (společnost)
m	defaultní hodnota – ostatní vlastní jména; také pro značkování funkčních slov vystupujících jako součást vlastních jmen	

Tabulka 3.2: Původní sada hodnot definovaných pro příponu lematu, která určuje typ vlastního jména. Příklady uvedené ve třetím sloupci jsou převzaty z manuálu [Hana et al., 2002].

výrazů, které by bylo možno chápat jako pojmenované entity. V první verzi Manuálu pro morfologickou anotaci ([Hana et al., 2002]) bylo definováno sedm hodnot této přípony. Šest z nich odpovídalo jednotlivým typům vlastních jmen: rozlišena jsou geografická jména (hodnota `G`), křestní jména (`Y`), příjmení (`S`), označení příslušníka národa (`E`), názvy produktů (`R`) a názvy společností (`K`); sedmá hodnota (`m`) je přiřazována ostatním vlastním jménům (viz tab. 3.2).

V revidované verzi tohoto manuálu ([Zeman et al., 2005]) byla uvedená sada rozšířena o dalších třináct hodnot. Zatímco původních sedm hodnot bylo přiřazováno pouze vlastním jménům, nové hodnoty pokrývají i výrazy, které za vlastní jména považována nejsou (o těchto výrazech můžeme souhrnně hovořit jako o pojmenovaných entitách a definovanou sadu hodnot považovat za jejich klasifikaci).

Rozšířená sada hodnot (viz tab. 3.3) byla vedle morfologické anotace Pražského závislostního korpusu 2.0 ([Hajič et al., 2006]) použita také při značkování Českého akademického korpusu 1.0 ([Hladká et al., 2007]) a korpusu SYN2000b, který se od korpusu SYN2000 liší právě podobou morfologických lemat.[1]

Rozhodnutí zachycovat typ vlastního jména (nebo šíře: typ pojmenované entity) v rámci anotace na morfologické rovině se může jevit jako problematické. Pojmenované entity totiž často tvoří velmi rozsáhlé a složité struktury (srov. např. *Nové Mesto nad Váhom*, *Filozofická fakulta Masarykovy univerzity v Brně*) a na morfologické rovině, kde je každý token analyzován samostatně, tedy nemohou být zachyceny adekvátně. Např. název *Nové Mesto nad Váhom* je v korpusu SYN2000b lematizován následovně:

```
nový Mesto_;G nad-1 Váh_;G
```

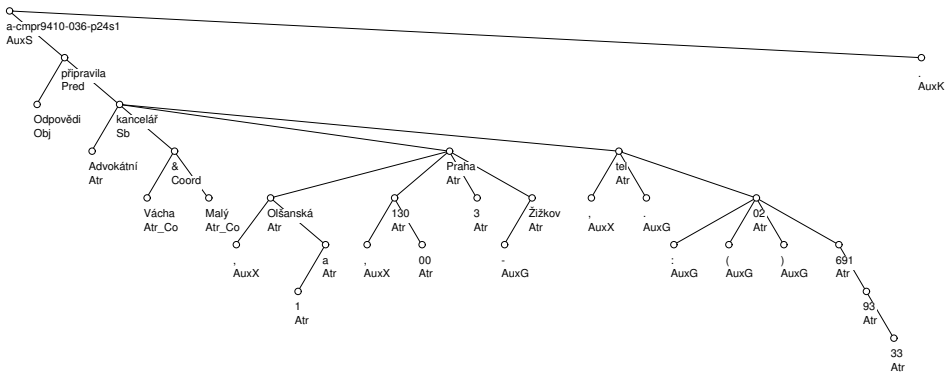
[1] Přípony lemat nejsou uvedeny v korpusu SYN2000 ani v dalších korpusech vytvořených v Ústavu Českého národního korpusu – morfologická lemata zde mají podobu základních slovních tvarů.

Hodnota	Popis	Příklad
Y	křestní jméno, (dříve jako defaultní hodnota)	*Petr, John*
S	příjmení	*Dvořák, Zelený, Agassi, Bush*
E	příslušník národa, obyvatel území	*Čech, Kolumbijec, Newyorčan*
G	geografické jméno	*Praha, Tatry* (hory)
K	společnost, organizace, instituce	*Tatra* (společnost)
R	produkt	*Tatra* (auto)
m	ostatní vlastní jména	
H	chemie	*sarin*
U	lékařství	*HIV*
L	přírodní vědy	*všivec*
j	právo	*disoluce*
g	technologie obecně	*flexografie, elektronvolt*
c	výpočetní technika a elektronika	*link, dvojpól*
y	koníčky, volný čas, cestování	*CKM* (Cestovní kancelář mládeže)
b	hospodářství, finance	*dolar*
u	kultura, vzdělávání, umění, ostatní vědy	*naturfilozofie, tritón*
w	sport	*Wimbledon*
p	politika, vláda, armáda	*ČSSD*
z	ekologie, životní prostředí	*pelagiál*
o	barvy	*bíločervený*

Tabulka 3.3: Rozšířená sada hodnot, jichž může nabývat přípona lematu určující typ vlastního jména ([Zeman et al., 2005]). Tato sada je použita v morfologické anotaci PDT 2.0, v ČAK a SYN2000b. Příklady uvedené ve třetím sloupci tabulky byly u prvních šesti hodnot převztaty z manuálu [Hana et al., 2002], příklady v ostatních řádcích pocházejí z PDT 2.0 a z korpusu SYN2000b.

Skutečnost, že tato slova tvoří dohromady název, není na této rovině zachycena. Anotace na této rovině rovněž neumožňuje popsat případnou strukturu složitých názvů (viz uvedený název fakulty, v němž vystupuje také název univerzity a název města). Výhrady je samozřejmě možné mít rovněž k samotné klasifikaci, s níž se pracovalo (např. počítá se s vyznačováním označení barev, ale nikoli jiných vlastností objektů, např. tvarů), ani v korpusu ovšem není těchto značek užíváno důsledně (např. v PDT 2.0 se přípona `;ـه` objevuje v lematech slov *červený* nebo *fialový*, ale nikoli třeba *modrý*). Je ovšem nutno připomenout, že při tvorbě ani jednoho z korpusů, kde byla tato klasifikace uplatněna, nebylo zpracování pojmenovaných entit ústředním úkolem, a této problematice tedy nebyla věnována soustavnější pozornost.

Pražský závislostní korpus 2.0 obsahuje kromě anotace na morfologické rovině, jejíž součástí je uvedená klasifikace, také anotaci na dvou syntaktických



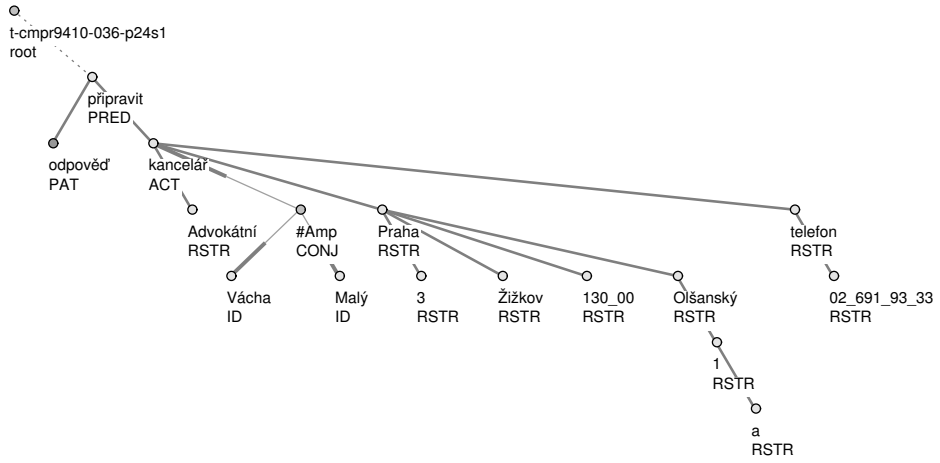
Obrázek 3.1: Závislostní strom reprezentující větu *Odpovědi připravila Advokátní kancelář Vácha & Malý, Olšanská 1a, 130 00 Praha 3 - Žižkov, tel.: (02) 691 93 33* na analytické rovině PDT 2.0.

rovinách: na analytické rovině, kde je zachycena povrchová syntaktická struktura dané věty (tj. které slovo je podmětem, předmětem atd.; tato informace se zachycuje v atributu `afun`), a na rovině tektogramatické, kde jsou určeny hloubkově syntaktické funkce jednotlivých částí věty (zda jde o aktor, patiens atd.; tato funkce se zachycuje v atributu `functor`). Na obou těchto rovinách byly výrazy vystupující v textu jako pojmenované entity analyzovány jako běžné součásti věty: byly zapojeny do závislostní stromové struktury a byla určena jejich funkce, kterou v této struktuře plní. Vztahy mezi jednotlivými částmi složitějších pojmenovaných entit (např. poštovních adres obsahujících jméno osoby, název instituce, ulici apod.) byly reprezentovány stejně jako vztahy závislostní, ačkoli zde o závislost v lingvistickém slova smyslu nejde (srov. např. vztah mezi číslem popisným a názvem ulice zachycovaný na tektogramatické rovině jako vztah přívlastkový). Reprezentace věty *Odpovědi připravila Advokátní kancelář Vácha & Malý, Olšanská 1a, 130 00 Praha 3 - Žižkov, tel.: (02) 691 93 33* na analytické rovině uvádíme na obr. 3.1, závislostní strom odpovídající této větě na tektogramatické rovině je uveden na obr. 3.2.

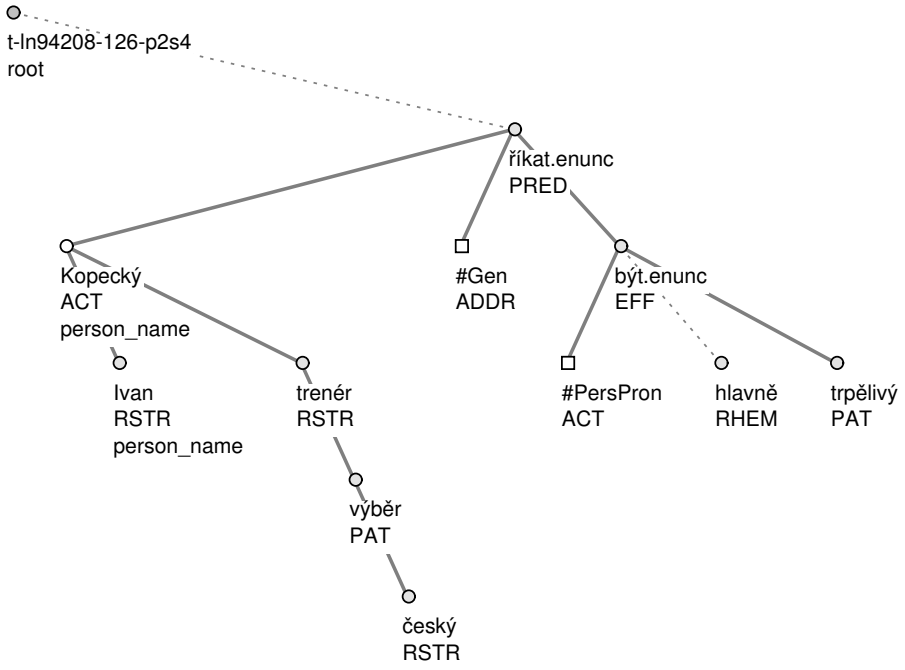
Na tektogramatické rovině byl navíc zaveden atribut `is_name_of_person`, který nabývá hodnoty `1` u jmen osob, hodnoty `0` pak jinde – srov. obr. 3.3. Hodnoty tohoto atributu byly určovány na základě seznamu osobních jmen (srov. také následující sekci 3.2), problematické případy pak byly rozhodnuty anotátorem.

