„Bohemistyka” 2013, nr 1, ISSN 1642–9893

Magdaléna RYSOVÁ
Praha

Jazykové prostředky vyjadřující textové vztahy v češtině a jejich zpracování v Pražském závislostním korpusu

Klíčová slova: Diskurz, konektory, Pražský závislostní korpus, textové vztahy, jazykové prostředky
Keywords: Connectives, discourse, Prague Dependency Treebank, text relations, linguistic means

Abstract

The paper discusses by which language means it is possible to express certain discourse relations in Czech. The analysis is based on the annotated data from the Prague Dependency Treebank (PDT) and deals with the language expressions that are not considered classic connectives (i.e. mainly conjunctions and adverbs as *therefore*, *and*, *but* etc.). Examples of such expressions (called alternative lexicalizations of discourse connectives or AltLex's) are *the reason is*, *with the justification that*, *simply speaking* etc. The paper presents the comparison of Czech and English AltLex's and their semantic characteristics – it analyzes their possibility of expressing an anaphoric reference. Finally, the paper demonstrates how discourse may be interlinked with coreference, as new Czech AltLex's in PDT may be found and annotated on the basis of their annotation under coreference.

Příspěvek se zabývá otázkou, jaké jazykové prostředky v češtině mohou vyjadřovat diskurzní vztahy v textu. Výzkum proběhl na základě analýzy anotovaných dat Pražského závislostního korpusu (PDT) a byl zaměřen na jazykové výrazy, které v pojetí PDT nejsou chápány jako klasické textové konektory (tj. především spojky a příslovce typu *proto*, *a*, *ale* atd.). Tyto prostředky jsou v odborné literatuře nazývány alternativní lexikální vyjádření diskurzních konektorů (zkráceně altlexy), příkladem jsou vyjádření jako *důvodem je*, *s odůvodněním*, *jednoduše řečeno* atd. Příspěvek přináší sémantickou charakteristiku českých altlexů (tj. analyzuje jejich možnost vyjadřovat na povrchové syntaktické rovině anaforickou referenci) a jejich srovnání s obdobným výzkumem pro angličtinu. Příspěvek dále poukazuje na překrývání hranic koreference a diskurzu (tj. na možnost automatického vyhledávání některých altlexů v korpusu na základě anotace koreference).

Soudržnost a srozumitelnost jsou základními atributy textu, jejichž zásluhou čtenáři a posluchači rozumějí psaným i mluveným komunikátům[1]. Jednotlivé jednotky koherentního textu jsou provázány explicitními i implicitními vztahy. Jak dokazují mnohé studie (srov. např. Irwinová a Pulverová 1984), lidé lépe rozumí textovým vztahům vyjádřeným explicitně (tj. pomocí konkrétních jazykových výrazů). Je proto třeba těmto jazykovým prostředkům, které se podílejí na soudržnosti a srozumitelnosti textu, věnovat pozornost.

Jazykové výrazy s konektivní funkcí jsou například zpracovávány ve stromových strukturách textů Pražského závislostního korpusu (PDT) v rámci anotace textových vztahů (tj. vtahů mezi nadvětnými celky).

Cílem tohoto článku je stručně popsat, jakým způsobem jsou v PDT textové vztahy zachycovány, a poté přiblížit jeden typ jejich explicitního vyjadřování – tzv. altlexy (tj. jazykové prostředky typu *důvodem je*; *to je důvod*, *proč* atd.).

V první fázi anotace diskurzních vztahů v PDT byly zachycovány pouze vztahy uvozené explicitními konektory (konektory jsou zde definovány jako prostředky patřící mezi určité slovní druhy – především spojky, částice, příslovce, některá užití zájmen typu *kromě toho*, srov. Mladová a kol. 2011). Ukázalo se však, že to není jediný způsob explicitního vyjadřování vztahů v textu. Na další typ prostředků vyjadřujících textové vztahy upozornili tvůrci jiného diskurzního korpusu – pensylvánského korpusu Penn Discourse Treeebank (Prasadová a kol. 2008). Jedná se o pokračování projektu Penn Treebank (PTB), tj. anglického syntaktického korpusu. Korpus Penn Discourse Treebank (PTB) poskytuje uživatelům anotaci diskurzních vztahů a obsahuje přibližně 49 000 vět. Texty přitom nejsou zpracovávány ve stromových strukturách (jako v Pražském závislostním korpusu), ale lineárně.

