Segmentace textu na věty

Helena Palátová – Marek Grác
Centrum počítačové lingvistiky, Filozofická fakulta MU Brno
helena.palatova@gmail.com

ABSTRACT: Pro češtinu v současné době existuje spousta nástrojů schopných na dobré úrovni popsat její morfologickou rovinu, ale co se týče její volnější, a proto hůře formálně popsatelné syntaxe, musí se (i přes všechny dosavadní pokusy o vytvoření kvalitního automatického syntaktického analyzátoru) nejen korpusoví lingvisté prozatím obejít bez nástrojů, které by byly schopny automaticky do textových korpusů vložit syntaktické značky, podle nichž by se lépe v textech vyhledávalo a zkoumalo jazyk na této rovině.

Aby byly nástroje schopny účinně rozpoznávat a popisovat vztahy mezi jednotlivými textovými slovy, potřebují nejprve jasně a především správně rozpoznat hranice jednotlivých autonomních celků, tedy hranice vět. V naší práci jsme se zabývali delimitací věty (*sentence*). Tato problematika byla sice již v minulosti řešena a prakticky každý korpus má vyznačené hranice vět, ale problematiku určování hranic vět v okrajových případech stále nepovažujeme za dořešenou. Až po dořešení tohoto problému je možné zjistit, jak kvalitně fungují existující automatické nástroje, a můžeme řešit jejich vylepšování.

Proto jsme zvolili následující postup. Nejprve byla stanovena formální pravidla, která byla poté použita v návodu pro anotátory. Při tvorbě pravidel jsme se zaměřili zvláště na formální zvláštnosti některých syntaktických konstrukcí používaných v českých textech. Právě těmito pravidly se řídili anotátoři při práci s korpusem současných blogových textů. Text ručně segmentovali na jednotlivé věty a vytvořili tím jednotná data. Takto vytvořená data byla základem pro upřesnění pravidel s cílem dosáhnout vyšší interanotační shody. V příspěvku prezentujeme formální a jednoznačná pravidla pro určování hranic vět spolu s jejich zdůvodněním na korpusovém materiálu. Vytvořená data plánujeme zpřístupnit tak, aby mohla sloužit nejen jako testovací data, ale i jako zdroj pro statistické (automatické) strojové učení.

1 Úvodem

V naší práci se zabýváme delimitací věty, resp. určením pravidel pro problematické případy stanovení hranic věty. Tato problematika byla sice již v minulosti řešena (Julinek, 1999: 4; Kiss, Strunk, 2006; O'Neil; Wikipedia) a prakticky každý korpus má odlišnými postupy a s různě kvalitními výsledky vyznačené hranice vět, ale právě kvůli nedostatečným definicím hranic vět (zvláště v problematických případech), které jsou příčinou nedostatečné přesnosti výsledných dat, působící komplikace při dalším zpracovávání textu, toto téma stále nepovažujeme za dořešené.

2 Strategie

Při tvorbě korpusů je ruční značkování hranic vět obvykle příliš nákladné, a proto se využívají automatické metody, které tento proces urychlují a činí lacinějším, i za cenu vyšší míry chybovosti. Konkrétní přístupy k automatickému určování hranic jednotlivých vět jsou různé. Prvním z nich je vytvoření automatického oddělovače vět – programu, který určuje hranice podle pravidel (např. pokud za interpunkčním znaménkem (tečkou) následuje velké písmeno,

vloží mezi ně program znak hranice věty; pokud je token[1] předcházející tečce obsažen v předpřipraveném seznamu zkratek, není za tečkou znak hranice věty). Pro angličtinu fungují tyto programy s přesností 95 % (O'Neil).

Další rozšířenou strategii, která má již vyšší úspěšnost (pro angličtinu až 99,5 %) využívají programy na (automatické) naučení konců vět. Z trénovacích dat, ve kterých jsou anotátory vyznačeny hranice vět, se naučí pravidla, potřebná pro rozhodování o hranicích vět. Řešení může být založené např. na modelu maximální entropie (Reynar, Ratnaparkhi, 1997). Na vstupu je na každém řádku věta, díky čemuž se snadno vytvoří seznam slov, za kterými je tečka, ale věta jí nekončí (zkratky, řadové číslovky, e-mailové adresy, emotikony, webové adresy, apod.).