3.2 Další zdroje pro identifikaci pojmenovaných entit v českých textech

Křestní jména i příjmení lidí, názvy obcí, jejich částí apod. lze v textu identifikovat také na základě seznamů. Pro češtinu jsou k některým typům vlastních jmen k dispozici seznamy v elektronické podobě. Tyto seznamy většinou nebyly sestaveny k (počítačově) lingvistickým účelům (jsou přístupné např. na stránkách Ministerstva vnitra České republiky nebo Českého statistického úřadu), ale pro zpracování pojmenovaných entit by je bylo možné využít. Nyní uve-



Obrázek 3.2: Závislostní strom reprezentující větu *Odpovědi připravila Advokátní kancelář Vácha & Malý, Olšanská 1a, 130 00 Praha 3 - Žižkov, tel.: (02) 691 93 33* na tektogramatické rovině PDT 2.0.



Obrázek 3.3: Závislostní strom odpovídající větě *Musíme být hlavně trpěliví, říká trenér českého výběru Ivan Kopecký* na tektogramatické rovině PDT 2.0. U uzlů reprezentujících osobní jména *Ivan* a *Kopecký* byla v atributu `is_name_of_person` vyplněna hodnota `1` – ve stromu je zobrazena jako `person_name`.

deme internetové adresy, kde jsou seznamy vlastních jmen jednotlivých typů dostupné.[2]

Seznamy křestních jmen a příjmení jsou k dispozici na internetových stránkách Ministerstva vnitra České republiky:

- 1764 mužských křestních jmen:
  `http://www.mvcr.cz/statistiky/jmena/muzi/index.html`
- 2373 ženských křestních jmen:
  `http://www.mvcr.cz/statistiky/jmena/zeny/index.html`
- 61 589 mužských podob příjmení:
  `http://www.mvcr.cz/statistiky/jmena/index.html`
- 60 417 ženských podob příjmení:
  `http://www.mvcr.cz/statistiky/jmena/index2.html`

Seznam názvů obcí je k dispozici např. na internetových stránkách Českého statistického úřadu: `http://www2.czso.cz/csu/edicniplan.nsf/p/1302-04` nebo na adrese `http://www.mestaobce.cz/`. Český statistický úřad vydal naposledy v roce 2005 *Statistický lexikon obcí* obsahující seznam obcí a jejich částí, tato publikace byla vydána v knižní podobě i na CD-ROM ([ČSÚ, 2005]).

Seznamy jmen ulic a ostatních městských veřejných prostranství jsou na internetu k dispozici pouze pro některá města – seznam odkazů uvádíme podle abecedního pořadí příslušných měst:

- Benešov:
  `http://www.benesov-city.cz/mapa/benesov.html`
- Blatná:
  `http://amber.feld.cvut.cz/user/sika/html/blatna/planek.htm`
- Brno:
  `http://www.brno.cz/toCP1250/download/ovv/ulice/index.htm`
- Doksy:
  `http://www.doksynet.wz.cz/doksy.htm#ulice`
- Chrudim:
  `http://www.chrudim.info/turistika/mapy/chrudim/rejstrik.php3`
- Jablonné v Podještědí:
  `http://www.inso.cz/adresar/mapy/jablonne/html/seznam_ulic.html`
- Jablunkov:
  `http://jablunkov.cz/mapa/cisla/cisla_mapa.html`
- Jičín:
  `http://jicin.tpc.cz/reg/regst.htm`

[2]Autorem seznamu internetových odkazů, které zde uvádíme, a informací o databázích atd. je PhDr. Pavel Štěpán z Ústavu pro jazyk český Akademie věd České republiky.

- Jihlava:
  http://www.mestaobce.cz/scripts/vismo/_obce/dokumenty2.asp?u=5967&id_org=5967&id=68902
- Jirkov:
  http://www.jirkov.cz/map.php
- Kaplice:
  http://www.ckrumlov.cz/cz1250/atlas/t_mapkap.htm
- Kladno:
  http://www.mestokladno.cz/seznam_ulic.asp
- Klatovy:
  http://www.retour.cz/mesta/klatovy/ulice.htm
- Kunovice:
  http://www.mesto-kunovice.cz/web/kunovice/Cz/seznam-ulic
- Mělník:
  http://www.melnik.cz/mapa/melnik.html
- Mladá Boleslav:
  http://www.mb-net.cz/plan_mesta/mboleslav.html
- Olomouc:
  http://www.olomouc.com/mapa/map_indx.html
- Ostrava - Poruba:
  http://www.moporuba.cz/poruba/ulice.htm
- Praha:
  http://vmp.bluehole.cz/Index.htm
- Rychnov nad Kněžnou:
  http://rychnov-city.cz/?id=m_mapa
- Sušice:
  http://www.retour.cz/mesta/susice/ulice.htm
- Teplice:
  http://old.teplice.cz/mapa/default.php?m=ulice&kvadrant=
- Úvaly:
  http://www.uvaly.cz/mapa/ulice.htm
- Valašské Meziříčí:
  http://www.mestovalmez.cz/data.php?adr=mapa&cislo=1
- Vimperk:
  http://www.retour.cz/mesta/vimperk/ulice.htm
- Zdice:
  http://www.mesto-zdice.cz/mapa/index.php?x=04&y=03

- Žďár nad Sázavou:
  `http://e-zdar.wz.cz/rej_ulic.htm`

V případě tzv. pomístních jmen z území Čech (tj. názvů neosídlených míst, především polí, luk, lesů, hor, vod a cest) je k dispozici počítačová databáze s více než 425 000 záznamy, která byla vytvořena v oddělení onomastiky Ústavu pro jazyk český Akademie věd České republliky v letech 1994 až 2004 na základě abecedního lístkového katalogu pomístních jmen v Čechách (tento katalog vznikl excerpcí soupisů pomístních jmen pro jednotlivé obce, shromážděných s pomocí dobrovolných místních spolupracovníků v letech 1963 až 1980).

Co se týká dalších typů pojmenovaných entit, seznam názvů firem je k dispozici např. na internetové adrese `http://adresarfirem.cz/`. Dílčí databáze názvů výrobků se nachází např. na adrese `http://www.ekoznacka.cz/firmy.asp`.

Založit zpracování pojmenovaných entit pouze na seznamech však není možné. Například na základě seznamu příjmení mohou být jako příjmení identifikována i obecná jména psaná s velkým počátečním písmenem proto, že stojí na začátku věty (*pan Dlouhý – Dlouhý kometář...* apod.). Identifikace výrazů zapsaných číslicemi a vystupujících v textu jako pojmenované entity (např. číslo telefonu) pak není na základě seznamu možná vůbec.

Kapitola 4

Klasifikace a anotace pojmenovaných entit navržená pro češtinu

Jedním z cílů projektu *Integrace jazykových zdrojů za účelem extrakce informací z přirozených textů* je automatické zpracování pojmenovaných entit v českých textech. Pro vývoj softwarových nástrojů je třeba mít k dispozici označkovaná data.

V první fázi projektu tedy byla navržena klasifikace pojmenovaných entit (viz sekce 4.1) a tato klasifikace byla použita při ruční anotaci několika tisíc vět (viz sekce 4.2). Kvantitativní údaje týkající se těchto vět jsou uvedeny v sekci 4.3. Experimenty s automatickým zpracováním pojmenovaných entit v českých textech, k nimž jsou označkovaná data používána, jsou podrobněji popsány v kapitole 5.

4.1 Klasifikace pojmenovaných entit

Klasifikace navržená v rámci projektu *Integrace jazykových zdrojů za účelem extrakce informací z přirozených textů* zahrnuje několik desítek *typů* pojmenovaných entit a řadí se tak mezi podrobnější klasifikace (srov. koncepce v sekci 2.4; např. oproti původní koncepci rozlišující jen několik typů, viz sekce 2.3). *Typu* pojmenované entity odpovídá v naší koncepci značka sestávající ze dvou znaků (zpravidla dvou malých písmen, např. `ps`, součástí klasifikace jsou ovšem i značky jako `p_`) – seznam značek typů uvádíme v tabulce 4.1, její rozšířenou verzi pak v tab. 4.2.

Vedle typu pojmenované entity v naší koncepci pracujeme také s pojmy *rozsah* pojmenované entity, *kontejner*, *nadtyp* a *instance*. *Rozsahem* pojmenované entity rozumíme posloupnost tokenů, které danou entitu tvoří – začátek entity se vyznačuje závorkou `<`, její konec závorkou `>`. *Kontejner* vždy zahrnuje několik pojmenovaných entit (popř. i další slova), kontejneru odpovídá značka tvořená jedním velkým písmenem (např. `P`); rozsah kontejneru se vymezuje stejně, jako je tomu u pojmenované enitity (`<...>`) – seznam značek kontejnerů uvádíme v tabulce 4.3. Kromě toho, že se pojmenované entity 'vnořují' do kontejnerů (př. `<P<pf Pavel> <ps Drda>>`), může se pojmenovaná entita v našem po-

jetí ‘vnořit’ i do jiné pojmenované entity (př. `<ic Právnická fakulta <ic<ps Masarykovy> univerzity>>`). *Nadtypu* pojmenované entity neodpovídá vlastní značka – budeme-li hovořit o pojmenovaných entitách určitého nadtypu, budou tím míněny pojmenované entity těch typů, jejichž značky začínají stejným písmenem: např. pojmenované entity `<pf Pavel>` a `<ps Drda>` jsou pojmenované entity nadtypu `p` (tedy jména osob).
Kromě toho bylo zavedeno ještě několik značek pro ošetření zvláštních případů: jednopísmenné značky `f` a `s` pro vyznačení, zda jde o cizí slovo nebo zkratku, značky `segm` a `cap` pro vyznačení toho, že se slovo s velkým počátečním písmenem vyskytuje uprostřed textu z důvodu špatné segmentace textu, popř. kvůli zdůraznění, a značky `lower` a `upper` pro vyznačení chyb v psaní velkých a malých písmen. Značka `?` znamená, že typ příslušné pojmenované entity nelze určit; značka `!` umisťovaná v závorkách před začátek věty (tedy `<!>`) říká, že věta je defektní (např. neúplná) a nebyla anotována. Tyto značky jsou uvedeny v tabulce 4.4.

Pojem *instance* pojmenované entity používáme jako souhrnný pojem, chceme-li hovořit o pojmenované entitě jakéhokoli druhu (tedy ať už jí byla přidělena značka typu, značka kontejneru nebo některá ze speciálních značek): v příkladu `<P<pf Pavel> <ps Drda>>` jsou v tomto smyslu přítomny tři instance pojmenovaných entit, z toho dvě o rozsahu jednoho slova (`Pavel` a `Drda`) a jedna o rozsahu dvou slov (`Pavel Drda`); `<io<s ODS>>` obsahuje dvě jednoslovné instance apod.

I když – v souladu se zahraničními koncepcemi – považujeme za pojmenované entity kromě vlastních jmen také číselné výrazy se specifickým významem (např. telefonní čísla), v prvním kole jsme anotaci pojmenovaných entit omezili pouze na vlastní jména (k anotacím viz následující sekce 4.2). Díky tomu bylo nejen snazší sestavit klasifikaci typů, ale usnadnili jsme tím také práci anotátorům (snadněji si mohli zapamatovat, jaké výrazy mají v textu vyznačovat jako pojmenované entity, navíc méně značek se lépe pamatuje, díky čemuž je anotace rychlejší a její výsledek konzistentní).

