Markus G i g e r – Pavel Š t ě p á n (Praha)

POJMENOVANÉ ENTITY V POČÍTAČOVÉ LINGVISTICE A VLASTNÍ JMÉNA[1)]

1. Pojmenované entity v počítačové lingvistice

V 90. letech 20. století se začala počítačová lingvistika (konkrétně její směry Information Extraction a Natural Language Processing) zajímat o tzv. pojmenované entity. Při počítačovém zpracování přirozeného jazyka se naráželo víc a víc na to, že mnoho lexikálních jednotek (a některých dalších, zejména číselných výrazů) běžně vystupujících v textech není obsaženo ve slovnících, a proto se jen obtížně automaticky zpracovává. Bylo zjištěno, že je nutné automaticky rozpoznat jednotky jako jména (osob, organizací, míst) a numerické výrazy jako čas, datum, peněžní obnosy, procenta atd. Později byly připojeny jako další kategorie „artefakty“, jako např. tituly knih (*Odyssey*) nebo názvy výrobků (*Windows*). Ještě později byly zavedeny kategorie „geografická a politická entita“ (*Francie*, *New York*), a „Facility“ (*Empire State Building*). Počítalo se se sedmi až deseti kategoriemi pojmenovaných entit; řešení identifikace se mělo opřít o seznamy pojmenovaných entit a o programy schopné naučit se vlastnosti těch kterých kategorií a identifikovat je na jejich základě[2)].

Kromě skutečnosti, že pojmenované entity nebývají uváděny v běžných slovnících, se uvádějí i další jejich vlastnosti, které způsobují potíže při počítačovém zpracování: soubory pojmenovaných entit podléhají stálým změnám; mají specifické syntaktické chování (nulový člen); alternují s apelativy a zájmeny při zachování reference (*Tony Blair visited ... The Prime Minister emphasized ... He made it perfectly clear* ....); jejich forma je značně proměnlivá (*The Royal Bank of Scotland Inc.; the Royal Bank of Scotland; the Royal* etc.)[3)].

Někteří autoři počítají mezi pojmenované entity i biomedicinské termíny, tedy označení léků (*codeine*, *aspirin*), diagnózy (*decreased capillary fragility*), výsledky testů (*Abnormal skin pH*), symptomy (*dyspnea; nausea; Homans' sign*), testy

[1)] Tento příspěvek vznikl v rámci projektu Informační společnosti Grantové agentury Akademie věd ČR „Integrace jazykových zdrojů za účelem extrakce informací z přirozených textů“ (reg. číslo 1ET101120503).

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

[3)] Srov. L. Lemnitzer: Named Entity Recognition, http://www.sfs.nphil.uni-tuebingen.de/~lothar/TT05/Slides/slidesNER.pdf (materiály pro výuku).

(*radioimmunoassay; acoustic rhinometry*), léčebné postupy (*hemodialysis; cesarean section*), označení genů (*c-Ha-ras*)[4]. Najdou se například i formulace jako následující: „Named Entity Recognition involves identification of proper names in texts, and classification into a set of predefined categories of interest: Person names, organizations (companies, government organisations, committees, etc), locations (cities, countries, rivers, etc), date and time expressions, other common types: measures (percent, money, weight etc), email addresses, Web addresses, street addresses, etc., some domain-specific entities: names of drugs, medical conditions, names of ships, bibliographic references etc.“[5].

2. Vztah pojmenovaných entit k vlastním jménům

Z hlediska onomastiky si můžeme v souvislosti s kategorií pojmenovaných entit v počítačové lingvistice položit dvě základní otázky: jednak, zda mají pojmenované entity, jak bývají definovány v počítačové lingvistice, nějaký společný jazykový jmenovatel, který umožňuje jejich shrnutí do jedné velké skupiny, a jednak, které pojmenované entity představují vlastní jména a které nikoliv.