Existují ještě jiné přístupy řešící tuto problematiku, například se k desambiguaci hranic vět využívá neuronových sítí (př. systém SATZ (Palmer, 1995; Palmer, Hearst, 1997)). Výsledky dosahují až 98,5% přesnosti (Wikipedia).

Pro brněnské korpusy se využívá nejdříve popsaný přístup, tedy program s pravidly fungující následovně:

- ke všem zkratkám, které jsou obsaženy v morfologickém analyzátoru (m)ajka (Sedláček, Smrž, 2001), se přidá varianta s velkým písmenem,
- z nich jsou vybrány ty, které jsou i platným slovem (gen. – genitiv / gen; plk. – plukovník / plk; ...),
- římské číslice představují problém; arabské (2., 5.) se berou stejně jako zkratky,
- strukturní značky a <doc></doc> jsou vždy hranice vět, ostatní se ignorují,
- poté se už používá jen testování, zda po interpunkci následuje velké písmeno (s ohledem na zkratky),
- pokud je věta příliš dlouhá, „někde" se rozdělí, přestože bychom ji normálně nechali vcelku.

Problematickými jsou tudíž především římské číslice a poslední zmíněná skutečnost. Kvůli již zmíněným nedostatkům našeho stávajícího systému jsme se rozhodli změnit strategii řešení segmentace textu na věty. Zvolili jsme druhý popisovaný obecný postup, tedy získat označkovaná data, která budou dále sloužit jako trénovací pro statistické strojové učení. Důležitým úkolem bylo tudíž stanovení formálních pravidel, jimiž by se anotátoři řídili, pokud možno co nejjednotnějším způsobem.

3 Hledání definice termínu věta

Nejprve je třeba zmínit, jak se na větu dívá tradiční lingvistika. Pro účely tohoto článku nebudeme terminologicky rozlišovat větu a výpověď. Pracujeme s reálnými výpověďmi, které ale potřebujeme vymezit, přičemž se nevyhneme generalizacím a abstraktivizacím, povrchovému popisu struktur. Proto vše zahrnujeme pod pojem věta v širokém slova smyslu (v anglické terminologii *sentence*), tedy z formálního hlediska může jít jak o samostatnou větu jednoduchou (klauzi), tak o souvětí.

Ještě v 19. století nebyla věta pokládána za systémovou jednotku, protože je díky rekurzi potenciálně možné vytvořit nekonečné množství vět. Oproti tomu je množství systémových jednotek na ostatních rovinách jazyka konečné. Během první poloviny 20. století se však

[1] token = textové slovo, řetězec ohraničený mezerami

hledání definice věty stalo jedním z hlavních témat lingvistiky (Palek, 1989: 155). Skutečnost, že se dodnes nikomu nepodařilo formulovat definici věty, která by se stala obecně přijímanou, ukazuje na problematičnost tohoto úkolu. Zřejmě největší komplikace při vymezování tohoto termínu způsobuje právě již zmiňovaný fakt, že je možné vytvořit nekonečně mnoho konečně dlouhých vět. Kdybychom chtěli být při delimitaci věty exaktní, museli bychom respektovat všechny její možné varianty, a to od nejelementárnějších vět, až ke složitým periodám, jejichž zákonitosti organizace nejsou dodnes popsány. Nemenší obtíží jsou i další typické rysy (zvláště českých) vět: velmi rozmanitá délka, velká míra volnosti slovosledu, potence formálního nevyjádření některých pozic, obtížná stanovitelnost hranice věty a rozmanitost mentálních projevů a funkcí jednotlivých vět.