Klasifikace, kterou jsme pro první kolo anotací sestavili, tedy zahrnuje především značky pro vlastní jména, zahrnuje však i některé další typy (např. několik typů časových údajů). Seznam značek všech typů pojmenovaných entit, které byly v prvním kole anotací použity, uvádíme v tabulce 4.1 – značka typu je uvedena vždy v prvním sloupci, ve druhém sloupci následuje vysvětlení, ve třetím sloupci je pak uveden příklad (v příkladu je vždy vyznačena pouze entita příslušného typu); značky typů jsou seskupeny podle nadtypu (tzn. značky začínající stejným písmenem jsou pohromadě); pořadí nadtypů i pořadí typů v rámci jednoho nadtypu se pak řídí podle četnosti odpovídajícího typu v datech anotovaných v prvním kole anotací (kolem 2000 vět, srov. sekci 4.2), značka nejčastějšího typu je tedy v daném nadtypu uvedena vždy jako první atd.

p	JMÉNA OSOB	
ps	příjmení	<ps Nováková>, <ps van Dyk>
pf	křestní jméno	<pf David> Uhl, <pf J.> Drda
p_	jméno osoby nespecifikovaného typu / nezařaditelné do ostatních typů	<p_ Slované>
pc	obyvatelská jména	<pc Afričan>, <pc Pražan>
pp	náboženské postavy, pohádkové a mytické postavy, personifikované vlastnosti	<pp sv. Jakub>, <pp Prozřetelnost>
pm	druhé křestní jméno	Georg <pm Friedrich> Händel
pd	titul (pouze zkratkou)	<pd Mgr.> J. Pola
pb	jména zvířat	pes <pb Fík>
g	GEOGRAFICKÉ NÁZVY	
gu	obce, hrady a zámky	<gu Praha>, <gu Opočno>
gc	státní útvary	<gc Česká republika>, <gc Svatá říše římská>
gr	menší územní jednotky	<gr Morava>, <gr Rychnovsko>, <gr Badensko-Württembersko>
gs	ulice, náměstí	<gs ul. Šikmá>, <gs nám. Míru>
gq	části obcí, pomístní názvy	<gq Smíchov>
gh	vodní útvary	<gh Vltava>, <gh Balaton>
gl	přírodní oblasti / útvary	<gl Sibiř>, <gl Apeninský poloostrov>, <gl Polabí>
gt	kontinenty	<gt Jižní Amerika>
g_	geografický název nespecifikovaného typu / nezařaditelný do ostatních typů	Učkuduk v <g_ Mojunkumech>
gp	planety, vesmírné útvary	<gp Země>
i	NÁZVY INSTITUCÍ	
ic	kulturní, vzdělávací a vědecké instituce, sportovní kluby,...	kino <ic Dlabačov>, hokejisté <ic Sparty>
if	firmy, koncerny, hotely,...	<if Unipetrol>, řetězec <if Edeka>
io	státní a mezinárodní instituce, politické strany a hnutí, náboženské skupiny	<io Evropská komise>, <io Policie>, <io ODS>
ia	přednášky, konference, soutěže,...	<ia Stanley Cup>
i_	instituce nespecifikovaného typu / nezařaditelná do ostatních typů	<i_ Studijní odd.>

Tabulka 4.1: Klasifikace pojmenovaných entit pro první kolo anotací

o	NÁZVY VĚCÍ	
oa	kulturní artefakty (knihy, filmy stavby,...)	Hugovi <oa Bídníci>
op	výrobky	mikroprocesory <op Intel Pentium>
om	měny (zapsané zkratkou, symbolem)	1000 <om Kč>, 30 <om $>
oe	měrné jednotky (zapsané zkratkou)	200 <om MHz>
o_	názvy nespecifikovaného typu / nezařaditelné do ostatních typů	čeleď <o_ Rubiaceae>, <o_ HIV>
or	předpisy, normy,..., jejich sbírky	ve <or Sbírce zákonů>
oc	názvy chemikálií, chemické vzorce	plyny <oc Ar> a <oc He>
t	ČASOVÉ ÚDAJE	
th	hodina	v <th 17> hodin
ty	rok	od roku <ty 1555>
tm	měsíc	1. <tm října> 2005
td	den	<td 1.> října 2005
ti	časový interval	<ti 24. až 30. března>
tf	svátky a významné dny	<tf Velikonoce>
m	NÁZVY MÉDIÍ	
mn	periodika, redakce, tiskové agentury	agentura <mn Interfax>
mt	televizní stanice	<mt ČT 2>
mr	rozhlasové stanice	<mr Frekvence 1>
a	ČÍSLA JAKO SOUČÁSTI ADRES	
ah	číslo popisné	ul. Šikmá <ah 356>
at	telefon, fax	tel. <at 256 458 588>
az	PSČ	<az 180 00> Praha 8

Tabulka 4.1: Klasifikace pojmenovaných entit pro první kolo anotací

Pro druhé kolo anotací, v němž jsme se zaměřili hlavně na číselné výrazy, byly v sadě značek představené v tab. 4.1 provedeny následující změny:

- do nadtypu časových údajů (`t`) byly přidány typy `tc` (století), `tn` (minuta), `tp` (období) a `ts` (sekunda);
- v nadtypu časových údajů byl zrušen typ `ti` (interval);
- do nadtypu `m` (názvy médií) byl přidán typ `mi` (internetový odkaz);
- byl zaveden nový nadtyp `c` (součásti bibliografických údajů), v jehož rámci se rozlišují typy `cb` (číslo dílu,...), `cn` (číslo kapitoly,...), `cp` (číslo strany), `cr` (číslo zákona,...) a `cs` (název článku,...);
- byl zaveden nový nadtyp `n` (čísla se specifickým významem) a v jeho rámci jsou rozlišovány typy `na` (věk), `nc` (skóre), `ni` (itemizátor), `nm` (vzorec), `np` (číslo jako součást jména osoby), `nq` (označení čtvrti), `nr` (poměr), `nw` (velikost bytu) a `n_` (pro jiná čísla se specifickým významem);
- byl zaveden nový nadtyp `q` (čísla s významem počtu a pořadí) a v něm rozlišeny typy `qc` (číslo s významem počtu) a `qo` (číslo s významem pořadí).

Rozšířenou sadu značek všech typů uvádíme v tabulce 4.2; značky jednotlivých typů jsou v této tabulce – stejně jako v tabulce 4.1 – seskupeny podle nadtypu, nadtypy i typy v rámci jednotlivých nadtypů jsou ovšem seřazeny abecedně; v příkladu uvedeném ve třetím sloupci je vždy vyznačena pouze pojmenovaná entita příslušného typu. Značky, které jsou v této sadě navíc oproti sadě původní, jsou zvýrazněny tučně.

Je však velmi pravděpodobné, že ani tato rozšířená sada není definitivní.

a	ČÍSLA JAKO SOUČÁSTI ADRES	
ah	číslo popisné	ul. Šikmá <ah 356>
at	telefon, fax	tel. <at 256 458 588>
az	PSČ	<az 180 00> Praha 8
c	SOUČÁSTI BIBLIOGRAFICKÝCH ÚDAJŮ	
cb	**číslo dílu,...**	Mluvnice češtiny <cb 2>
cn	**číslo kapitoly, obrázku,...**	viz obr. <cn 15>
cp	**číslo strany**	na straně <cp 3>
cr	**číslo zákona, nařízení,...**	zákon č. <cr 586>
cs	**název článku,...**	v článku <cs On pronouns>
g	GEOGRAFICKÉ NÁZVY	
gc	státní útvary	<gc Česká republika>, <gc Svatá říše římská>
gh	vodní útvary	<gh Vltava>, <gh Balaton>
gl	přírodní oblasti / útvary	<gl Sibiř>, <gl Apeninský poloostrov>, <gl Polabí>
gp	planety, vesmírné útvary	<gp Země>
gq	části obcí, pomístní názvy	<gq Smíchov>
gr	menší územní jednotky	<gr Morava>, <gr Rychnovsko>, <gr Badensko-Württembersko>
gs	ulice, náměstí	<gs ul. Šikmá>, <gs nám. Míru>
gt	kontinenty	<gt Jižní Amerika>
gu	obce, hrady a zámky	<gu Praha>, <gu Opočno>
g_	geografický název nespecifikovaného typu / nezařaditelný do ostatních typů	Učkuduk v <g_ Mojunkumech>
i	NÁZVY INSTITUCÍ	
ia	přednášky, konference, soutěže,...	<ia Stanley Cup>
ic	kulturní, vzdělávací a vědecké instituce, sportovní kluby,...	kino <ic Dlabačov>, hokejisté <ic Sparty>
if	firmy, koncerny, hotely,...	<if Unipetrol>, řetězec <if Edeka>
io	státní a mezinárodní instituce, politické strany a hnutí, náboženské skupiny	<io Evropská komise>, <io Policie>, <io ODS>
i_	instituce nespecifikovaného typu / nezařaditelné do ostatních typů	<i_ Studijní odd.>

Tabulka 4.2: Rozšířená klasifikace pojmenovaných entit

m	NÁZVY MÉDIÍ	
mi	**internetové odkazy**	<mi www.cinestar.cz>
mn	periodika, redakce, tiskové agentury	agentura <mn Interfax>
mr	rozhlasové stanice	<mr Frekvence 1>
mt	televizní stanice	<mt ČT 2>
n	ČÍSLA SE SPECIFICKÝM VÝZNAMEM	
na	**věk**	je mu <na 18>
nc	**skóre**	vyhráli <nc 5:0>
ni	**itemizátor**	<ni 1)> DPH, <ni 2)> daň z příjmu
nm	**vzorec**	<nm q = 3m + p>
np	**číslo jako součást jména osoby**	Karel <np IV.>
nq	**označení čtvrti**	Praha <nq 8>
nr	**poměr**	v poměru <nr 2:3>
nw	**velikost bytu**	byty <nw 3+1>
n_	**číslo se specifickým významem, jehož typ nebyl vyčleněn jako samostatný / nelze identifikovat**	autobusová linka <n_ 158>
o	NÁZVY VĚCÍ	
oa	kulturní artefakty (knihy, filmy stavby,...)	Hugovi <oa Bídníci>
oc	názvy chemikálií, chemické vzorce	plyny <oc Ar> a <oc He>
oe	měrné jednotky (zapsané zkratkou)	200 <om MHz>
om	měny (zapsané zkratkou, symbolem)	1000 <om Kč>, 30 <om $>
op	výrobky	mikroprocesory <op Intel Pentium>
or	předpisy, normy,..., jejich sbírky	ve <or Sbírce zákonů>
o_	názvy nespecifikovaného typu / nezařaditelné do ostatních typů	čeleď <o_ Rubiaceae>, <o_ HIV>
p	JMÉNA OSOB	
pb	jména zvířat	pes <pb Fík>
pc	obyvatelská jména	<pc Afričan>, <pc Pražan>
pd	titul (pouze zkratkou)	<pd Mgr.> J. Pola
pf	křestní jméno	<pf David> Uhl, <pf J.> Drda

Tabulka 4.2: Rozšířená klasifikace pojmenovaných entit

pm	druhé křestní jméno	Georg <pm Friedrich> Händel
pp	náboženské postavy, pohádkové a mytické postavy, personifikované vlastnosti	<pp sv. Jakub>, <pp Prozřetelnost>
ps	příjmení	<ps Nováková>, <ps van Dyk>
p_	jméno osoby nespecifikovaného typu / nezařaditelné do ostatních typů	<p_ Slované>
q	ČÍSLA S VÝZNAMEM POČTU A POŘADÍ	
qc	**číslo s významem počtu**	<qc 168> stran, <qc 3> %
qo	**číslo s významem pořadí**	<qo 6.> děkan, přišel jako <qo 3.>
t	ČASOVÉ ÚDAJE	
tc	**století**	<tc 20.> století
td	den	<td 1.> října 2005
tf	svátky a významné dny	<tf Velikonoce>
th	hodina	v <th 17> hodin
tm	měsíc	1. <tm října> 2005
tn	**minuta**	<tn 25> minut
tp	bf období	<tp 70.> léta, od <tp dvacátých> let
ts	**sekunda**	<ts 3> sekundy
ty	rok	od roku <ty 1555>

Tabulka 4.2: Rozšířená klasifikace pojmenovaných entit

A	adresa: <A<if KOMO>, <gs Knížecí> <ah 12/173>, <az 709 00> <gu Ostrava-Nová Ves>, tel.: <at 069 6621773>, <at 6621375>, <at 601527588>, fax: <at 069 6621773>>
C	bibliografický údaj: <C<P<pf G.> <ps Lukács>>: <oa<f Die Theorie des Romans>>, <gu Berlin>, <ic<f Verlag <P<pf Paul> <ps Cassirer>>>> <ty 1920>>
P	jméno osoby: <P<pd Doc.> <pd MUDr.> <pf Přemysl> <ps Doberský>, <pd DrSc.>>
T	časový údaj: <T<td 21.> <tm června> <ty 2003> <th 20.00>>

Tabulka 4.3: Značky kontejnerů užívaných v prvním i ve druhém kole anotací. Ve druhém sloupci jsou uvedené značky vysvětleny a doloženy příkladem. Příklady jsou značkovány tak, jak se vyskytují v anotovaných datech.

s	zkratka	je členem <io<s ODS>>
f	slovo z cizího jazyka	<if<f Deutsche Bank>>
segm	slovo napsáno velkým písmenem např. v důsledku chybné segmentace textu (zde název článku a jeho první věta spojeny v korpusu do jedné věty)	Revoluční zvrat v pohledu na život <segm Základem> života není...
cap	slovo napsáno velkými písmeny např. z typografických důvodů	A jak <cap TO> vysvětlíte?
lower	slovo napsáno chybně s velkým písmenem	ekonomicky ovládnout <lower Střední> <gt Evropu>
upper	slovo napsáno chybně s malým písmenem	dalším zajímavým <upper britem> je...
?	entita nespecifikovaného typu / nezařaditelná do ostatních typů	<? Asmara> se odmítá stáhnout z území...
!	věta neanotována	<!> 70: 15 Písní na S. Georga, op.

