Oponentský posudek práce

ROZŠÍŘENÁ TEXTOVÁ KOREFERENCE A ASOCIAČNÍ ANAFORA

(Koncepce anotace českých dat v Pražském závislostním korpusu)

předložené k disertačnímu řízení na

Filozofické fakultě Univerzity Karlovy

Annou Nedoluzhko

Obecné poznámky k zadání disertační práce a k obtížnosti zadaného úkolu

Záměrem předložené disertační práce bylo – pokud dobře rozumím – vypracovat podrobný přehled jevů českého jazyka vázaných na ty případy koreference a anafory, které nepodléhají přímo gramatickým pravidlům českého jazyka, a na tomto základě navrhnout a (alespoň částečně) prakticky ověřit pravidla, jimiž by se řídila anotace koreferenčních a anaforických vztahů tohoto typu v Pražském závislostním korpusu.

Složitost a neurčitost obecných koreferenčních a anaforických vztahů (resp. jejich závislost na mimojazykových faktorech) je jedním z hlavních problémů, kvůli kterým je velmi těžké anotovat takové vztahy automaticky. Anotace rozsáhlých jazykových dat je přitom velmi podstatná podmínka jak pro jakýkoliv skutečně objektivní výzkum takových vztahů, tak pro praktické aplikace (za všechny je možno typicky jmenovat strojový překlad, který se bez správného určení koreferenčních a anaforických vztahů ve zdrojovém textu vůbec nemůže obejít): ať už tedy v aplikacích uvažujeme o metodách strojového učení, nebo o metodách symbolických, je dostupnost anotovaných jazykových dat zcela zásadní otázkou.

Pokud předchozí fakta stručně shrneme, je možno konstatovat, že úkol, před kterým disertantka stála, byl značně náročný, jeho splnění by však přineslo výrazný pozitivní efekt jak pro jazykovědný výzkum, tak pro jeho aplikace.

Vlastní posudek

Předtím, než se budu (pohříchu stručně) věnovat obsahu práce, bych se chtěl podrobněji vyjádřit k její formální stránce.

Je mi samozřejmě známo, že disertantka není rodilou mluvčí češtiny a ani neprošla systémem českého základního a středního školství, pokládám nicméně za nutné se v posudku její práce i přesto zabývat jazykovou úrovní předloženého textu, a to zejména z toho důvodu, že i při vší toleranci k pochopitelné disertantčině neznalosti složitých zákoutí českého jazyka budí tento text bohužel dojem velmi nedbalého zpracování (je např. zřejmé, že k finální redakci nebyl využit ani tak jednoduchý nástroj, jako je spelling-checker, nemluvě o nástrojích dokonalejších nebo o jistě možné, proveditelné a pro kvalitní výsledek v podstatě nevyhnutelné kontrole jazykové správnosti rodilým mluvčím). Můj odhad počtu jazykových (nikoliv obsahových) problémů se

tak v průběhu čtení ustálil na cca 4 (jazykových) chybách na stránku, což je opravdu podstatně více, než pokládám za možné tolerovat (i pokud vezmu v úvahu uvedené okolnosti).

Na asi nejjednodušší úrovni obsahuje text opravdu velmi dlouhou řadu překlepů či dalších triviálních chyb. Mezi ty, které by bylo možno opravit prostým spell-checkingem, patří například (vybírám z mnoha, číslo v závorce vždy odkazuje na stranu) *počitač* místo *počítač* (11), *sekretařku* místo *sekretářku* (13), *třida* místo *třída* (31), *Rasputín* (sic!) místo *Rasputin* (34), *pojmenováaní* místo *pojmenování* (41), *spectrum* místo *spektrum* (46), *ztrátí* místo *ztratí* (53), *prácích* místo *pracích* (55), *stávájící* místo *stávající* (76, dvakrát), *tekrogramatickou* místo *tektogramatickou* (99), *koreferefovaných* místo *koreferovaných* (102), mezi neopravitelné spelling-checkingem patří např. *stráně* místo *straně* (1), *vetší* místo *větší* (37), *mezí* místo *mezi* (105), *měřící* místo *měřicí* (115, na stejné straně je ovšem totéž i správně, jen o 2 řádky níže). Podobné chyby jsou ale i v angličtině a v ruštině: *principals* místo *principles* (xv), *preceeding* místo *preceding* (7), *referencial* místo *referential* (282), *Sopostavitelnaja* místo *Sopostavitel'naja* (273), *kultury* místo *kul'tury* (282) atd. atd.