John Ries shromáždil ve dvacátých letech minulého století všechny v té době známé definice věty. V práci *Was ist ein Satz?* jich popsal 139 (Ries, 1931: 208–224). O další vymezení se později pokusili např. E. Seidel, D. J. Allerton či B. L. Müller (Grepl, Karlík, 1998: 17). Uveďme si zde alespoň několik ukázkových vlastností věty, v nichž se tyto definice často shodují:

- Každý prvek věty má vztah k ostatním prvkům, neexistuje prvek věty, který by nevstupoval do vztahu alespoň s jedním dalším prvkem téže věty.
- Věta je znak.
- Pořadí prvků ve větě může být využito k vyjádření jejich syntaktické funkce.
- Prvky věty nemusí být vždy vyjádřené formou, ale i bez znakového vehikula mohou zastupovat syntaktickou pozici, nést nějakou funkci ve větě.
- Věta je myšlenka vyjádřená slovy, ucelená ve vztahu k této myšlence (Palek, 1989: 155–8).

Vladimír Skalička nesouhlasí s tím, že formální vlastnosti věty jsou její podstatou, definuje větu obecněji – jako *elementární sémiologickou reakci* (Skalička, 1935). Tato definice podle něj již implikuje fakt, že věta má jakousi nutně normalizovanou formu, není tedy nutné definici stavět na formálních vlastnostech věty (jejichž přesný popis je ztížený výše popsanými skutečnostmi).

Na druhou stranu – není většina uživatelů jazyka i bez přesného vymezení tohoto termínu schopná intuitivně rozpoznat větu? K dosažení našeho cíle jsou zmíněné definice nepoužitelné – příliš abstraktní, neplatící obecně pro všechny věty jazyka (obsahují nezávazné výrazy jako „většinou“, „často“, „někdy“, „moci“, „nemuset“, opatrné formulace s kondicionálovými tvary sloves apod.), pro anotátory, kteří text segmentují na věty, nesrozumitelné, zbytečně komplikované a často nejednotně interpretovatelné. A pokud nezvládá ani člověk podle daných pravidel jednotně rozhodovat hranice vět, nemůže očekávat dobré výsledky od automatu. Proto jsme se zaměřili na sestavení intuitivnějších srozumitelných pravidel pro jednotnější určování hranic vět, i za cenu toho, že definice věty nebude v hraničních případech odpovídat tradičním lingvistickým představám. Nemáme totiž dobré zkušenosti s příliš dlouhými a detailními manuály, interanotační shoda je podmíněna co nejstručnějšími, nejpřehlednějšími a nejjasněji stanovenými pravidly (Grác, 2011).

4 Pravidla pro anotátory

Prvním krokem bylo vytvoření definice termínu *věta*, takové, která by odpovídala výše popsaným nárokům. Zvolili jsme proto velmi intuitivní a čistě formální pohled: *Věta začíná velkým písmenem (před nímž stojí interpunkce) a končí tečkou, vykřičníkem, otazníkem, případně trojtečkou či horními uvozovkami (tedy věta jednoduchá i souvětí, každý autonomní celek).*

Poté jsme věnovali zvláštní pozornost krajním ale zdaleka ne ojedinělým případům, které vykazují z formálního i sémantického hlediska jisté zvláštnosti a stanovili jsme pravidla pro tyto specifické případy. Jimi jsme rozšířili předešlou triviální definici. Tato pravidla se týkala především výrazů a konstrukcí obsažených uvnitř závorek či uvozovek (přímá řeč), či celků oddělených od ostatního textu středníkem, pomlčkou nebo dvojtečkou, různých výčtů apod. Také jsme se zabývali problematikou nadpisů a nevětných i polovětných konstrukcí (větných ekvivalentů jako je třeba oslovení, různé nápisy, samostatné infinitivy apod.).

Náš postup při sestavování anotačního manuálu byl následující:

1. Anotace vět podle intuice,
2. příprava anotačního manuálu,
3. anotace vět podle manuálu,
4. vyhodnocení shody mezi anotátory,
5. opakování 2.–4. kroku podle potřeby,
6. anotace korpusu.