Tabulka 4.4: Ostatní značky použité v anotaci. Ve druhém sloupci je uvedeno vysvětlení, ve třetím sloupci příklad.

4.2 Anotace trénovacích a testovacích dat

4.2.1 První kolo anotací

Anotace, která probíhala na konci roku 2005, se zaměřovala na identifikaci a klasifikaci vlastních jmen v českých textech. Značkami uvedenými v tab. 4.1 bylo anotováno 2000 vět. Tyto věty byly náhodně vybrány ze souboru 5.364.071 vět, které byly v korpusu SYN2000 nalezeny na základě dotazu `([word=".*[a-z0-9]"] [word="[A-Z].*"])`, tzn. byly hledány dvojice slov, z nichž první končí jakým-

koli malým písmenem nebo číslicí a druhé začíná velkým písmenem (tj. chtěli jsme nalézt slova začínající velkým písmenem, která ovšem nestojí na začátku věty).

Anotaci prováděly paralelně dvě anotátorky. Jejich úkolem bylo vymezit rozsah pojmenované entity (tzn. umístit do textu levou závorku `<` vyznačující začátek entity a pravou závorku `>` vyznačující její konec) a určit typ pojmenované entity (tzn. za levou závorku uvést dvoumístnou značku z uvedené sady), případně vymezit rozsah kontejneru a určit jeho druh nebo do textu umístit některou ze speciálních značek (např. `segm`).
Po odevzdání všech anotací byly anotace obou anotátorek porovnány. V místech, kde se anotace lišily, byla v dalším průchodu zvolena konečná anotace. Při tomto průchodu také byly vymazány věty, které nebyly anotovány (opatřené značkou <!>). Z původních 2000 anotovaných vět tak zbylo 1923 vět. Do této sady bylo následně přidáno dalších 87 anotovaných vět (tentokrát jsme anotovali 100 vět náhodně vybraných z celkem 4.204.857 vět, které byly na základě výše uvedeného dotazu nalezeny v korpusu SYN2005; po odstranění 13 vět, které byly opatřeny značkou <!>, jsme získali uvedených 87 vět). Konečná sada tedy obsahuje 2010 vět. Tyto věty se skládají celkem z 51921 slovních jednotek (tj. slovních tvarů, číselných výrazů a interpunkčních znamének). Při anotaci v nich bylo identifikováno celkem 11644 pojmenovaných entit. Ukázka anotovaného textu je uvedena na str. 39.

Sada anotovanych vět byla rozdělena do tří částí, vznikla tak sada trénovacích dat (*train data*) a dvě sady dat testovacích (testovací data vývojová, *d-test data*, a testovací data evaluační, *e-test data*). Kvantitativní vlastnosti dat jsou popsány v sekci 4.3.

4.2.2 Druhé kolo anotací

V listopadu a prosinci 2006 proběhlo další kolo anotací. Tentokrát šlo především o anotaci číselných výrazů. V této souvislosti byla stávající sada značek rozšířena (rozšířená sada je uvedena v tab. 4.2). Anotováno bylo 2000 vět náhodně vybraných z celkem 1.356.321 vět, které byly v korpusu SYN2005 nalezeny na základě dotazu `[word=".*[0-9].*"]`, tzn. dotazu na všechny řetězce obsahující alespoň jednu číslici.

Anotaci prováděla jedna anotátorka, jejím úkolem bylo vyznačit v textu výrazy obsahující aspoň jednu číslici (tj. např. *1998*, ale také *MP3*) a určit, zda se jedná o číslo s běžným významem počtu nebo pořadí nebo zda jde o číslo, popř. 'slovo' obsahující číslo se specifickým významem (a tedy v našem pojetí o pojmenovanou entitu). Úkolem anotátorky dále bylo vymezit rozsah těch pojmenovaných entit, které byly anotovány už v prvním kole anotací (tedy hlavně vlastních jmen), a určit jejich typ. Dále anotátorka u obou zpracovávaných druhů pojmenovaných entit (tzv. entit obsahujících číslice i u vlastních jmen) při anotaci vymezila rozsah kontejneru a určila jeho druh nebo do textu umístila některou ze speciálních značek (např. `segm`). Odevzdané anotace zatím neprošly kontrolou. Ukázka anotovaného textu je uvedena na str. 40.

Ukázka textu anotovananého v prvním kole anotací s použitím značek uvedených v tab. 4.1:

72: Britský multimediální umělec <p_ Sting> , vlastním jménem <P<pf Gordon> <ps Sumner>> , který má vystoupit <T<td 14 .> <tm června>> v pražské <ic Sportovní hale> , bude s největší pravděpodobností bydlet se svým devětadvacetičlenným týmem pod krycím jménem v některém z pražských hotelů .

73: Na našem trhu se nejvýznamněji podílí <if Supraphon> s <if PolyGramem> , jehož páteř tvoří tři labely : <if<f Deutsche Grammophon>> , <if Decca> a <if Philips> .

74: V <ic Galerii <P<pf Václava> <ps Špály>>> bude dnes zahájena výstava obrazů německého umělce <P<pf Herberta> <ps Achternbusche>> , připravená ve spolupráci s pražským <ic GoetheInstitutem> .

75: Ostrov <gl u Černé <pp Matky Boží>> .
76: <ic Galerie u <ps Klicperů>> (Divadlo) : <P<pf Karel> <ps Sládek>> - obrazy .

77: Kdybychom si však mohli vybrat z týdenního programu ještě další představení , určitě bychom šli na <oa Sen noci svatojánské> londýnského divadla <ic<f Set Up Theater>> , na <ps Beckettův> <oa Konec hry> sarajevského souboru , na španělské představení <ps Müllerovy> hry <oa Kvartet> , na vystoupení souboru <ic<f Victory Sonqoba Theatre Company>> z jihoafrického <gu Johanesburgu> , na <ps Gombrowiczovu> <oa Operetu> z krakovského <ic Stareho Teatru> , na inscenaci <P<pf Dino> <ps Mustafice>> <oa Král umírá> , na francouzského <oa Vojcka> . . . a to nejsou zdaleka všechna představení a doprovodné akce festivalu .

78: Na začátku sedmdesátých let uveřejnili dva mladí básníci " nové vlny " <P<pf Adam> <ps Zagajewski>> a <P<pf Julian> <ps Kornhauser>> literární manifest " <oa Svět nezobrazený> " , požadující nový realismus , který zapůsobil také na mladé polské filmaře .

79: Kompilace <C<oa Dalekonosné husle> (<mr Český rozhlas <gu Brno>> - <if Gnosis> , <ty 1998>)> čerpá z nahrávek houslisty a primáše <P<pf Jožky> <ps Kubíka>> (<ti 1907 - 78>) , který jako první zavedl do horňácké hudby revoluční novinku - cimbál .

80: Okres <gu Los Angeles>
81: Hlavní role v černobílém širokoúhlém (!) filmu si zahráli <P<pf Václav> <ps Koubek>> , <P<pf Pavel> <ps Landovský>> , <P<pf Eliška> <ps Sirová>> , <P<pf Jana> <ps Dolanská>> , <P<pf Jiří> <ps Soukup>> (pouhá shoda jmen se scenáristou) a <P<pf Matěj> <ps Hádek>> .

82: Hrají : <P<pf Jason> <pm Scott> <ps Lee>> , <P<pf Cary> <ps Elwes>> , <P<pf Lena> <ps Headeyová>> . - ci 9 / 95 <mt NOVA>

Ukázka anotovaného textu z druhého kola anotace, při němž byly používány značky uvedené v tab. 4.2:

151: Registrace domény s koncovkou . cz byla do <T<tm srpna> <ty 1999>> zdarma , nyní se platí <qc 1600> <om Kč> za první a polovinu za každý další rok .

152: Například rodinná vstupenka určená pro dva dospělé a alespoň jedno dítě pro nejdelší prohlídkový okruh <o_ A> stojí <qc 300> korun .

153: <gu OSTRAVA> (tch) - Už dříve trestaný lotr (<na 21>) bezdůvodně napadl o rok staršího mladíka před restaurací <if Karlos> v <gu Ostravě>-<gq Zábřehu> .

154: Francouzská lyžařka (nar . <gu Saint-Maxime> , <g_ Var> , <ty 1945>) .
155: Neodporují si navíc časové údaje , když srovnáme posouvání u <gl Islandu> a u <gl ostrova <P<pf Jana> <ps Mayena>>> směrem k <gl Sibiři> (obr . <cn 58>) ?

156: Pro přepravu hromadných substrátů je zapracováno <qc 111> pravidelných tras odesilatelských vlaků , z toho pro přepravu uhlí <qc 65> .

157: / <ni 7>/ Sémickou analýzu je třeba doplnit , testovat a verifikovat analýzou oblasti užití (v . dále) , v níž se sémém manifestuje .

158: Primáš cikánské kapely , jistě že <pf Lájoš> jest jméno jeho , <na 50> let , umí primášovat

159: Každý rok jich na následky této chudoby <qc 18 miliónu> zemře (<qc 50 tisíc> denně) , z toho <qc 12 miliónů> dětí do věku pěti let (více než <qc 30 tisíc> denně) .

160: Od počátku roku <ty 2003> bylo rovněž zahájeno osm nových řízení .
161: Je logické , že dokud ji mít nebude , nepotečou do <if TV<n_ 3>> žádné větší investice , protože jejich návratnost je nejistá .

162: Podle <io<s MMF> a <io G<n_ 7>> vyvolaly četné krize ve světě , především ale platební neschopnost <gc Argentiny> , nutnost vytvořit konkurzní zákon i pro státy .

163: Banka rovněž napříště nakoupí každý měsíc za <qc 1,2 bil .>
164: Broskve ponoříme na <qc 1> minutu do vroucí vody , pak je osušíme .
165: <P<pd Ing .> <pf Karel> <ps Hennhofer> , <pd PhD>> , zastupující ostravskou divizi <ic Technické inspekce " DOM-ZO <n_ 13> "> popsal nové pojetí systému managementu jakosti podle normy <or ČSN EN ISO <nr 9001:2001>> .