Z citovaných definic celkem jasně vychází, že pro autory zabývající se pojmenovanými entitami stála v popředí mnohem více praktická obtížnost jejich zpracování než jejich uspokojivá teoretická definice. Tato obtížnost souvisí jednak s tím, že nejsou (a ani nemohou být) k dispozici úplné seznamy těchto jednotek, a jednak s tím, že jejich struktura je velmi různá (srov. antroponymum typu *Karel Novák* s pojmenováním instituce typu *Ústav pro jazyk český Akademie věd České republiky, veřejná výzkumná instituce*). Uvedená proměnlivost pojmenovaných entit je zřejmě různá a závisí také na vlastnostech jednotlivých jazyků, resp. jejich systémů pojmenování: v oblasti antroponym se v běžných evropských pojmenovacích systémech soubor rodných jmen proměňuje jenom pozvolna (hlavně přejetím z jiných jazyků a osamostatněním hypokoristických podob)[6]. Soubor příjmení je teoreticky uzavřen a rozšiřuje se hlavně migrací nositelů příjmení z jedné jazykové oblasti do druhé. Oproti tomu je nárůst počtu nových pojmenování výrobků rychlý a z hlediska teoretického zcela neomezen. I proměnlivost vnitřní struktury je různá: v oblasti antroponym sleduje určitá pravidla *(Karel Novák, Karel, Kája, Novák, pan Novák* atd.), v oblasti toponym je nejčastější u urbanonym (*Kaprova ulice, Kaprova,*

---

[4] Srov. P. Zweigenbaum, Question Answering in Biomedicine, 2003, http://staff.science.uva.nl/~mdr/NLP4QA/ZweigenbaumEACLW2003.exp.8up.pdf (referát v rámci workshopu „Natural Language Processing for Question Answering“ na konferenci Evropské pobočky Asociace počítačových lingvistů).

[5] Srov. D. Maynard, http://gate.ac.uk/talks/bcs-03-cheltenham.ppt (podklady referátu na shromáždění Britské počítačové společnosti, 2003).

[6] Srov. M. Knappová, Jak se bude vaše dítě jmenovat, 4. vyd., Praha 2006, s. 17 – 23.

Kaprovka; *náměstí I. P. Pavlova*, *Pavlák*, *Ípák*), nezřídka se však vyskytují i neoficiální podoby místních jmen (*Valašské Meziříčí – Valmez*, *České Budějovice – Budějovice*, *Buděj(i)ce*, *Budějky*), u institucí je situace ještě složitější a méně přehledná *(Ústav pro jazyk český Akademie věd České republiky, veřejná výzkumná instituce, Ústav pro jazyk český Akademie věd České republiky, Ústav pro jazyk český)*. Uvedená syntaktická zvláštnost pojmenovaných entit (nulový člen) je jednak jazykově specifická (týká se pouze některých jazyků disponujících systémem určitého a neurčitého členu) a jednak se ani tam netýká všech názvů, jak vychází z příkladu *the Royal Bank of Scotland* téhož autora. Koreference se zájmeny je vlastní i názvům obecným, ty se ovšem mohou střídat s vlastními jmény pouze v některých případech (pokud odkazují k individuu, které je nositelem vlastního jména). Navíc časté střídání uvedeného typu (*Tony Blair*, *The Prime Minister*) je závislé na druhu textu, charakteristické je převážně pro publicistické texty.

Zajímavější z hlediska onomastiky je však otázka, které typy pojmenovaných entit v chápání počítačové lingvistiky představují vlastní jména z pohledu onomastiky. Lze zde rozlišit minimálně dvě skupiny, v závislosti na přístupu i tři: pojmenované entity, které jsou jednoznačně a bez pochyb vlastními jmény, pojmenované entity, které patří do skupin, jejichž status je sporný i v rámci onomastiky, a popř. pojmenované entity, které z hlediska onomastiky vlastními jmény nejsou. Do první skupiny patří osobní jména (antroponyma), zeměpisná jména (toponyma) a jména věcí (chrématonyma).