Na úrovni o něco vyšší obsahuje předložená disertační práce řadu hrubých chyb ve shodě, a to ve všech typech gramatické shody, které v češtině existují: *výběrem projektů, které nám posloužili* (58), *svůj hyperonymum* (120), *anafora, jejíž klasifikace* (53), což se týká dokonce i příkladů převzatých z anotovaných textů *dítě ... se nám ozval* (118).

Velmi problematická je interpunkce – kladení čárek: to odpovídá daleko spíše úzu anglickému než českému (příklady neuvádím, jsou jich ale desítky).

Za problematické pokládám velké množství rusismů, z nichž řadu nelze kvalifikovat jinak než jako skutečně brutální; mezi nimi nad jiné vyniká mnohonásobně použité prepozitivní postavení rozvinutého shodného přívlastku: *odkaz na již uvedený v předcházejícím textu antecedent* (4), *vypracované pro účely extrakce informace* (37), *všechny problematické z tohoto hlediska jmenné fráze* (45) a další. O mnoho přívětivější k českému čtenáři nejsou ani celé „ruské" lexikální konstrukce hovořící např. o *nakládání negace* na větu (29), mezi další patří *v rámcích* místo *v rámci* (např. na str. 27, ale i jinde). Jiné, méně nápadné rusismy, jsou pak naopak o to nebezpečnější, že svou zdánlivou "českostí" zastírají skutečný smysl textu, např. používání slova *význam* namísto patřičného *hodnota* (76 i jinde), *představit* místo *uvést* (87) apod., a čtenáře tak matou.

Kromě čistě jazykových chyb trpí práce i „spřízněnými" chybami v užívání terminologie.

Nejčastějším problémem je nejasné používání rodu u zcela základního pojmu *anafora* (o které se ovšem většinou hovoří jako o *anaforu*, tedy v mužském neživotném rodě, někdy je ale *anafor* i životný – viz *referent nepřímého anafora* (7)), ale i u slov dalších: *antecedent daného kandidátu* (místo ... *kandidáta*) atd.

Chybná (a bez znalosti angličtiny nesrozumitelná) je terminologie ohledně *definitivní jmenné fráze* (69) a *určené jmenné fráze* (72) – obojí zřejmě místo *určitých jmenných frází*, podobně nesprávná je ovšem i *definitivní deskripce* (6).

Nesprávně jsou používána adjektiva v řadě spojení typu *otázkové slovo* (99), *číslové výrazy* (44), zcela nejasná je věta na str. 95 *Textové koreference se mohou zúčastnit výrazy v asertivních, otázkových, rozkazovacích i negovaných větách.*, kde se jednak překvapivě uvádí výčet kategorií navzájem nezávislých (rozkazovací věta může být, ba často dokonce je negativní, např. *Nechoď tam !*), zejména však není jasné, co se rozumí pod pojmem *asertivní věta* (když už pomineme výše zmíněnou záměnu náležitého *tázací* za *otázkový*). Nejasná je i terminologie a klasifikace adjektiv na str. 80 a 81, kde jsou sice v textu adjektiva rozdělena do skupin, ale všem z nich je pak na formální rovině připsána hodnota (v disertantčině formulaci *význam*) atributu

sempos = *adj.denot*, a není tak tedy jasné, proč byla klasifikace (která následně nemá žádné formální vyjádření) zavedena.

Stále snad v oblasti terminologické se s disertantkou neshoduji v označení spojení *an Edinburgh jeweller* jako apozice (45).