4.3 Kvantitativní vlastnosti anotovaných dat

Soubor dat, který byl označkován v prvním kole anotací, obsahuje celkem 2010 vět, což činí 51921 slovních jednotek. Věty byly rozděleny do tří sad v poměru 8:1:1. 80 % všech dat slouží jako data trénovací (1608 vět, 41710 slovních jednotek), 10 % dat jako vývojová testovací data (*d-test data*; 201 vět, 5296 slovních jednotek), 10 % dat jako evaluační testovací data (*e-test data*; 201 vět, 4915 slovních jednotek). Ve všech větách bylo identifikováno 11644 instancí pojmenovaných entit, tzn. do vět bylo umístěno celkem 11644 značek typů (viz tab. 4.1), značek kontejnerů (viz tab. 4.3) a ostatních značek (viz tab. 4.4). Z toho 9263 instancí je součástí trénovacích dat. Podrobnější kvantitativní charakteristiky už budeme uvádět pouze pro trénovací data.

Z 9263 instancí pojmenovaných entit obsažených v trénovacích datech má 6109 instancí rozsah jednoho tokenu, 3154 entit se skládá z více tokenů – podrobnější údaje viz tabulka 4.5. Seznam všech značek setříděný podle počtu jejich výskytů v trénovacích datech je uveden v tab.4.6; tentýž seznam setříděný podle abecedního pořadí značek je uveden v tab. 4.8. 1065 značek z nich jsou značky kontejnerů – seznam značek kontejnerů setříděný podle počtu výskytů je uveden v tab. 4.7, tentýž seznam setříděný podle abecedy viz tab. 4.9.

Délka instance	Počet výskytů
1	6109
2	2018
3	619
4	243
5	104
6	64
7	34
8	31
9	9
10	9
13	8
11	7
15	3
16	1
21	1
23	1
28	1
34	1

Tabulka 4.5: Rozložení instancí vzhledem k jejich délce (počet slovních jednotek) v sadě trénovacích dat

Značka instance	Počet výskytů	%	Značka instance	Počet výskytů	%
ps	1380	14,90 %	gq	48	0,52 %
pf	1087	11,73 %	gh	42	0,45 %
P	945	10,20 %	om	40	0,43 %
gu	754	8,14 %	ti	38	0,41 %
oa	664	7,17 %	oe	37	0,40 %
th	505	5,45 %	pm	36	0,39 %
s	406	4,38 %	gl	34	0,37 %
ic	392	4,23 %	pd	30	0,32 %
gc	372	4,02 %	mt	28	0,30 %
io	260	2,81 %	gt	27	0,29 %
segm	256	2,76 %	ah	22	0,24 %
if	253	2,73 %	or	21	0,23 %
f	223	2,41 %	A	20	0,22 %
p_	147	1,59 %	mr	19	0,21 %
ty	141	1,52 %	oc	14	0,15 %
op	124	1,34 %	o_	14	0,15 %
tm	99	1,07 %	at	8	0,09 %
td	94	1,01 %	pb	7	0,08 %
T	93	1,00 %	lower	7	0,08 %
ia	92	0,99 %	gp	7	0,08 %
mn	79	0,85 %	g_	7	0,08 %
pc	74	0,80 %	C	7	0,08 %
gr	69	0,74 %	tf	6	0,06 %
gs	62	0,67 %	upper	4	0,04 %
cap	61	0,66 %	i_	4	0,04 %
?	51	0,55 %	az	4	0,04 %
pp	49	0,53 %			

Tabulka 4.6: Všechny značky vyskytující se v trénovacích datech setříděné podle počtu výskytů

Značka kontejneru	Počet výskytů	%
P	945	88,73 %
T	93	8,73 %
A	20	1,88 %
C	7	0,66 %

Tabulka 4.7: Značky kontejnerů vyskytující se v trénovacích datech setříděné podle počtu výskytů

Značka instance	Počet výskytů	%	Značka instance	Počet výskytů	%
?	51	0,55 %	o_	14	0,15 %
A	20	0,22 %	oa	664	7,17 %
ah	22	0,24 %	oc	14	0,15 %
at	8	0,09 %	oe	37	0,40 %
az	4	0,04 %	om	40	0,43 %
C	7	0,08 %	op	124	1,34 %
cap	61	0,66 %	or	21	0,23 %
f	223	2,41 %	p_	147	1,59 %
g_	7	0,08 %	P	945	10,20 %
gc	372	4,02 %	pb	7	0,08 %
gh	42	0,45 %	pc	74	0,80 %
gl	34	0,37 %	pd	30	0,32 %
gp	7	0,08 %	pf	1087	11,73 %
gq	48	0,52 %	pm	36	0,39 %
gr	69	0,74 %	pp	49	0,53 %
gs	62	0,67 %	ps	1380	14,90 %
gt	27	0,29 %	s	406	4,38 %
gu	754	8,14 %	segm	256	2,76 %
i_	4	0,04 %	T	93	1,00 %
ia	92	0,99 %	td	94	1,01 %
ic	392	4,23 %	tf	6	0,06 %
if	253	2,73 %	th	505	5,45 %
io	260	2,81 %	ti	38	0,41 %
lower	7	0,08 %	tm	99	1,07 %
mn	79	0,85 %	ty	141	1,52 %
mr	19	0,21 %	upper	4	0,04 %
mt	28	0,30 %			

Tabulka 4.8: Všechny značky vyskytující se v trénovacích datech setříděné podle abecedy

Značka kontejneru	Počet výskytů	%
A	20	1,88 %
C	7	0,66 %
P	945	88,73 %
T	93	8,73 %

Tabulka 4.9: Všechny značky vyskytující se v trénovacích datech setříděné podle abecedy

Kapitola 5

Experimenty s automatickým rozpoznáváním pojmenovaných entit v češtině

5.1 Vymezení úkolu

V této kapitole je popsán systém, který slouží k automatickému rozpoznávání pojmenovaných entit v českém textu. Úkol byl řešen specificky pro anotační schéma popsané v předcházející kapitole. Pro trénovaní a testování systému byla použita data vytvořená v prvním kole anotací.

Úlohu rozpoznávání pojmenovaných entit jsme se rozhodli rozdělit na dvě části: určení rozsahu pojmenovaných entit a určení jejich typu. *Rozsahem* entity rozumíme souvislou posloupnost tokenů (slov a interpunkčních znamének), která se vyskytla ve vstupní větě a která tvoří pojmenovanou entitu. *Typ* entity je hodnota z výčtu v tabulce 4.1. Za správně rozpoznanou entitu se považuje taková, která má správně určený rozsah i typ. Podrobnější popis vyhodnocování úspěšnosti bude uveden v sekci 5.4.

V trénovací i vyhodnocovací části systém využívá ručně anotovaná data převedená z původního textového formátu určeného pro ruční anotaci do formátu XML. Anotovaná data jsou rozdělena do dvou souborů: jeden reprezentuje původní text obohacený o morfologické značky a lemata (m-soubor, ukázka v příloze A.1), druhý pak odděleně reprezentuje jednotlivé instance pojmenovaných entit. V druhém souboru je pro každou z instancí uveden její typ a dva odkazy (začátek a konec rozsahu instance) do prvního souboru (ne-soubor, ukázka v příloze A.2).

5.2 Metoda

Určení typu a rozsahu entity řešíme jako oddělené úlohy. Určení typu je typická klasifikační úloha – daná posloupnost tokenů se zařadí do některého z definovaných typů. Naproti tomu určit rozsah entity znamená nalézt v textu posloupnost tokenů, která nějakou entitu tvoří. Taková posloupnost může být libovolně dlouhá, omezená jen délkou věty.

Značného zjednodušení se dá dosáhnout omezením délky entit, které se pokusíme rozpoznávat. Tím lze převést i určování rozsahu entit na klasifikační úlohu, ovšem za cenu, že nerozpoznáme entity, jejichž délka přesahuje stanovené maximum. V našem systému jsou takto rozpoznávány jednoslovné a dvouslovné pojmenované entity.

Vlastní klasifikace je založena na sledování atributů (*features*) u tokenů a bigramů.[1] Atributy jsou – formálně řečeno – efektivně vyčíslitelné funkce z množiny tokenů (resp. bigramů) do množiny racionálních nebo reálných čísel nebo do libovolné konečné množiny. Je to tedy cokoliv, co můžeme na tokenu (bigramu) sledovat, strojově vyhodnotit a co nabývá číselné nebo kategoriální hodnoty.[2] Třídu, do které bude ta která posloupnost tokenů zařazena, určí externí klasifikátor.

Celkem je úloha rozdělena na pět podúloh:

- Určení rozsahu jednoslovných entit
- Určení typu jednoslovných entit
- Určení rozsahu dvouslovných entit
- Určení typu dvouslovných entit
- Rozpoznání některých druhů víceslovných pojmenovaných entit

První čtyři podúlohy jsou řešeny pomocí klasifikace na základě atributů. Poslední je pak řešena zvláštními algoritmy. Sady atributů jsou různé pro každou klasifikační podúlohu.

Ukázky trénovacích vektorů atributů ve formátu pro c5.0 (viz níže) jsou uvedeny v přílohách A.3 a A.4. Vzorek vytvořeného rozhodovacího stromu (výstup z c5.0) je v příloze A.5.

5.3 Implementace

Systém je implementován v jazyce Perl a využívá klasifikátor *Rulequest c5.0.*[3] Jeho činnost je rozdělena na trénovací a analytickou část.

5.3.1 Trénování

Trénovací část se dělí na přípravu dat pro klasifikátor a využití klasifikátoru ke konstrukci klasifikační rutiny. Jak bylo řečeno v sekci 5.2, úloha zahrnuje čtyři klasifikační podúlohy. Pro každou z nich probíhá příprava dat pro klasifikátor stejně, liší se jen sadami použitých atributů. Rozlišujeme dva typy atributů, *kategoriální* a *pravdivostní.*

Pro určení rozsahu jednoslovných entit používáme následující atributy:

[1] Bigramem rozumíme dvojici sousedních tokenů v jedné větě.

[2] Mnoho atributů nabývá pravdivostních booleovských hodnot pravda/nepravda. To je považováno za zvláštní případ kategoriální hodnoty.

[3] `http://www.rulequest.com/`

- Výskyt v trénovacích datech `(in_data)`

  Kategoriální atribut, který udává, zda se token vyskytuje v trénovacích datech a tvoří tam pojmenovanou entitu.

  Algoritmus: Z trénovacích dat se shromáždí tokeny, jejichž lemata jsou stejná jako u vyhodnocovaného tokenu. Z těchto tokenů se vyberou ty, které tvoří pojmenovanou entitu. Pokud žádný takový není, vrátí se `0`. Pokud jsou více než čtyři, vrátí se `N` a typ některé z entit. Jinak se vrátí počet nalezených tokenů a typ některé z entit, které jsou těmito tokeny tvořeny.

- Začíná velkým písmenem `(capped)`

  Pravdivostní atribut. Pravda, pokud forma tokenu začíná velkým písmenem a není na začátku věty. Nepravda ve všech ostatních případech.

- Jméno podle lematu `(YSG)`

  Některá lemata jsou označena jako jména. Obvyklá křestní jména obsahují v technické příponě lematu řetězec `;Y`, příjmení `;S` a místopisné názvy `;G`. Tento kategoriální atribut nabude hodnoty `Y`, `S` nebo `G`, pokud lema obsahuje odpovídající označení. V opačném případě nabude hodnoty `0`.

- Jediný kandidát `(only_cap_num)`

  Pravdivostní atribut. Pravda, pokud je token jediným číslem ve větě (forma je tvořena samými číslicemi) nebo pokud je jediným slovem s velkým písmenem, nepravda jinak (není číslem, nemá velké písmeno nebo se ve větě vyskytuje jiné slovo s takovými vlastnostmi). Pro první slovo ve větě tento atribut vždy nabývá hodnoty nepravda.

- Je velké písmeno `(cap_letter)`

  Pravdivostní atribut, který nabývá hodnoty pravda, pokud je forma tokenu tvořena jediným velkým písmenem.

- Měsíc `(month)`

  Pravdivostní atribut. Pravda, pokud je lema jedno ze slov *leden*, *únor*, ..., *prosinec*.