Podívejme se nyní na jednotlivé problémy, pro které bylo třeba stanovit pravidla určování hranic vět. Nezáleží nám primárně na tom, zda jsou správná a zda komplexně postihují všechny aspekty věty, či nikoli. Důležitější je konzistence anotování, jednotnost označkovaných dat.

Jedním ze specifických případů jsou nadpisy, které jsme zařadili mezi samostatné sentence. Anotátory jsme na ně upozornili především proto, že nejčastěji nebývají od okolního textu odděleny interpunkcí. K nadpisům mají blízko i výrazy (ne celé věty) před dvojtečkami. (Při exemplifikaci budeme užívat strukturní značky <s>, </s> pro znázornění hranic mezi větami.)

(1) <s> Barevné oblečení </s>

<s> O barevném oblečení jsem vůbec nepsala, protože to je až další level. </s>

(2) <s> Vlastnosti a přednosti**:** </s>

<s> okamžitě hydratuje pokožku, obsahuje aktivní výtažky z aloe. </s>

(3) <s> Mini-recenze**:**</s>

<s> Jak používat**:**</s>

<s> aplikujte na umyté, ručníkem prosušené vlasy. </s>

Dalším jevem, který naše pravidla zmiňují, jsou výrazy v textu obsažené uvnitř závorek. Pokud jimi je pouze větný člen patřící syntakticky i sémanticky do nějaké klauze, či samostatná klauze patřící do nějaké širší sentence, obsah závorek neoznačujeme jako samostatnou sentenci. Když ale tato samostatná klauze v závorkách nepatří k okolní sentenci (nedá se interpretovat jako věta vedlejší, jedná se o parenteze), je samostatnou sentencí.

(4) <s> K véčkovému výstřihu se někdy dávají ramínka za krk **(**halterneck**)**, ale na to pozor – máte-li kýtovité paže, akorát je zvýrazníte. </s>

(5) <s> Pohostinství U Brejšků je jediným obnoveným podnikem z jakéhosi bermudského trojúhelníku pražských restaurací, tvořeného Unionkou na Národní **(**dnes zde stojí panelák, který patřil nakladatelství Albatros**)**, U Ježíška **(**dnes záchodky metra, stanice Národní**)**, kde se v Haškově době pravidelně ztrácelo množství lidí z českého uměleckého světa. </s>

(6) <s> Tunika na první a druhé fotografii je mírně tvarovaná na tělo pěkně provedenými záševky </s>

<s> **(**žehlí! </s>

<s> ona žehlí!**)** </s>

<s> , takže štíhlá modelka nevypadá jako krabice, přesto nepochybuji o tom, že má v tunice dostatek volnosti (tolik k obhajování beztvarých pytlů řečičkami o pohodlnosti). </s>

Cílem psaní středníku v textu je grafické znázornění hlubšího předělu, než na jaký poukazuje čárka. Do jisté míry tedy osamostatňuje, proto podle nás odděluje dvě sentence. Výjimkou z tohoto pravidla je případ, kdy středník stojí při výčtu mezi souřadně spojenými nevětnými výrazy (člení je na větší skupiny). Celý výčet potom bereme jako součást klauze.

(7) <s> První výmluva zavání názorem, že jakákoli péče o svůj zevnějšek je povrchní**;** </s>

<s> dovolila bych si též rýpnout, že člověk (počítám ženy i muže), který nenosí nic jiného než džíny a bundu a jeho pokusy o styl končí u špičatých bílých kozaček nebo kecek Nike, nemá na demonstrativní pohrdání hadrama nejmenší právo. </s>

(8) <s> Je pravda, že třeba u košil je výběr v zahraničních e-shopech a kamenných obchodech mnohem širší**;** </s>

<s> u běžného oblečení jako trička se zase běžně stává, že v butiku stejné značky mají totéž zboží kvalitnější ve Vídni než v Brně. </s>