[1] Tento příspěvek vznikl za podpory Vnitřního grantu Filozofické fakulty Univerzity Karlovy v Praze „Diskurzní vztahy v textu” (VG146).

V průběhu anotací diskurzní roviny v pensylvánském korpusu Penn Discourse Treebank upozornili Prasadová a kol. (2010) na širokou škálu vyjádření, kterou nazvali alternativní vyjádření konektorů[2]. Jednalo se o takové jazykové prostředky, které v textu mají stejnou funkci jako (výše vymezené) konektory. Tato vyjádření stejně jako konektory signalizují, že dané dva slovesné argumenty jsou v určitém diskurzním vztahu, ale jejich lexikálně-syntaktická povaha se od konektorů liší. Příkladem takovýchto českých prostředků (altlexů – alternativních lexikálních vyjádření diskurzních konektorů) jsou např. spojení *hlavním důvodem je…*; *odůvodnil to tím, že*; *důsledkem jejich odjezdu může být…* U nás na existenci podobných jazykových prostředků upozorňují zejména Hrbáček (1994) a Hoffmannová (1983). Pokud tedy chceme souborně obsáhnout všechny prostředky podílející se na výstavbě textu, nemůžeme se omezovat pouze na výrazy z určitých tříd slovních druhů.

Příspěvek si proto klade za cíl prozkoumat, jaké jazykové prostředky (vedle „klasických" textových konektorů typu *a proto*, *ale*, *a*) mohou v češtině vyjadřovat diskurzní vztahy v textu. Představuje sémantickou klasifikaci českých altlexů na základě analýzy anotovaných dat Pražského závislostního korpusu (PDT). Jeho předmětem je také srovnání českých altlexů vyskytujících se v PDT a anglických altlexů z PDTB (anotovaného pensylvánského korpusu Penn Discourse Treebank).

[2] Jsme si vědomi toho, že tento termín není pro označení konektivních prostředků zcela vhodný. Slovo „alternativní" implikuje, že altlex je alternativou ke konektoru. Jak ovšem sami autoři americké studie (Prasadová a kol. 2010) uvádějí, existují i takové vztahy, pro které daný jazyk nemá příslušný konektor – srov. např. vztah příčiny a důsledku, který může být v angličtině vyjádřen pouze altlexem, jako je *a major reason is* (*hlavním důvodem je*), nikoli konektorem (srov. Prasadová a kol. 2010). Ne zcela adekvátní může být také slovo „lexikální". Ukazuje se totiž, že se do vyjadřování textových vztahů zapojují i některé gramatické kategorie ve spojení s pozicí ve větě (srov. např. instrumentál na začátku věty – *cílem je*; *důvodem je*; *podmínkou je*). Z těchto důvodů tedy není pojem alternativní lexikální vyjádření diskurzních konektorů zcela vyhovující. V našem článku však toto pojmenování přesto zachováváme, protože se jedná o označení již poměrně zavedené v odborné zahraniční literatuře.

Detailní zpracování těchto jazykových prostředků (tj. výrazů typu *to je důvod*, *proč*; *kvůli těmto skutečnostem*; *stručně řečeno*) bylo podníceno obdobným výzkumem na pensylvánské univerzitě University of Pennsylvania (srov. Prasadová a kol. 2010).

1. Alternativní vyjádření konektorů v odborné literatuře

Alternativní vyjádření konektorů je termín převzatý ze studie *Realization of Discourse Relations by Other Means: Alternative lexicalizations* (Prasadová a kol. 2010). Autoři v ní popisují, jak zjistili existenci těchto vyjádření při anotaci jazykových dat v PDTB. V první fázi anotace anotátoři označovali pouze takové diskurzní vztahy, které byly signalizovány explicitními konektory – ty byly vymezeny jako výrazy patřící mezi vybrané slovní druhy (souřadicí a podřadicí spojky, předložkové fráze, adverbia). V další fázi byly anotovány vztahy implicitní, tj. dané výpovědi nejsou propojeny žádným explicitním jazykovým výrazem a vztah mezi nimi vyplývá pouze z významu obou argumentů. Explicitní diskurzní vztah najdeme v příkladu (1), srov.:

(1) Mám velký hlad. A proto už jdu na oběd.

Naopak diskurzní vztah u stejných výpovědí bez konektoru a proto chápeme jako implicitní, srov.:

(2) Mám velký hlad. Už jdu na oběd.