Mezi vlastní jména nepatří numerické výrazy (číslice označující počet nebo pořadí, procentuální údaje aj.). Vlastními jmény dále nejsou např. internetové adresy typu *http://www.ujc.cas.cz*, které sice označují jednoznačně určitou webovou stránku, ale z jazykového hlediska bychom je sotva přiřadili k vlastním jménům (byť vlastní jména často obsahují, srov. *http://www.praha.cz*). Adresy – ať e-mailové či poštovní – jsou jednotkami určité soustavy, která dovoluje identifikovat místo dodání zásilky; jako takové obsahují i vlastní jména, ale nepředstavují samy vlastní jméno. Obdobně je tomu s bibliografickými údaji. Celá řada pojmenovaných entit patří k jednotkám, které jsou chápány různými způsoby. Jsou to třeba označení úseků času: podle Van Langendoncka mohou mít označení jako *winter*, *Monday*, *Easter* charakter jak apelativ, tak i vlastních jmen, podle toho, v jakých konstrukcích vystupují[7)]. V rámci označení času uvádí Van Langendonck typologicky periferní skupinu názvů událostí jako *Second World War*, *French Revolution*[8)]. R. Šrámek uvádí názvy tohoto typu v rámci skupiny „vlastní jména společenského jevu“[9)]. I tyto entity hrají důležitou úlohu

[7)] W. Van Langendonck, Theory and Typology of Proper Names, Berlin – New York, 2007, s. 186.
[8)] W. Van Langendonck, o. c. v pozn. 7, s. 228.
[9)] Srov. R. Šrámek, Úvod do obecné onomastiky, Brno 1999, s. 165.

v rámci identifikace pojmenovaných entit v elektronických korpusech. Speciální problém představují názvy čísel (a písmen, které ovšem vystupují oproti prvním velmi zřídka): ačkoliv jsou podle Van Langendoncka také vlastními jmény (*Three is a sacred number, Five is a figure*), vystupují v drtivé většině případů v apelativním, resp. atributivním užití (*Is this a six?, The room contains two tables*). Asi z tohoto důvodu názvy čísel a písmen u jiných autorů mezi skupinami vlastních jmen nenajdeme[10]. Další skupinou sporných vlastních jmen jsou názvy jazyků, barev a nemocí. Opět je nacházíme u Van Langendoncka, ovšem nikoliv u ostatních autorů.

U Van Langendoncka tím vzniká velmi široké pojetí vlastních jmen, která jsou klasifikována na vlastní jména prototypická[11] a neprototypická[12]. K prototypickým jménům patří nejen antroponyma a toponyma, ale např. také jména budov a lodí[13], organizací a institucí. Mezi neprototypická jména jsou řazeny názvy časových úseků, uměleckých děl, obchodních značek apod.

Je jasné, že v korpusu, zejména při automatické morfologické anotaci, není možné rozlišovat onymické a apelativní užití určité jednotky. Není ani možné zařazovat určité jednotky do skupin více či méně prototypických vlastních jmen. Je však žádoucí zvolit si třídění pojmenovaných entit, které odpovídá co nejvíce požadavkům lingvistickým (včetně onomastických) i praktickým: mělo by umožnit anotovacímu programu co nejpřesnější anotaci pojmenovaných entit a mělo by dovolit najít jednoduchým způsobem určité typy pojmenovaných entit pro praktické účely (osoby, adresy, telefonní čísla etc.), ale i pro lingvistickou analýzu (jaká jména institucí či výrobků se vyskytují v korpusu?).

3. Klasifikace pojmenovaných entit

V rámci projektu Informační společnosti Grantové agentury Akademie věd ČR „Integrace jazykových zdrojů za účelem extrakce informací z přirozených textů“ byl první návrh klasifikace pojmenovaných entit vypracován M. Ševčíkovou, Z. Žabokrtským a O. Krůzou[14]. Zavádí deset skupin pojmenovaných entit: čísla jako součást adres, součásti bibliografických údajů, geografické názvy, názvy institucí, názvy médií, čísla se specifickým významem, názvy věcí, jména osob, čísla