(9) <s> Vždy poznáváme člověka určitým způsobem začleněného do společnosti (širší – černoch, Američan, Evropan**;** užší – rodina, kroužek, třída). </s>

Další pravidlo se týká také interpunkčního znaménka – dvojtečky, za níž často následuje výčet (někdy strukturovaný do odrážek). Pokud tedy za dvojtečkou stojí nevětný výčet (neobsahuje verba finita), tvoří tento výčet jednu sentence spolu s „uvozovací větou“ před dvojtečkou. Jestliže se však v textu vyskytne výčet většího rozsahu – obsahující věty, jde v rámci tohoto výčtu buď o jednotlivé klauze jedné sentence (dohromady s „uvozovací větou“), nebo o celé sentence oddělené tečkami (potom je „uvozovací věta“ také samostatnou sentencí). Pokud dvojtečka odděluje samostatnější větší celky (tj. když se dá nahradit tečkou), jsou oba sentencemi.

(10) <s> Co bych si z polyesteru nekoupila**:** trička, halenky a jiné oblečení „na tělo“. </s>

(11) <s> Svou vinu bude mít také zvláštní způsob, jímž autorka postupuje (popisuje ho zde)**:** </s>

<s> nepostupuje od celku k detailům a nevyužívá různé pomocné linie, ale kreslí shora dolů a pečlivě poměřuje vzdálenosti jednotlivých prvků v obličeji. </s>

V případě, že se v textu vyskytne jedna pomlčka (neplnící funkci zámlky, odmlčení se), platí pro ni při rozdělování stejná pravidla jako pro dvojtečku. Dvojice pomlček má v textu funkci podobnou čárkám či závorkám, proto v jejich případě aplikujeme „závorkové“ pravidlo (tzn. patří-li části sentence kolem pomlček syntakticky k sobě a může-li výraz mezi pomlčkami stát mimo tuto sentenci bez sémantické změny propozice, oddělíme ho, jinak jej ponecháme v jedné větě).

(12) <s> Podle něj totiž nikdo nedokázal otravu v potoce ohlásit včas (takže se už nic nezjistí), a co víc **–** rybáři nemají deset tisíc na odebrání a analýzu speciálních vzorků. </s>

(13) <s> Nadstandardní službu **–** výpis dob pojištění na vlastní žádost **–** tedy ČSSZ v minulosti poskytovala nad rámec zákona. </s>

(14) <s> To bylo pro Půlpána, teď jen mezi námi **–** </s>

<s> Půlpáne, nečti! </s>

<s> **–** vůbec se mi tam nechtělo. </s>

(15) <s> O problematice převodů už tady řeč byla **–** </s>

<s> Jirko, děkuji za vzornou reakci! </s>

<s> **–** a Aleš dnes jasně dal najevo, jakou šajbu Kozlíková jezdí. </s>

Ani tečka a trojtečka nemusí vždy signalizovat konec věty. Tečky se píší i ve zkratkách, dále v řadových číslovkách či e-mailových a webových adresách (což by při ručním značkování nemělo činit potíže). Trojtečka zase může být užita uvnitř věty jako zámlka, výpustka či v přerývané řeči.

(16) <s> Pokud řidič sledovaného vozu zabloudí, má možnost prostřednictví dispečinku NASEC zjistit přesnou polohu místa, kde se nachází a nechat se navigovat do jemu známé lokality, příp. cíle své cesty. </s>

(17) <s> Ty lze později využívat pro další zpracovávání, např. pro tvorbu knihy jízd. </s>

(18) <s> Přednáška další členky kolektivu Ing. Jarmily Neugebauerové, Ph.D. ze Zahradnické fakulty, Lednice Mendelovy univerzity, byla věnována tématu pěstování léčivých rostlin, jejich konkrétním nárokům na stanovištní podmínky, na agrotechniku a na problémy při ekologickém způsobu jejich pěstování. </s>

(19) <s> Domácí krutě zaspali začátek a v 17. minutě už se mohli shánět po ručníku. </s>