Několik vět je vůbec nesrozumitelných, např. *Oddíl III.4.2.4. je věnován problematice správného určování antecedentů a obsahuje veškeré konvence a rozhodnutí výběru.* (98) nebo *Její základní myšlenka spočívá v tom, že na rozdíl od predikátů, které nemají vlastní referenci, celé propozice (včetně svých aktantů), a to na situaci, kterou pojmenovávají.* (121).

Potíže se srozumitelností textu mám ale i v oblastech, které na problémy s jazykovou správností pouze volně navazují nebo se s ní překrývají.

Na str. 66 pokládám za nesrozumitelnou formulaci, ve které se praví, že *se subjekty* (z kontextu se důvodně domnívám, že jsou míněny skutečně větné podměty, a to v italštině) ... *používají predikativně.* Přestože italštinu ovládám, jak doufám, dosti obstojně, uvedenou formulaci nechápu, resp. nejsou mi známy žádné případy italských konstrukcí, kde by podmět věty byl použit současně jako její přísudek

Na str. 74 se píše o *substantivech s odděleně reprezentovaným příznakem negace.* Velmi pravděpodobně jde o špatný překlad ruského slova *otdel'no* (správně by mělo zřejmě být uvedeno české *samostatně*), situaci však ještě dále komplikuje o řádku níže uvedené tvrzení *Do této skupiny* (tj. do skupiny substantiv s odděleně reprezentovaným příznakem negace) *patří deverbativní substantiva zakončená na -ní/-tí (hlasování).* Kde se ale u substantiva *hlasování* vyskytuje příznak negace, resp. jak je od něj takový příznak oddělen a kde v takovém případě stojí, není dále vysvětleno.

Často nepřesné či jinak chybné jsou překlady cizojazyčných příkladů, a to kupodivu zejména z ruštiny: správný překlad příkladu (19) na str. 13 by zřejmě měl být ... *s jakoukoliv/libovolnou cizinkou* ..., a dále jde přinejmenším o překlady příkladu (44) na str. 27, příkladů (60) a (64) na str. 30 a příkladů (77), (78) a (79) na str. 125.

Nejasné je použití nikde předtím nedefinovaného pojmu termínu *kontejner* (63), které je ovšem jenom příkladem toho, že srozumitelnost práce vskutku masivně trpí četnými odkazy na pojmy zavedené podstatně později, než jsou použity. Podobně nevhodná, ba ke čtenáři přímo nepřátelská je volba slova *Martin* (74) jako příkladu přídavného jména přivlastňovacího (byť formálně se skutečně o takové přídavné jméno v jednom z významů slova jedná).

Problematická je v některých případech exemplifikace disertantčina řešení koreference: tak např. na str. 65 je v příkladu (6) koreferováno zájmeno *to* na podstatné jméno *Němci* v předchozí větě, je však zřejmé, že toto zájmeno vůbec koreferováno být nemá (nebo by mohlo být případně – ale velmi nepravděpodobně – koreferováno k výrazu *děti).*

Zásadně nesouhlasím s tvrzením disertantky na str. 60 *V případě gramatické koreference je možnost rovnocenného výběru mezi více antecedenty pro jeden anaforický člen prakticky vyloučena – gramatická pravidla jazyka (z definice gramatické koreference) předurčí pouze správný antecedent.* Takovéto razantní tvrzení je nejen zcela chybné (viz mnohočetné případy, kdy gramatická pravidla triviálně neurčují antecedent jednoznačně – typ *obhájci rodičů týraných dětí, které jsme tam potkali*: vztažné zájmeno *které* se může vztahovat k *obhájcům, rodičům* nebo *dětem*, o vyloučení rovnocenného výběru nemůže být ani řeč), ale trivialitou své nesprávnosti vzbuzuje, a to velmi silně, dojem, že disertantka o problematice gramatické koreference vůbec nepřemýšlela.

Takto bych mohl pokračovat ještě dlouho, předpokládám však, že jako ilustrace toho, že práce je – přinejmenším – obtížně srozumitelná, výše uvedený výčet problémů postačuje.