[10] Srov. R. Šrámek, o. c. v pozn. 9; L. Olivová-Nezbedová, K hierarchii onomastických termínů podle objektů pojmenovaných vlastními jmény, Acta onomastica 39, 1998, s. 46 – 54.
[11] W. Van Langendonck, o. c. v pozn. 7, s. 186 – 222.
[12] W. Van Langendonck, o. c. v pozn. 7, s. 223 – 246.
[13] W. Van Langendonck, o. c. v pozn. 7, s. 220.
[14] Viz M. Ševčíková – Z. Žabokrtský – O. Krůza, Zpracování pojmenovaných entit v českých textech. Technická zpráva Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy, Praha 2007.

s významem počtu a pořadí a časové údaje[15)]. Není zde místo pro podrobnou analýzu dané klasifikace, zejména jejích početných podskupin, chceme se zde omezit na ty skupiny vlastních jmen, pro které jsme navrhli určitou revizi. Týká se to názvů geografických, názvů institucí, názvů věcí a jmen osob.

Původní klasifikace Ševčíkové, Žabokrtského a Krůzy v oblasti geografických názvů vypadá takto:

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>

Pro tuto klasifikaci jsme navrhli následující úpravy:

gb **názvy budov** (*Kramářova vila, Wiehlův dům*)**, hradů** (*Pernštejn*)**, zámků** (*Kozel*) **a dalších staveb a objektů** (*Europark, Důl Hlubina*)

gc **státní útvary** (*Česká republika, Estonsko*)

gh **vodní útvary**

go **terénní útvary** (*Petřín, Krkonoše, Kokořínský důl, Macocha, Koněpruské jeskyně*)

gp **vesmírné objekty**

gr **územní celky (dané jak administrativně a/nebo historicky, tak přírodně)** (*Morava, Bavorsko, Kosovo*; *Příbramsko*; *Blízký východ*; *Polabí, Apeninský poloostrov*)

[15)] M. Ševčíková – Z. Žabokrtský – O. Krůza, o. c. v pozn. 12, s. 34 – 36.

gs **ulice a ostatní městské komunikace a veřejná prostranství** (*Mánesova, Nové náměstí, Zelný trh, Masarykovo nábřeží, Čechův most, Havlíčkovy sady*)

gt **kontinenty**

gu **obce a části obcí**

gv **nádraží, stanice a zastávky veřejné dopravy** (*Masarykovo nádraží, Aš město, Ládví, Koleje Jižní Město, Šeberák*)

gz **ostatní pomístní jména – pozemky, komunikace aj.** (*Vidrholec, Podvesní pole, Černá louka*; *Umrlčí cesta, Posázavský pacifik, D1, Transsibiřská magistrála*; *Poušť Gobi*)

Přehled a odůvodnění změn:

1) Zřízena nová skupina **gb** zahrnující názvy budov (včetně stavebních památek) a některých dalších člověkem vytvořených objektů (např. dolů), nezařaditelných do jiných skupin. Do této skupiny budou přesunuty takové pojmenované entity ze skupin **ic** a **if** (viz níže), které se spíše než k institucím vztahují k budovám, případně dalším objektům. Ze skupiny **oa** (viz níže) sem budou přesunuty entity označující stavební památky (charakter těchto „artefaktů“ je zcela jiný než u artefaktů ostatních, důležité je zde především zakotvení na určitém místě). Kostely, chrámy atd. se podle dosavadní klasifikace zařazovaly do jiné skupiny než hrady a zámky[16)].

2) Skupinu **gl** navrhujeme zrušit a pojmenované entity z ní přesunout do skupin **gr** a **go**, případně též **gz**. Nezdá se totiž účelné rozlišovat územní celky přírodní a jiné – vést přesnou hranici mezi nimi je velmi obtížné. Názvy přírodních oblastí budou tedy zařazeny mezi názvy ostatních územních celků (**gr**), názvy přírodních útvarů mezi názvy terénních útvarů (**go**), případně – pokud by se jednalo o jiné než terénní útvary (např. názvy pouští) – mohli bychom uvažovat o zařazení mezi ostatní pomístní jména (**gz**).