- Jedno z čísel `(num_brother)`

  Pravdivostní atribut. Pravda, pokud je forma tvořena samými číslicemi a forma alespoň jednoho sousedního tokenu téže věty je také tvořena samými číslicemi.

- Hodina `(hour)`

  Pravdivostní atribut, který nabývá hodnoty pravda, pokud forma vypadá jako určení času na hodinách, např. *16:30*. Zda forma 'vypadá jako určení času', je dáno následujícím regulárním výrazem (notace podle Perlu):

  ```
  ^([01]?[0-9]|2[0-3])[.:][0-5][0-9]([ap]m)?$
  ```

- Lema `(lemma)`

  Kategoriální atribut. V případě, že je lema tokenu časté, je hodnotou atributu toto lema. Jinak řetězec `OTHER`. Mezi častá lemata byla vybrána ta, která se v trénovacích datech vyskytla aspoň desetkrát nebo která v trénovacích datech tvořila aspoň čtyři (jednoslovné) pojmenované entity.

- Město `(town)`

  Pravdivostní atribut, který nabývá hodnoty pravda, pokud je slovo jménem některého českého města. Seznam českých měst byl převzat z české Wikipedie.[4]

- Atributy založené na tagu `(T*)`

  Pro každou pozici v tagu (až na nevyužité pozice 13 a 14) je zaveden jeden atribut, který nabývá hodnot jako tag na dané pozici.

- Kontextové atributy `([yn]([LR]C)?[lf][1N]w)`

  Tyto atributy jsou motivovány pozorováním, že výskyt některých slov souvisí s tím, zda se kolem nich nachází pojmenovaná entita. Vyskytuje-li se v textu např. *MUDr.*, pak je velká pravděpodobnost, že následovat bude křestní jméno nebo příjmení. Rozlišujeme dvacet čtyři kontextových atributů. Jsou sledovány korelace mezi formou (resp. lematem) a jednou z následujících vlastností:

  - výskyt jednoslovné pojmenované entity před slovem
  - výskyt víceslovné pojmenované entity před slovem
  - výskyt jednoslovné pojmenované entity za slovem
  - výskyt víceslovné pojmenované entity za slovem
  - výskyt jednoslovné pojmenované entity tvořené slovem
  - výskyt víceslovné pojmenované entity obsahující slovo
  - absence jednoslovné pojmenované entity před slovem
  - absence víceslovné pojmenované entity před slovem
  - absence jednoslovné pojmenované entity za slovem
  - absence víceslovné pojmenované entity za slovem
  - absence jednoslovné pojmenované entity tvořené slovem
  - absence víceslovné pojmenované entity obsahující slovo

  Každý z těchto atributů je pravdivostní, takže jsou určeny pevné hranice, jaké formy a jaká lemata s výskytem či absencí pojmenované entity korelují a jaké nikoliv. Tato hranice je dána jednoduchým vzorcem.

  Pro každou formu (resp. lema) z trénovacích dat a pro každou z následujících vlastností:

  - forma je vlevo[5] od jednoslovné entity

[4] `http://cs.wikipedia.org/wiki/Seznam_českých_měst`

[5] Slovy 'být vlevo / vpravo od entity' myslíme být bezprostředním sousedem v rámci téže věty.

– forma je vpravo od jednoslovné entity
– forma tvoří jednoslovnou entitu
– forma je vlevo od víceslovné entity
– forma je vpravo od víceslovné entity
– forma je součástí víceslovné entity

se vyhodnotí tato čísla:

– kolikrát se slovo vyskytlo v trénovacích datech (označme ALL)
– kolikrát slovo mělo tu kterou vlastnost (označme HIT)

Forma nebo lema podle daného atributu pak ukazuje na tu kterou vlastnost, pokud $\frac{HIT}{ALL} > max(\frac{A}{ALL+C}+B, MIN)$. Jestliže $\frac{HIT}{ALL} < LOW$, forma nebo lema ukazuje na negaci. Použité koeficienty byly empiricky nastaveny takto: $A = 2.4$, $B = 0.4$, $C = 1$, $MIN = 0.5$, $LOW = 0.03$.

Atributy pro trénování typu jednoslovných entit jsou shodné s atributy uvedenými výše až na to, že atribut `lemma` je nahrazen kategoriálním atributem `suffix`. Ten se vyhodnocuje následovně:

Projde se seznam vybraných častých přípon (v technickém smyslu, tzn. n posledních znaků) od nejdelších k nejkratším (nejdelší jsou čtyřpísmenné, nejkratší jednopísmenné). Pokud základní část lematu vyšetřovaného slova končí některou z těchto častých přípon, nabude atribut `suffix` její hodnoty. Vždy se použije první nalezená přípona. Jestliže není žádná taková přípona nalezena, atribut `suffix` nabude hodnoty `OTHER`. Základní částí lematu rozumíme část od začátku k prvnímu výskytu některého ze znaků `^`, `‘`, `-`, `_`.

Do seznamu suffixů byly vybrány takové, které se u nějaké pojmenované entity vyskytly alespoň pětkrát a poměr počtu výskytů v entitě ku počtu všech výskytů suffixu byl alespoň 0.6.

Atributy pro trénování rozsahu a typu dvouslovných entit sdílejí tutéž sadu atributů:

- Vzorec slovních druhů `(pos_pat)`

  Kategoriální atribut. Hodnoty jsou dvojice slovních druhů (podle první pozice tagu) slov ve zkoumaném bigramu. Pokud daná dvojice slovních druhů tvoří v trénovacích datech méně než deset entit, použije se řetězec `OTHER`.

- Vzorec predikcí jednoslovných entit `(sw_pat)`

  Kategoriální atribut. Pro obě slova bigramu se vyhodnotí, zda je analyzátor určí jako pojmenované entity a jakého typu. Dvojice těchto typů (kde `0` znamená, že nejde o entitu) se pak použije jako hodnota atributu. Opět se definuje výčet častých hodnot a pro jiné hodnoty se použije implicitní řetězec `OTHER`. Přípustné dvojice typů jsou tyto:

  `(pf,0) (ps,0) (s,0) (0,s) (0,gu) (pf,ps) (0,0)`

- Vzorec velkých písmen (`cap_pat`)

  Kategoriální atribut. Pro obě slova bigramu se vyhodnotí, zda začínají velkým písmenem nebo jsou na začátku věty. Každé slovo má tedy jednu z vlastností 'je na začátku věty', 'začíná velkým písmenem' nebo 'nezačíná velkým písmenem'. Dvojice těchto vlastností obou slov je pak použita jako hodnota atributu.

- Shoda (`agree`)

  Pravdivostní atribut, který nabývá hodnoty pravda, jestliže se obě slova bigramu shodují v rodě, čísle a pádě.

- Výskyt v datech (`in_data`)

  Kategoriální atribut. Pokud vyšetřovaná dvojice lemat tvoří v trénovacích datech méně než tři pojmenované entity, pak je hodnotou tento počet, jinak `N`.

5.3.2 Analýza

Během analýzy jsou rozpoznávány a klasifikovány entity s využitím informace získané z trénovacích dat. Vstupem analýzy je tentokrát pouze m-soubor, výstupem je ne-soubor, který obsahuje nalezené instance pojmenovaných entit a který by měl být přirozeně co nejshodnější s ne-souborem vzniklým konverzí ruční anotace (pokud je pro daný text k dispozici).

Vstupní m-soubor je zpracováván po větách ve třech fázích. Zvlášť probíhá rozpoznávání

1. jednoslovných entit,
2. dvouslovných entit
3. a víceslovných entit.

Pro každé slovo se vyhodnotí atributy pro rozsah jednoslovné entity. Podle těchto atributů klasifikační rutina určí, zda slovo tvoří entitu či nikoliv. Pokud ano, vyhodnotí se atributy pro typ jednoslovné entity a klasifikační rutina určí podle hodnot atributů typ.

Pro každý bigram se provede totéž jako pro slova, jen s použitím atributů pro dvouslovné entity. V poslední fázi analýzy se věta prohledá na výskyt víceslovného českého města. Jestliže se nalezne, označí se jako entita typu `gu`.

5.4 Vyhodnocování

Podle zadání úlohy je úspěšnost měřena třemi hodnotami – precision, recall a F-measure. Precision je poměr počtu správně určených entit ku počtu všech určených entit. Pohybuje se tedy mezi 0 a 1; je rovna 1, pokud všechny entity, které systém určí, jsou určeny správně (ale třeba jsou některé vynechány), a nemá smysl, pokud systém neurčí žádnou entitu. Recall je poměr počtu správně určených entit ku počtu entit v textu (které 'měly' být rozpoznány). Pohybuje se taktéž mezi 0 a 1; je roven 1, pokud systém správně určí všechny entity,

které v textu jsou (ale navíc třeba chybně určí nějaké jiné), a nemá smysl, pokud v textu žádná entita není. F-measure je definována tak, aby shrnovala precision i recall v jednom čísle. Je nulová, když recall nebo precision jsou nulové, je jednotková, pokud precision i recall jsou jednotkové. Vzorec na výpočet F-measure je

$$\frac{2 \cdot precision \cdot recall}{precision + recall}$$

Za správně určenou entitu se považuje taková, která má správně určený rozsah i typ.

Aby byl zřetelnější podíl různých typů chyb, které při automatickém značkování pojmenovaných entit nastávají, zavádíme tři stupně 'přísnosti' vyhodnocování:

- predikovaná entita se musí s ručně anotovanou shodovat v rozsahu i typu,
- musí se shodovat v rozsahu a nadtypu, např. křestní jméno (`pf`) a příjmení (`ps`) mají týž nadtyp (`p`),
- musí se shodovat pouze v rozsahu.

Celkem tedy udáváme dvacet sedm čísel: tři veličiny (precision, recall, F-measure) pro každou z devíti kategorií (jednoslovné, víceslovné a všechny entity podle typu, jednoslovné, víceslovné a všechny entity podle nadtypu a jednoslovné, víceslovné a všechny entity podle rozsahu). U klasifikátorů používajících baseline (viz níže) udáváme méně podrobné výsledky.

Pro získání přehledu o tom, jaké výsledky považovat za úspěch a jaké nikoliv, byly navrženy dva triviální klasifikátory, tzv. baseline. První z nich za entitu označí každé slovo jinde než na začátku věty, které začíná velkým písmenem. Typ určí všem stejný, a to nejčastější v trénovacích datech (`ps`). Výsledky takového klasifikátoru jsou uvedeny v tabulce 5.1.

Druhá baseline rozhoduje na základě jediného atributu – výskyt v datech – jak je popsán v sekci 5.3.1. Za pojmenovanou entitu tedy označí každé slovo nebo dvojici slov, jejichž lema (lemata) tvoří v trénovacích datech pojmenovanou entitu. Typ pak určí takový, jaký má některá nalezená entita v trénovacích datech. Výsledky jsou uvedeny v tabulce 5.2.

Je vidět, že výsledky obou baseline se liší. První z nich má stejnou precision jako recall. Znamená to, že slov s velkým písmenem, která nejsou entitami, je stejně jako entit s malým písmenem. Zatímco rozsah určuje s velkou přesností, uniformně přidělený typ je správný spíše výjimečně. Druhý baseline klasifikátor naopak pokud nějakou entitu najde, pak jí skoro vždycky určí správný typ. To není žádné překvapení, protože tvoří-li lema nějakou entitu, pak jinde bude

	Podle typu entity	Podle nadtypu	Podle rozsahu
Precision	0.11	0.29	0.68
Recall	0.11	0.29	0.68
F-measure	0.11	0.29	0.68

Tabulka 5.1: Vyhodnocení první baseline ('velké písmeno')

	Podle typu entity	Podle nadtypu	Podle rozsahu
Precision	0.54	0.57	0.59
Recall	0.33	0.34	0.36
F-measure	0.40	0.43	0.45

Tabulka 5.2: Výhodnocení druhé baseline ('in-data')

	Všechny entity	Jednoslovné entity	Dvouslovné entity
Podle typu	0.74 / 0.54 / 0.62	0.72 / 0.69 / 0.70	0.93 / 0.22 / 0.35
Podle nadtypu	0.81 / 0.59 / 0.68	0.79 / 0.76 / 0.78	0.95 / 0.22 / 0.36
Podle rozsahu	0.88 / 0.64 / 0.75	0.87 / 0.84 / 0.86	0.98 / 0.23 / 0.37

Tabulka 5.3: Precision, recall a F-measure klasifikátoru natrénovaných na ručně anotovaných datech

pravděpodobně tvořit stejnou (pokud nějakou). Rozsah entit určuje druhý klasifikátor méně spolehlivě. Dá se předpokládat, že toto je do značné míry ovlivněno velikostí trénovacích dat. Čím větší budou, tím bude větší recall a menší precision (určí se víc entit, některé špatně).