(20) <s> Výzva pro bývalé členy oddílu juda: </s>

<s> kdo by měl zájem pomáhat s výukou juda jako asistent trenéra nebo studovat jiu-jitsu, dejte vědět na: petras.r@centrum.cz. </s>

(21) <s> Ty jsi ale ... nevím jak to slušně napsat. </s>

(22) <s> Herečka jsem dala schválně do úvozovek....neboť tato žena se k herectví dostala pouze zásluhou svého otce ...jinak na to,aby hrála absolutně nemá.....jako uvaděčka zklamala taky...teda alespoň mě. </s>

Jeden z největších problémů při určování hranic vět představuje přímá řeč. Predikátor věty uvozovací má totiž ve své elementární struktuře (často obligatorní) pozici, kterou naplňuje právě přímá řeč. Proto by právem patřily k sobě, mohli bychom výpověď ohraničenou uvozovkami brát jako zvláštní případ vedlejší věty rozvíjející větu uvozovací. My jsme se přesto nakonec rozhodli oddělovat věty uvozovací jako samostatné sentence, zvláště za účelem možnosti dále na jednotlivé sentence segmentovat také samotný text přímé řeči (v případě delšího promluvového úseku), ale také kvůli jednotnosti značkování.

(23) <s> Přijde paní k obchodníkovi a říká: </s>

<s> „Chtěla bych si objednat 3, co umí plavat a 2, kteří umí hrát na kytaru.“ </s>

(24) <s> „Musíte to vidět, musíte vidět, jak tam dát tu ruku, musíte si to osahat,“ </s>

<s> upozornil Krška na nutnost zkusit si tuto metodu v praxi. </s>

(25) <s> „Do týdne bude vypsáno standardní výběrové řízení a do měsíce budou výsledky,“ </s>

<s> řekl ČTK Šnajdr. </s>

<s> „Zatím bude pojišťovnu řídit dosavadní statutární zástupce a člen managementu. </s>

<s> To znamená jasnou kontinuitu, nic se neděje,“ </s>

<s> zdůraznil. </s>

(26) <s> Tady byl v mém případě evokován biblický příběh o stvoření: </s>

<s> „I řekl Bůh: </s>

<s> »Buď světlo!« </s>

<s> A bylo světlo.“ </s>

V neposlední řadě naše pravidla upozorňují anotátory na možnost výskytu některých interpunkčních znamének v závorkách. Ty pak mají jinou funkci, než je oddělování větných celků (standardně zdůrazňování předcházející lexikální jednotky, v případě otazníku vyjádření pochybnosti, nejistoty), proto za nimi neznázorňujeme hranice vět.

(27) <s> Popadla tedy své osobní věci, odstrčila dva **(!)** policisty, kteří jí stáli v cestě, a vyskočila z okna v prvním patře. </s>

(28) <s> Nýt rozhodně nebyl ve spáře a někdo **(?)** ho spolu asi se 3 skobami po pár týdnech ukradl. </s>

Nakonec jsme věnovali pozornost problematice eliptických výpovědí, dodatečně připojených osamostatnělých větných členů, různých nápisů a samostatných infinitivních konstrukcí. Všechny tyto určujeme jako oddělené sentence, jelikož splňují nárok na formální autonomnost.

(29) <s> Ležérní saka? </s>
<s> Pouze manšestrová. </s>

(30) <s> Nekouřit! </s>

(31) <s> Velmi dobře. </s>

(32) <s> Jak je od začátku jasné, je osudem, aby se Josseran s Chutlun střetli. </s>
<s> Rytíř Templu a tatarská princezna. </s>
<s> Vděčný námět. </s>

(33) <s> Hustě sněžilo a párkrát to houplo. </s>
<s> Nic příjemného. </s>

(34) <s> No, někdy prostě nemůžeme mít všechno, co bychom si přáli. </s>
<s> Bohužel. </s>

(35) <s> Rodinný miláček. </s>
<s> Prodej z časových důvodů. </s>
<s> Cena dohodou. </s> (Baisa, Suchomel, 2012; czTenTen12, czes)

5 Závěrem

Samotný anotační proces probíhal tak, že v první fázi dostali anotátoři korpusové texty obsahující automaticky (naším starým systémem) předznačkované hranice vět. Každá věta byla ověřována dvěma anotátory. Interanotační shoda poté činila 96,07 %. Následně byly hranice vět, na nichž se anotátoři neshodli, znovu označkovány jinými anotátory.