Pokud se – s přiznaným omezením svého porozumění – mám dále obecně vyjádřit k obsahové úrovni práce, pak bych svůj postoj shrnul asi takto:

- disertantka vložila do práce nepochybně velké úsilí a jistě jí obětovala i značný čas a dokázala tak, že má dostatek trpělivosti, houževnatosti i cílevědomosti,
- vlastní úroveň práce však zanechává nejasnosti v tom, zda jde v případě této disertace o vskutku přesvědčivou ukázku promyšlené práce vědecké: kromě již uvedených nepřesností formulačních a kromě toho, že práci pokládám za skutečně obtížně srozumitelnou, nejsem plně přesvědčen o tom, že disertantka naplnila vědeckovýzkumný záměr definovaný na str. 257 slovy *The purpose of this thesis is to describe the theoretical basis of annotation of coreference and the bridging anaphora in the Prague Dependency Treebank.* Práce je do značné míry totiž „pouhým" manuálem co a jak anotovat: problém takového vypracování disertace vidím pak zejména v tom, že (dle mého porozumění cílům práce) by takový anotační návod měl být založen na podrobné studii a popisu koreferenčních a anaforických jevů v češtině. Pokud jsem schopen to posoudit, takovou studii ale předložená práce neobsahuje, nebo přinejmenším neobsahuje žádnou systematickou studii české anafory a koreference. Úvod práce je sice věnován přehledu koreference či její anotace v jiných projektech, ale jde o popisy či projekty zpracované na materiálu ruštiny, angličtiny, italštiny a němčiny. Problematikou jevů koreference a anafory v češtině (a jejich popisu v odborné literatuře) se ani v úvodu, ani nikde jinde předložená disertace nijak podrobně a zejména nijak systematicky nezabývá, cituje jen několik všeobecných prací. Části práce, která se věnuje samotnému popisu anotačních postupů, pak rozumím tak, že je v ní sice prezentována jistá klasifikace koreferenčních a anaforických konstrukcí, tato klasifikace však není založena na výzkumu českého jazykového materiálu. Dle mého názoru totiž nikde v práci není uvedeno, že disertantka skutečně provedla korpusový výzkum a na něm založenou „inventarizaci" českých anaforických a koreferenčních konstrukcí, kterou by bylo lze označit za vyčerpávající nebo alespoň reprezentativní (a že tento výzkum skutečně proveden nebyl, soudím předběžně z toho, že v disertaci nejsou vůbec zmíněny ani některé konstrukce diskutované např. v práci *Hajičová, Oliva, Sgall: Odkazování v gramatice a v textu, in: Slovo a slovesnost, vol. 48, pp. 199-212, Praha, 1987*, ba tato práce není ani citována v literatuře, což je zarážející minimálně už proto, že je uvedena např. v seznamu obecné doporučené literatury pro magisterské a doktorské studium na webové stránce disertantčiny „mateřské" instituce, viz http://ufal.mff.cuni.cz/literatura.html). Domnívám se proto, že klasifikace, o kterou se v hlavní části disertace uvedený „anotační manuál" opírá, je založena na výzkumu, zpracování a klasifikaci koreferenčních a anaforických jevů v jiných jazycích, než je čeština, což – pokud je to pravda – pokládám za velmi závažný problém předložené disertační práce, a to nejen v rovině teoretické, ale zejména v rovině praktické: anotovat češtinu na základě zjištění o (např.) angličtině je jednak teoreticky a metodicky špatně, a jednak a hlavně to k vyhovující anotaci vůbec nemůže vést.

Závěrem si tedy dovoluji poněkud neradostně shrnout, že po formální stránce je předložená práce vyhotovena podle mého soudu nedostatečně pečlivě a ani její obsah mne nepřesvědčil o disertantčině schopnosti skutečně vědecky pracovat (a výsledky své práce i odpovídajícím způsobem prezentovat). Je samozřejmě možné, že jsem práci nebyl schopen plně porozumět, a proto – v kombinaci s výše uvedenými výtkami – se vzdávám jakýchkoliv doporučení – pozitivních i negativních – ohledně toho, zda by disertantce na základě uvedené práce měl či neměl být Univerzitou Karlovou udělen titul Philosophiae Doctoressa, ve zkratce Ph.D.

V Praze 27. května 2010

karel oliva