Ačkoli se mezi výpověďmi v příkladě (2) nevyskytuje explicitní spojovací výraz, z významu obou vět je zřejmé, že jde o vztah příčiny a důsledku.

Anotátoři Penn Discourse Treebanku zároveň k příslušnému implicitnímu vztahu doplňovali, který konektor by vztahu mezi danými argumenty nejlépe odpovídal (autoři je nazvali konektory implicitní). Stávalo se ovšem, že anotátoři mezi příslušnými argumenty viděli diskurzní vztah, ale nebyli schopni k nim přiřadit odpovídající konektor, protože by vzniklé výpovědi „nezněly dobře". Autoři studie tyto případy blíže zkoumali a zjistili, že vzniklé výpovědi „nezní dobře" kvůli

tomu, že diskurzní vztah je v nich již signalizovaný jiným, alternativním vyjádřením. Pokud bychom mezi dané argumenty ještě vložili příslušný konektor, diskurzní vztah by v nich byl signalizován dvakrát. Dochází zde tedy ke zdvojení signalizace diskurzního vztahu. Tuto skutečnost můžeme ilustrovat na příkladu z PDT, srov. příklad (3).

(3) Téměř každý vytěžený diamant má kvalitu drahokamu, a *to je důvod, proč* se tak nákladná těžba z moře firmě vyplácí.

Mezi danými argumenty je vztah příčiny a důsledku signalizovaný spojením *to je důvod, proč*, které podle kritérií PDTB nepatří mezi textové konektory. Pokud bychom mezi dané argumenty vložili ko- nektor typický pro daný vztah (*proto*), vznikla by následující výpověď:

(4) Téměř každý vytěžený diamant má kvalitu drahokamu, a *proto je to důvod, proč* se tak nákladná těžba z moře firmě vyplácí.

Z příkladu (4) je patrné, že vztah příčiny a důsledku je mezi argumenty signalizován dvakrát. Příslušný konektor je nadbytečný, protože jeho funkci ve výpovědi již plní jiné spojení. Toto spojení a další podobné výrazy autoři studie nazvali alternativní vyjádření konektorů (tzv. altlexy). *Realization of Discourse Relations by Other Means: Alternative lexicalizations* (Prasadová a kol. 2010) je první studie věnovaná analýze anglických altlexů. Nicméně zmínky o tom, že diskurzní vztahy mohou být uvozeny různými jazykovými prostředky, nacházíme i jinde, srov. např. Hoffmannová (1983), která pod pojmem *konektory* chápe jakýkoli jazykový výraz či spojení signalizující vztah v rámci textu. O existenci koherenčních jazykových prostředků z jiných tříd, než jsou konektory typu *proto*, *ale*, *a* atd., se zmiňuje také Eugene Winter ve své studii *A clause-relational approach to English texts: A study of some predictive lexical items in written discourse* (1977). Winter zde uvádí, že lexikální jednotky z otevřených tříd (jako jsou podstatná jména, slovesa či přídavná jména) mohou mít funkci ukazatelů mezivětných vztahů a mohou mít vliv na organizaci a strukturaci textu. Z dalších autorů zabývajícími se diskurzními vztahy můžeme jmenovat např. Diane Blakemoreovou (2002). Autorka ve své knize popisuje různé jazykové prostředky, kterými lze vyjádřit některé vztahy, např. vztah přeformulování. Za diskurzní konektory přitom považuje výrazy jako *in other words* (*jinými slovy*) nebo *that is* (*to je*) a jako sémanticky komplexní protějšek k nim uvádí např. *to put it in other words* (*řečeno jinými slovy* / *abych to vyjádřil jinými slovy*). Je tedy patrné, že myšlenky o kategorizaci různých způsobů, jak vyjádřit diskurzní vztahy, se objevují v několika pojetích. Někteří autoři (srov. např. Hoffmannová 1983) tyto prostředky nazývají souhrnně konektory, jiní (srov. např. Prasadová a kol. 2010) pracují se dvěma kategoriemi: konektory a jejich alternativní vyjádření. Vzhledem k tomu, že tento příspěvek přímo navazuje na studii o anglických altlexech (Prasadová a kol. 2010), pojem *alternativní lexikální vyjádření diskurzních konektorů* zachováváme i zde.