3) Zřízena nová skupina **go** zahrnující názvy terénních útvarů. Pokud existuje skupina **gh** – názvy vodních útvarů (hydronyma), je vhodné zavést také samostatnou skupinu pro názvy útvarů terénních (oronyma), která by zahrnovala jak vyvýšeniny (vrchy, hory, pohoří), tak sníženiny (údolí, rokle, propasti; volně sem lze přiřadit také jeskyně).

[16)] Problematické a velmi obtížně zařaditelné jsou názvy částí budov – jednotlivých místností, sálů. síní apod. (např. *Španělský sál*, *Dvořákova síň*, *Zelená komnata*). Tyto názvy jsou na pomezí mezi toponymy a chrématonymy, ovšem ani do jedné z těchto skupiny nejsou zařaditelné bez jistých obtíží. Vzhledem k tomu, že označují objekty pevně fixované na určitém místě, navrhujeme je volně přiřadit k názvům budov, tedy mezi toponyma.

4) Přikláníme se ke zrušení skupiny **gq**, protože je často velmi obtížné rozpoznat, zda se jedná o obec, nebo její část. Není to vždy snadné ani u městských čtvrtí – srov. např.:
98: Telefonát jednoho ze zakladatelů <ia Porty> <P<pf Miloslava> <ps Langra>> , přenášený na ***<gu Lochotín>*** *z <gu Mnichova>...*
(Anotátor přiřadil název *Lochotín* do skupiny **gu**, protože bez znalosti místa nemohl rozpoznat, že se jedná o plzeňskou čtvrť.)

5) Skupina **gr** bude rozšířena na všechny druhy územních celků – celky administrativní, celky dané historicky i celky přírodní. Do této skupiny bude přesunuta část pojmenovaných entit ze zrušené skupiny **gl**. Název této skupiny nemůže být „menší územní jednotky“, neboť sem budou zahrnuty i územní celky rozsáhlejší (např. *Západ*, *Blízký východ*). Patří sem všechny územní celky s výjimkou státních útvarů (**gc**) a kontinentů (**gt**).

6) Skupina **gu** zahrnuje jak samostatné obce, tak jejich části (včetně městských čtvrtí).

7) Zřízena nová skupina **gv** – názvy letišť, přístavů, nádraží, stanic a zastávek veřejné dopravy.

8) Zřízena nová skupina **gz**, do níž spadají pomístní jména (názvy neosídlených objektů) nezařaditelná do ostatních skupin – jedná se především o jména pozemků (včetně lesů) a komunikací. Alternativně by bylo také možné rozčlenit **pozemky** a **komunikace** do dvou samostatných skupin.

Názvy institucí byly původně tříděny Ševčíkovou, Žabokrtským a Krůzou následujícím způsobem:

i	NÁZVY INSTITUCÍ	
ia	přednášky, konference, soutěže	<ia Stanley Cup>
ic	kulturní, vzdělávací a vědecké instituce, kino, sportovní kluby	<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.>

Zde[17] jsme navrhli vydělit samostatnou skupinu, do které by patřily akce (přednášky, konference, soutěže apod.), protože v těchto případech se nejedná o instituce. Důležitým kritériem je časové vymezení/ohraničení takovýchto pojmenovaných entit. Do této skupiny by bylo vhodné zařadit i historické události, které nejsou zařaditelné do žádné ze stávajících (pod)skupin. Výsledek by byl následující[18]:

NÁZVY INSTITUCÍ

(ia – přednášky, konference, soutěže: ruší se, resp. převádí se do ec)

ic – kulturní (divadla, kina, ...), vzdělávací, vědecké instituce, sportovní kluby

if – firmy, koncerny, hotely, restaurace, nakladatelství...

io – státní a mezinárodní instituce, politické strany a hnutí, náboženské skupiny, občanská sdružení, mezinárodní nestátní organizace

NÁZVY AKCÍ A UDÁLOSTÍ

ec – vědecké, kulturní, společenské a sportovní akce

eh – historické události

Skupina názvy věcí měla původně tuto strukturu:

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 <oe MHz>
om	měny (zapsané zkratkou, symbolem)	1000 <om Kč>, 30 <om $>
op	výrobky	mikroprocesory <op Intel Pentium>