Takto jsme vytvořili ručně označkovaná data, která sice nepopisují věty jazyka ve všech případech způsobem, který je intuitivní pro lingvisty, jsou ale konzistentnější (mají vyšší shodu mezi anotátory). To má příznivý dopad na jejich využití v oblasti strojového učení, resp. ověřování kvality nástrojů automatického rozdělování na věty. Důležitým přínosem bylo i vyzkoušení možnosti tvorby jednoduchého anotačního manuálu iteračními metodami, protože v minulosti byly tyto metody ověřované jen na jiných jazykových rovinách.

REFERENCE

BAISA, V. – SUCHOMEL, V. (2012): Detecting Spam in Web Corpora. In: Aleš Horák, Pavel Rychlý, *6th Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Tribun EU. 69–76.

GRÁC, M. (2011): Case study of BushBank concept. In: *PACLIC 25th Pacific Asia Conference on Language, Information and Computation*. 353–361.

GREPL, M. – KARLÍK, P. (1998): *Skladba češtiny*. 1. vyd. Olomouc: Votobia.

JULINEK, R. (1999): *Automatická detekce hranic vět* [online]. [cit. 2012-10-27]. Diplomová práce. Brno: Masarykova univerzita, Fakulta informatiky. Vedoucí diplomové práce doc. PhDr. Karel Pala, CSc. Dostupné z: <http://is.muni.cz/th/3500/fi_m/>.

KISS, T. – STRUNK, J. (2006): Unsupervised Multilingual Sentence Boundary Detection. In: *Computational Linguistics*. Cambridge: MIT Press Cambridge, roč. 32, č. 4, 485–525.

O'NEIL, J. (2012): *Doing Things with Words, Part Two: Sentence Boundary Detection.* [online]. [cit. 2012-10-13]. Dostupné z: <http://www.attivio.com/blog/57-unified-information-access/263-doing-things-with-words-part-two-sentence-boundary-detection.html>.

PALEK, B. (1989): *Základy obecné jazykovědy*. 1. vyd. Praha: Státní pedagogické nakladatelství.

PALMER, D. D. (1995): SATZ – An Adaptive Sentence Segmentation System. In: *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing.* Berkley, California: University of California, 78–83.

PALMER, D. D. – Hearst, M. A. (1997): Adaptive Multilingual Sentence Boundary Disambiguation. In: *Computational Linguistics*. Cambridge: MIT Press Cambridge, roč. 23, č. 2, 240–267.

REYNAR, J. C. – RATNAPARKHI, A. (1997): A Maximum Entropy Approach to Identifying Sentence Boundaries. In: *ANLC '97 Proceedings of the fifth conference on Applied natural language processing.* Stroudsburg: Association for Computational Linguistics, 16–19.

RIES, J. (1931): *Was ist ein Satz? (Beiträge zur Grundlegung der Syntax III)*. Praha: Taussig & Taussig.

SKALIČKA, V. (1935): K problému věty. In *Slovo a slovesnost*. Praha: Ústav pro jazyk český Akademie věd České republiky, roč. 1, č. 4, 212–215.

SEDLÁČEK, R. a SMRŽ, P. (2001): A New Czech Morphological Analyser ajka. In: *Proceedings of the 4th International Conference TSD 2001*. Berlin: Springer-Verlag, 100–107.

Wikipedia (2012): *Sentence boundary disambiguation*. [online]. [cit. 2012-10-13]. Dostupné z: <http://en.wikipedia.org/wiki/Sentence_boundary_disambiguation>.