2. Alternativní vyjádření konektorů v Pražském závislostním korpusu (PDT) a pensylvánském korpusu Penn Discourse Treebank (PDTB)

Anotace explicitních konektorů a jejich alternativních vyjádření se v pojetí pensylvánského korpusu Penn Discourse Treebank a Pražského závislostního korpusu liší. Hlavním rozdílem mezi oběma pojetími je, v rámci jakých slovních druhů jsou konektory vymezovány (srov. Penn Discourse Treebank – souřadicí a podřadicí spojky, předložkové fráze, adverbia; Pražský závislostní korpus – souřadicí a podřadicí spojky, adverbia, částicové výrazy, některá užití zájmen, některé ustálené víceslovné konektivní prostředky vzniklé spojením různých výrazů, písmenné nebo číselné prvky pro vyjádření vztahu výčtu; srov. Mladová a kol. 2011), ale především také v metodě, jakou při vyhledávání konektorů v textech autoři použili.

Autoři projektu Penn Discourse Treebank (Prasadová a kol. 2008) nejprve vytvořili seznam anglických konektorů, na základě kterého byly konektory v textech vyhledávány a označovány. Anotátoři ovšem v průběhu anotací nacházeli i jiné výrazy se stejnou funkcí, které ovšem na seznamu nebyly. Pro tyto výrazy proto byla zavedena zvláštní kategorie s označením alternativní lexikální vyjádření diskurzních konektorů. Zavedení dvou kategorií (konektory vs. altlexy)

tedy vyvstalo z potřeb praktických anotací, resp. bylo dáno prvotní existencí seznamu konektorů a potřebou odlišit spojovací prostředky na seznamu neuvedené. Pro anotaci Pražského závislostního korpusu předem žádný seznam českých konektorů sestaven nebyl (srov. Mladová a kol., 2011). Anotátoři tedy dostali za úkol diskurzní konektory sami v textu nacházet a označovat, aniž by přitom používali předem daný seznam. Zároveň měli vyhledávat i alternativní vyjádření konektorů, tj. převážně víceslovná spojení s konektivní funkcí. Anotátoři tedy nedostali předem striktní definice konektorů a altlexů, ale interpretace daných vyjádření byla nechána na jejich uvážení. Cílem bylo nashromáždit dostatečný materiál výrazů, které anotátoři neinterpretují jako explicitní konektory, ale které mají podle nich podobnou funkci, a posléze rozhodnout, zda rozlišení konektorů a altlexů má své opodstatnění. Vzhledem k tomu, že v současné době – pokud je nám známo – neexistuje komplexní charakteristika jazykových prostředků, které mají konektivní diskurzní funkci, ale nepatří mezi „tradiční" slovní druhy, které jsou u konektorů nejčastěji zmiňované (tj. především spojky a příslovce), a vzhledem k tomu, že tuto charakteristiku si náš příspěvek klade přímo za cíl, zachováváme zde obě kategorie (tj. konektory i alternativní vyjádření konektorů). Je ovšem možné, že v jedné z budoucích verzí Pražského závislostního korpusu budou obě kategorie sloučeny pod souhrnný název diskurzní konektory.

3. Prvotní zpracování alternativních vyjádření konektorů v Pražském závislostním korpusu

Pražský závislostní korpus (PDT) nabízí uživatelům práci s jazykem na různých rovinách (morfologické, syntaktické a sémantické) a jako jediný český korpus v nejbližší době poskytne i práci s rovinou nadvětnou, diskurzní.

Nejnovější verze tohoto korpusu (PDT 2.5 – ještě bez diskurzní anotace) byla představena v prosinci 2012 (srov. Bejček a kol., 2012). Diskurzní rovina je anotovaná na datech této verze PDT 2.5, vyšla samostatně pod názvem Pražský diskurzní korpus 1.0 (srov. Poláková a kol. 2012). Pražský diskurzní korpus 1.0 tedy obsahuje anotaci

mezivýpovědních vztahů (typu příčina – důsledek, předčasnost – následnost, opozice, konfrontace atd.), přičemž anotace diskurzní roviny by se měla stát součástí připravované verze PDT 3.0.

V rámci anotace diskurzních vztahů v PDT již probíhá i prvotní zpracování alternativních vyjádření konektorů. Anotátoři dostali pokyn u každého jazykového vyjádření, které signalizuje určitý diskurzní vztah a které nepatří mezi konektory (podle kritérií stanovených pro účely PDT), napsat komentář „altlex". V první fázi anotace se tedy mezi danými výpověďmi neznačí typ diskurzního vztahu (jako u výpovědí uvozených „klasickými" konektory), ale jde zatím spíše o shromáždění dostatečného materiálu pro výzkum altlexů.