[17] V této části jsme vycházeli z následujících prací: M. Majtán, Klasifikácia chrématonymie. In: R. Šrámek – L. Kuba (eds.), Chrématonyma z hlediska teorie a praxe, Brno 1989, s. 7 – 13; R. Šrámek, Namen von Sachen (Chrematonymie) I. In: E. Eichler – G. Hilty – H. Löffler – H. Steger – L. Zgusta (eds.), Namenforschung. Name Studies. Les noms propres 2, Berlin – New York 1996, s. 1562 – 1567; M. Knappová, Namen von Sachen (Chrematonymie) 2. In: E. Eichler – G. Hilty – H. Löffler – H. Steger – L. Zgusta (eds.), Namenforschung. Name Studies. Les noms propres 2, Berlin – New York 1996, s. 1567 – 1572.

[18] Změny oproti původní klasifikaci jsou vyznačeny podtržením.

or	předpisy, normy, jejich sbírky	ve <or Sbírce zákonů>
o_	názvy nespecifikovaného typu	čeleď <o Rubiaceae>, <o HIV>

Zde navrhujeme následující změny:

oa – kulturní artefakty (knihy, filmy, pořady...). Stavby se převádějí se do gb.
oc – názvy chemikálií
oe – měrné jednotky (zapsané zkratkou)
om – měny (zapsané zkratkou, symbolem)
op – (sériové) výrobky (výrobní typy, značky atd., bez názvů individuálních výrobků)
or – smlouvy, doktríny, ústavy, předpisy, normy, jejich sbírky
os – ocenění, řády, vyznamenání, plakety
ot – názvy individuálních výrobků (letadel, lodí, drahokamů atd.) a dopravních spojů

Z hlediska prototypičnosti jména se jeví jako účelné rozlišovat názvy výrobků sériových a výrobků individuálních. Skupina op je proto zúžena na výrobky sériové, nově je z ní vydělena skupina **ot** (individuální výrobky).

Do žádné z původních skupin nelze bez obtíží zařadit názvy dopravních spojů, např. rychlík *Krakonoš, Bečva*..., metro *A*. Volně je přiřazujeme k názvům individuálních výrobků.

Obtížně zařaditelné jsou také názvy cen a ocenění (označeny u Majtána jako faleronyma)[19], např. *Oscar* – v rámci názvů věcí se pro ně vyděluje nová skupina **os**.

4. Závěr

Pod označením „pojmenované entity“, používaným v počítačové lingvistice od 90. let 20. stol., se skrývají různorodé skupiny slov a dalších (numerických) výrazů, které v běžných slovnících nebývají zpracovány. Tato skutečnost je primárně spojuje, avšak ani při velmi širokém chápání vlastního jména nepatří všechny skupiny pojmenovaných entit k propriím. Vymezení pojmenovaných entit na základě jejich vlastností není pro značnou různorodost jednotlivých skupin pojmenovaných entit možné. Pojmenované entity proto bývají zpravidla vymezovány víceméně negativně – jako jednotky v textu, které nejsou obsaženy ve slovníku, a jsou proto při automatickém zpracování přirozeného jazyka (natural language processing) obtížně rozpoznatelné. V rámci anotace textů z Českého národního korpusu[20] byla vyvinuta klasifikace pojmenova-

[19] Srov. M. Majtán, o. c. v pozn. 17, s. 11.
[20] Viz http://ucnk.ff.cuni.cz/.

ných entit, která se snaží brát ohled jak na potřeby korpusové a praktické, tak na potřeby onomastické.

Summary

Named Entities in Computational Linguistics and Proper Names

The article deals with the relationship between the so-called „named entities“, the concept of which has been established by the computational linguistics (namely its fields of Information Extraction and Natural Language Processing), and the proper names, as they are understood by onomasticians. Named entities are understood as those text units that are not included in dictionaries, and therefore cause difficulties during the automatic processing. Named entities include not only proper names, but also numeric expressions, Internet and e-mail addresses, during to some conceptions also biomedical terms, etc. The longest part of the article is devoted to the classification of named entities, as it has been proposed by computational linguists and modified by onomasticians.