Výsledky jednotlivých klasifikátorů natrénovaných a vyhodnocených na ručně označkovaných datech jsou uvedeny v tabulce 5.3.

Kapitola 6

Závěr

V této zprávě jsme shrnuli výsledky naší práce na tématu pojmenovaných entit v češtině. Podařilo se vytvořit jednoduché schéma pro anotaci pojmenovaných entit, získat konkrétnější empirickou představu o jejich rozložení v českých textech a s využitím ručně značkovaných dat provést první experimenty, při nichž byly uplatněny metody strojového učení pro automatické značkování pojmenovaných entit. V blízké budoucnosti bychom chtěli více 'vytěžit' anotovaná data (zejména s pomocí metod už ověřených na angličtině) a systematicky shromažďovat lexikální zdroje související s pojmenovanými entitami (např. geografické rejstříky). Dalším cílem je také zkoumat vhodnou reprezentaci pojmenovaných entit na abstraktnějších rovinách jazykové reprezentace, zejména na rovině tektogramatické.

Literatura

[NED, 1995] (1995). Named Entity Task Definition. Version 2.0. http://cs.nyu.edu/cs/faculty/grishman/NEtask20.book_1.html.

[Tok, 1995] (1995). Tokenization Rules. http://cs.nyu.edu/cs/faculty/grishman/tokenization-v12.book_1.html.

[TEI, 2003] (2003). Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange (P4). http://etext.lib.virginia.edu/standards/tei/teip4/index.html.

[Brunstein, 2002] Brunstein, A. (2002). Annotation Guidelines for Answer Types. http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html.

[Collins and Singer, 1999] Collins, M. and Singer, Y. (1999). Unsupervised Models for Named Entity Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, University of Maryland.

[Cucerzan and Yarowsky, 1999] Cucerzan, S. and Yarowsky, D. (1999). Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of 1999 Joint SIGDAT Conference on EMNLP and VLC*, pages 90–99, MD. University of Maryland.

[Fleischman and Hovy, 2002] Fleischman, M. and Hovy, E. (2002). Fine Grained Classification of Named Entities . In *COLING02*.

[Florian et al., 2003] Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named Entity Recognition through Classifier Combination. In *Proceedings of CoNLL-2003*, pages 168–171.

[Grishman and Sundheim, 1996a] Grishman, R. and Sundheim, B. (1996a). Design of the MUC-6 Evaluation. In *Annual Meeting of the ACL - Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996*, pages 413–422.

[Grishman and Sundheim, 1996b] Grishman, R. and Sundheim, B. (1996b). Message Understanding Conference - 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, volume I, pages 466–471.

[Hajič et al., 2001] Hajič, J., Hajičová, E., Pajas, P., Panevová, J., Sgall, P., and Hladká, B. V. (2001). Prague Dependency Treebank 1.0. Linguistic Data Consortium, CAT LDC2001T10, ISBN 1-58563-212-0.

[Hajič et al., 2006] Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., and Ševčíková Razímová, M. (2006). Prague Dependency Treebank 2.0.

[Hana et al., 2002] Hana, J., Hanová, H., Hajič, J., Hladká, B., and Jeřábek, E. (2002). Manual for Morphological Annotation. Technical Report TR-2002-14.

[Hladká et al., 2007] Hladká, B., Hajič, J., Hana, J., Hlaváčová, J., Mírovský, J., and Votrubec, J. (2007). *Czech Academic Corpus 1.0 Guide*. Karolinum - Charles University Press.

[Hlavsa et al., 1998] Hlavsa, Z., Hrušková, Z., Hůrková, J., Kraus, J., Martincová, O., Polívková, A., Sedláček, M., Svobodová, I., and Vlková, V. (1998). *Pravidla českého pravopisu*. Academia, Prague.

[Karlík et al., 2002] Karlík, P., Nekula, M., and Pleskalová, J., editors (2002). *Encyklopedický slovník češtiny*. Nakladatelství Lidové noviny.

[Klein et al., 2003] Klein, D., Smarr, J., Nguyen, H., and Manning, C. D. (2003). "named entity recognition with character-level models". In Daelemans, W. and Osborne, M., editors, *Proceedings of CoNLL-2003*, pages 180–183. Edmonton, Canada.

[Mikulová et al., 2005] Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová-Řezníčková, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., and Žabokrtský, Z. (2005). Anotace Pražského závislostního korpusu na tektogramatické rovině: pokyny pro anotátory. Technical report, ÚFAL MFF UK, Prague, Czech Republic.

[Palmer and Day, 1997] Palmer, D. D. and Day, D. S. (1997). A statistical profile of the named entity task. In *Proceedings of Fifth ACL Conference for Applied Natural Language Processing (ANLP-97)*, pages 190–193, Washington, D.C.

[Panevová, 1980] Panevová, J. (1980). *Formy a funkce ve stavbě české věty*. Academia, Praha.

[Rychlý, 2003] Rychlý, M. (2003). Získávání metainformací z textové podoby elektronicky dostupných článků. Bakalářská práce. Masarykova univerzita, Fakulta informatiky.

[ČSÚ, 2005] ČSÚ (2005). Statistický lexikon obcí České republiky 2005. Český statistický uřad a Ministerstvo vnitra České republiky.

[Sassano and Utsuro, 2000] Sassano, M. and Utsuro, T. (2000). Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition. In *COLING00*, volume 2, pages 705–711, Saarbrüken.

[Sekine, 2003] Sekine, S. (2003). Sekine's Extended Named Entity Hierarchy. http://nlp.cs.nyu.edu/ene/.

[Sekine, 2004] Sekine, S. (2004). Named Entity: History and Future. http://www.cs.nyu.edu/ sekine/papers/NEsurvey200402.pdf.

[Sgall, 1967] Sgall, P. (1967). *Generativní popis jazyka a česká deklinace*. Academia, Prague.

[Sgall et al., 1986] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

[Shinyama and Sekine, 2004] Shinyama, Y. and Sekine, S. (2004). Named Entity Discovery Using Comparable News Articles. In *Proceedings of 20th International Conference on Computational Linguistics*.

[Sundheim, 1995] Sundheim, B. (1995). Overview of Results of the MUC-6 Evaluation. In *Proceedings of the Sixth Message Conference (MUC-6)*, pages 13–31.

[Zeman et al., 2005] Zeman, D., Hana, J., Hanová, H., Hajič, J., Hladká, B., and Jeřábek, E. (2005). A Manual for Morphological Annotation, 2nd edition. Technical Report TR-2005-27, ÚFAL MFF UK, Prague, Czech Republic.

Příloha A

Ukázka zpracování anotovaných dat

A.1 Ukázka souboru s automatickou morfologickou anotací (m-soubor)

```
<s id='train-s161'>
<m id='train-s161m1'>  <form>Naopak</form>       <lemma>naopak</lemma>            <tag>Db-------------</tag> </m>
<m id='train-s161m2'>  <form>"</form>            <lemma>"</lemma>                 <tag>Z:-------------</tag> </m>
<m id='train-s161m3'>  <form>béčko</form>        <lemma>béčko</lemma>             <tag>NNNS1-----A----</tag> </m>
<m id='train-s161m4'>  <form>"</form>            <lemma>"</lemma>                 <tag>Z:-------------</tag> </m>
<m id='train-s161m5'>  <form>Sparty</form>       <lemma>Sparta_;K</lemma>         <tag>NNFS2-----A----</tag> </m>
<m id='train-s161m6'>  <form>patří</form>        <lemma>patřit_:T</lemma>         <tag>VB-S---3P-AA---</tag> </m>
<m id='train-s161m7'>  <form>spolu</form>        <lemma>spolu</lemma>             <tag>Db-------------</tag> </m>
<m id='train-s161m8'>  <form>s</form>            <lemma>s-1</lemma>               <tag>RR--7----------</tag> </m>
<m id='train-s161m9'>  <form>Litoměřicemi</form> <lemma>Litoměřice_;G</lemma>     <tag>NNFP7-----A----</tag> </m>
<m id='train-s161m10'> <form>a</form>            <lemma>a-1</lemma>               <tag>J^-------------</tag> </m>
<m id='train-s161m11'> <form>Esoxem</form>       <lemma>Esoxum</lemma>            <tag>NNNS7-----A----</tag> </m>
<m id='train-s161m12'> <form>Brno</form>         <lemma>Brno_;G</lemma>           <tag>NNNS1-----A----</tag> </m>
<m id='train-s161m13'> <form>k</form>            <lemma>k-1</lemma>               <tag>RR--3----------</tag> </m>
<m id='train-s161m14'> <form>nejvážněji</form>   <lemma>vážně_^(*1ý)</lemma>      <tag>Dg-------3A----</tag> </m>
<m id='train-s161m15'> <form>ohroženým</form>    <lemma>ohrožený_^(*4zit)</lemma> <tag>AAIP3----1A----</tag> </m>
<m id='train-s161m16'> <form>týmům</form>        <lemma>tým</lemma>               <tag>NNIP3-----A----</tag> </m>
<m id='train-s161m17'> <form>.</form>            <lemma>.</lemma>                 <tag>Z:-------------</tag> </m>
</s>
<s id='train-s162'>
<m id='train-s162m1'> <form>Petržela</form> <lemma>Petržela_;S</lemma>             <tag>NNMS1-----A----</tag> </m>
<m id='train-s162m2'> <form>učí</form>      <lemma>učit_:T</lemma>                 <tag>VB-S---3P-AA---</tag> </m>
<m id='train-s162m3'> <form>hráče</form>    <lemma>hráč</lemma>                    <tag>NNMP4-----A----</tag> </m>
<m id='train-s162m4'> <form>Sparty</form>   <lemma>Sparta_;K</lemma>               <tag>NNFS2-----A----</tag> </m>
<m id='train-s162m5'> <form>lépe</form>     <lemma>dobře</lemma>                   <tag>Dg-------2A----</tag> </m>
<m id='train-s162m6'> <form>bránit</form>   <lemma>bránit_^(klást_překážky)</lemma> <tag>Vf--------A----</tag> </m>
</s>
<s id='train-s163'>
<m id='train-s163m1'>  <form>Praha</form>       <lemma>Praha_;G</lemma>                  <tag>NNFS1-----A----</tag> </m>
<m id='train-s163m2'>  <form>-</form>           <lemma>-</lemma>                         <tag>Z:-------------</tag> </m>
<m id='train-s163m3'>  <form>Balík</form>       <lemma>balík-1_^(předmět)</lemma>        <tag>NNIS1-----A----</tag> </m>
<m id='train-s163m4'>  <form>žalob</form>       <lemma>žaloba</lemma>                    <tag>NNFP2-----A----</tag> </m>
<m id='train-s163m5'>  <form>ohledně</form>     <lemma>ohledně</lemma>                   <tag>RR--2----------</tag> </m>
<m id='train-s163m6'>  <form>privatizace</form> <lemma>privatizace</lemma>               <tag>NNFS2-----A----</tag> </m>
<m id='train-s163m7'>  <form>hutí</form>        <lemma>huť</lemma>                       <tag>NNFP2-----A----</tag> </m>
<m id='train-s163m8'>  <form>Poldi</form>       <lemma>Poldi_;K</lemma>                  <tag>NNFXX-----A----</tag> </m>
<m id='train-s163m9'>  <form>Ocel</form>        <lemma>ocel</lemma>                      <tag>NNFS1-----A----</tag> </m>
<m id='train-s163m10'> <form>se</form>          <lemma>se_^(zvr._zájmeno/částice)</lemma> <tag>P7-X4----------</tag> </m>
<m id='train-s163m11'> <form>stále</form>       <lemma>stále_^(*1ý)</lemma>              <tag>Db-------------</tag> </m>
<m id='train-s163m12'> <form>rozrůstá</form>    <lemma>rozrůstat_:T</lemma>              <tag>VB-S---3P-AA---</tag> </m>
<m id='train-s163m13'> <form>.</form>           <lemma>.</lemma>                         <tag>Z:-------------</tag> </m>
</s>
```

A.2 Ukázka souboru s instancemi pojmenovaných entit (ne-soubor)

```
<ne type='ic' start='train-s161m5' end='train-s161m5'/>
<ne type='ic' start='train-s161m9' end='train-s161m9'/>
<ne type='gu' start='train-s161m9' end='train-s161m9'/>
<ne type='ic' start='train-s161m11' end='train-s161m12'/>
<ne type='gu' start='train-s161m12' end='train-s161m12'/>
<ne type='ps' start='train-s162m1' end='train-s162m1'/>
<ne type='ic' start='train-s162m4' end='train-s162m4'/>
<ne type='gu' start='train-s163m1' end='train-s163m1'/>
<ne type='segm' start='train-s163m3' end='train-s163m3'/>
<ne type='if' start='train-s163m8' end='train-s163m9'/>
```

A.3 Ukázka trénovacích vektorů pro rozpoznávání jednoslovných pojmenovaných entit v c5.0

```
OTHER,OTHER,D,b,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m1.
DQUOT,OTHER,Z,COLON,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m2.
OTHER,cko,N,N,N,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m3.
DQUOT,OTHER,Z,COLON,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m4.
Sparta_K,OTHER,N,N,F,S,2,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,N,1,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,train-s161m5.
patit_T,OTHER,V,B,N/A,S,N/A,N/A,N/A,3,P,N/A,A,A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m6.
spolu,OTHER,D,b,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m7.
s1,OTHER,R,R,N/A,N/A,7,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,3,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m8.
OTHER,OTHER,N,N,F,P,7,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,1,1,G,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,train-s161m9.
a1,OTHER,J,ROOF,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,2,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m10.
OTHER,OTHER,N,N,N,S,7,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m11.
Brno_G,rno,N,N,N,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,N,1,G,0,0,0,0,0,1,1,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,train-s161m12.
k1,OTHER,R,R,N/A,N/A,3,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,2,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m13.
OTHER,OTHER,D,g,N/A,N/A,N/A,N/A,N/A,N/A,N/A,3,A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,train-s161m14.
OTHER,OTHER,A,A,I,P,3,N/A,N/A,N/A,N/A,1,A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m15.
tm,OTHER,N,N,I,P,3,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m16.
DOT,OTHER,Z,COLON,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s161m17.
OTHER,OTHER,N,N,M,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,1,S,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,train-s162m1.
OTHER,OTHER,V,B,N/A,S,N/A,N/A,N/A,3,P,N/A,A,A,N/A,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s162m2.
hr,OTHER,N,N,M,P,4,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s162m3.
Sparta_K,OTHER,N,N,F,S,2,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,N,1,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,train-s162m4.
dobe,OTHER,D,g,N/A,N/A,N/A,N/A,N/A,N/A,N/A,2,A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s162m5.
OTHER,OTHER,V,f,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s162m6.
Praha_G,OTHER,N,N,F,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,N,1,G,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,train-s163m1.
MINUS,OTHER,Z,COLON,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m2.
OTHER,OTHER,N,N,I,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,train-s163m3.
OTHER,OTHER,N,N,F,P,2,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m4.
OTHER,OTHER,R,R,N/A,N/A,2,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m5.
OTHER,OTHER,N,N,F,S,2,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m6.
OTHER,OTHER,N,N,F,P,2,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m7.
OTHER,di,N,N,F,X,X,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m8.
OTHER,OTHER,N,N,F,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m9.
se_zvr_zjmenostice,OTHER,P,7,N/A,X,4,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m10.
stle_1,OTHER,D,b,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,train-s163m11.
OTHER,OTHER,V,B,N/A,S,N/A,N/A,N/A,3,P,N/A,A,A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m12.
DOT,OTHER,Z,COLON,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,N/A,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,train-s163m13.
```

A.4 Ukázka trénovacích vektorů pro určení typu jednoslovných pojmenovaných entit v c5.0

```
Sparta_K,OTHER,N,N,F,S,2,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,N,1,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,ic,train-s161m5.
OTHER,OTHER,N,N,F,P,7,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,1,1,G,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,gu,train-s161m9.
Brno_G,rno,N,N,N,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,N,1,G,0,0,0,0,0,1,1,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,gu,train-s161m12.
OTHER,OTHER,N,N,M,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,1,S,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,ps,train-s162m1.
Sparta_K,OTHER,N,N,F,S,2,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,N,1,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,ic,train-s162m4.
Praha_G,OTHER,N,N,F,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,N,1,G,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,gu,train-s163m1.
OTHER,OTHER,N,N,I,S,1,N/A,N/A,N/A,N/A,N/A,A,N/A,N/A,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,segm,train-s163m3.
```

A.5 Ukázka rozhodovacího stromu vygenerovaného z c5.0

```
yl1w = 1:
:...nRCf1w = 0: 1 (1820/76)
:   nRCf1w = 1: 0 (4/1)
yl1w = 0:
:...YSG = S:
    :...yfNw = 0: 1 (1216/122)
    :   yfNw = 1: 0 (10)
    YSG = Y:
    :...cap_letter = 0: 1 (498/56)
    :   cap_letter = 1:
    :   :...in_data in {3,4,N}: 0 (0)
    :       in_data = 0: 0 (53)
    :       in_data = 1: 1 (6)
    :       in_data = 2: 0 (1)
    YSG = G:
    :...ylNw = 1: 0 (121/7)
    :   ylNw = 0:
    :   :...capped = 0: 0 (3)
    :       capped = 1:
    :       :...yfNw = 1: 0 (3)
    :           yfNw = 0:
    :           :...Tgrade in {2,3}: 1 (0)
    :               Tgrade = N/A: 1 (772/161)
    :               Tgrade = 1: 0 (16/2)
    YSG = 0:
    :...hour = 1: 1 (220/11)
```

A.6 Ukázka vygenerovaného souboru s automatickou anotací pojmenovaných entit

```
<ne type='ps' start='dtest-s5m4' end='dtest-s5m4'/>
<ne type='ps' start='dtest-s5m6' end='dtest-s5m6'/>
<ne type='gu' start='dtest-s5m8' end='dtest-s5m8'/>
<ne type='tm' start='dtest-s5m15' end='dtest-s5m15'/>
<ne type='pf' start='dtest-s6m1' end='dtest-s6m1'/>
<ne type='ps' start='dtest-s6m5' end='dtest-s6m5'/>
<ne type='pf' start='dtest-s6m11' end='dtest-s6m11'/>
<ne type='ps' start='dtest-s6m12' end='dtest-s6m12'/>
<ne type='P' start='dtest-s6m11' end='dtest-s6m12'/>
<ne type='ps' start='dtest-s6m43' end='dtest-s6m43'/>
<ne type='ps' start='dtest-s7m4' end='dtest-s7m4'/>
<ne type='pf' start='dtest-s7m8' end='dtest-s7m8'/>
<ne type='P' start='dtest-s7m8' end='dtest-s7m9'/>
<ne type='tm' start='dtest-s7m16' end='dtest-s7m16'/>
```

```
<ne type='ps' start='dtest-s8m3' end='dtest-s8m3'/>
<ne type='pf' start='dtest-s8m4' end='dtest-s8m4'/>
<ne type='P' start='dtest-s8m3' end='dtest-s8m4'/>
<ne type='pf' start='dtest-s8m5' end='dtest-s8m5'/>
<ne type='P' start='dtest-s8m4' end='dtest-s8m5'/>
<ne type='ps' start='dtest-s8m7' end='dtest-s8m7'/>
<ne type='pf' start='dtest-s8m10' end='dtest-s8m10'/>
<ne type='P' start='dtest-s8m10' end='dtest-s8m11'/>
<ne type='gl' start='dtest-s9m7' end='dtest-s9m7'/>
<ne type='gu' start='dtest-s9m8' end='dtest-s9m8'/>
<ne type='tm' start='dtest-s9m11' end='dtest-s9m11'/>
<ne type='gu' start='dtest-s9m25' end='dtest-s9m25'/>
```

A.7 Srovnání ruční a automatické anotace na vzorku vět

Uvnitř závorek před tokeny v následujícím vzorku je před lomítkem uveden typ pojmenované entity, který vyplývá z ruční anotace, a za lomítkem typ, který byl přiřazen automatickým značkováním.

```
(ia/)        Katolický
(ia/)        sjezd
             v
(p_,ic/ps)   Tailor
(ic/)        '
(ic/ps)      Hall
             v
(gu/gu)      Dublinu
             (
(ti/)        3
(ti/)        .
(ti/)        -
(ti/)        8
(ti/)        .
(ti/tm)      prosinec
             )
             .

(pf,P/pf)    W
(pf,P/)      .
(pm,P/)      W
(pm,P/)      .
(ps,P/ps)    Tomek
             o
             něm
             ve
             své
             monografii
(pf,P/pf,P)  Jan
(ps,P/ps,P)  Žižka
             mj
             .
             praví
             :
             "
             Ostatně
             také
             ze
             čtyř
             artikulů
             pražských
             jeden
             se
             vztahoval
             k
             stavování
             a
             kárání
             hříchů
             smrtelných
             a
             jiné
             nešlechetnosti
             při
             lidech
             světských
             a
             duchovních
             ,
             a
(ps/ps)      Žižka
             z
             něho
             odvozoval
             povolení
             své
             ke
             mstění
             nepravostí
             osob
             stavu
             kněžského
             ,
             které
             za
             předešlých
             časů
```

```
            obyčejně
            zůstávaly

            Britský
            multimediální
            umělec
(p_/ps)     Sting
            ,
            vlastním
            jménem
(P,pf/pf,P) Gordon
(ps,P/P)    Sumner
            ,
            který
            má
            vystoupit
(td,T/)     14
(td,T/)     .
(tm,T/tm)   června
            v
            pražské
(ic/)       Sportovní
(ic/)       hale
            ,
            bude
            s
            největší
            pravděpodobností
            bydlet
            se
            svým
            devětadvacetičlenným
            týmem
            pod
            krycím
            jménem
            v
            některém
            z
            pražských
            hotelů
            .

            Hrají
            :
(pf,P/ps,P) Jason
(P,pm/pf,P,P) Scott
(P,ps/pf,P) Lee
            ,
(pf,P/ps)   Cary
(ps,P/)     Elwes
            ,
(P,pf/pf,P) Lena
(P,ps/P)    Headeyová
            .
            -
            ci
            9
            /
            95
(mt/)       NOVA

            Na
            základě
            usnesení
            valné
            hromady
            založila
(if/gl)     Solo
(if,gu/gu)  Sušice
(td,T/)     1
(td,T/)     .
(tm,T/tm)   července
            z
            bývalých
            divizí
            čtyři
            akciové
            společnosti
(if/)       Solo
(if/)       Sirkárna
            ,
(if/)       Sololit
            ,
(if/)       Dřevařská
(if/)       výroba
(gu,if/gu)  Sušice
            a
(if/)       Sušická
(if/)       strojírna
            se
            stoprocentní
            účastí
            mateřské
            společnosti
            .
```