Celkově bylo v PDT (který na tektogramatické, tj. na hloubkové, sémantické rovině obsahuje 43 955 vět; srov. Hajič a kol. 2006) nalezeno 306 výskytů, které byly opatřeny prvotní anotátorskou poznámkou *altlex*. Toto číslo je ovšem spíše orientační, protože některá označená vyjádření byla podle našeho názoru interpretována nesprávně (tj. daná vyjádření nepatří mezi altlexy, protože nesignalizují žádný diskurzní vztah), srov. příklad (5), ve kterém bylo spojení *v případě zájmu o* označeno jako altlex, ačkoli zde podle našeho názoru nemá konektivní funkci:

(5) V ceně je ubytování na sedm dní s bohatou selskou snídaní, uvítací přípitek, malý dárek a pobytová taxa.
V případě zájmu o pokoj bez vlastního příslušenství (sprcha a WC na chodbě) by cena za dospělého činila 3180 Kč a za dítě 1265 Kč.

Výsledné číslo, se kterým zde pracujeme, tedy bylo zredukováno na 261 výskytů.

Na druhou stranu je zřejmé, že PDT jistě obsahuje i další altlexy, které ovšem anotátory označeny nebyly. Mohlo se tedy stát, že některý anotátor jeden konkrétní altlex označoval, jiný neoznačoval, a celkový počet výskytů daného typu altlexu tak zatím není konečný. Stejně tak je pravděpodobné, že PDT obsahuje zcela nové typy altlexů, které v rámci první fáze anotace nebyly zachyceny.

Pro ilustraci jsme podle lemmatu vyhledali jeden typ altlexů v celém PDT (vyjádření obsahující slovo *řečeno*) a zjišťovali jsme,

jak jsou anotovány. Mezi danými vyjádřeními byly např. výrazy *stručně řečeno* či *jednoduše řečeno*. Celkově obsahuje nadcházející verze PDT 53 takových vyjádření, z nichž 23 slouží jako diskurzní částice (tj. slouží k usouvztažnění dvou argumentů), a měly by proto být v diskurzních anotacích zachyceny, srov. příklad (6) s diskurzním vztahem explikace:

(6) Odvrácenou stranou podobného stylu práce je nesystematičnost, takže často dochází – slovy Miroslava Macka – k adhocracii.
Jinak řečeno, problémy se řeší, až když hoří, nebo jsou stranickým bagrem odsouvány na zítřek.

Oproti tomu zbylé výrazy s *řečeno* altlexy nejsou (tj. nemají konektivní funkci), srov. příklad (7):

(7) Vše podstatné již bylo *řečeno* a mám za to, že nezazněl jediný důvod, proč majetek církví, který má konkrétního vlastníka, nevrátit.

Anotace těchto prostředků (které mají konektivní funkci a měly by být anotovány) je zatím ovšem nekonzistentní. 3 z daných výrazů byly označeny jako altlexy, 7 jako konektory a zbytek (tedy 13) zůstal zcela bez anotace, srov. tabulka 1.

Tabulka 1. Anotované a neanotované příklady altlexů

Vyjádření	Celkové číslo	Použito jako diskurzní částice	Anotováno		Neanotováno
			jako konektor	jako altlex	
(*jednoduše, krátce, obecně...*) *řečeno*	53	23	7	3	13

Znamená to tedy, že současná předběžná anotace českých altlexů v PDT je pravděpodobně spíše nekonzistentní. Důvodem je, že anotace těchto vyjádření je zatím v první fázi a jejich výzkum právě probíhá. Detailnější a propracovanější anotace altlexů je zamýšlená pro jednu z budoucích verzí PDT a bude mimo jiné založena na této analýze.

4. Sémantická charakteristika českých altlexů

Mezi 261 výskyty jsme identifikovali 94 typů českých altlexů (za jeden typ např. považujeme altlexy se slovesem *následovat*, které se ve funkci altlexu objevily celkem v 10 výskytech). Následně jsme provedli jejich sémantickou klasifikaci a porovnali jsme ji s charakteristikou anglických altlexů z PDTB 2.0.

Ze sémantického hlediska mají diskurzní částice (tj. konektory i altlexy) mezi ostatními kohezními prostředky (tj. např. vedle refe- rence, substituce či elipsy) speciální pozici. Signalizují totiž daný diskurzní vztah a zároveň obsahují anaforický výraz, který odkazuje k prvnímu argumentu (srov. Forbes-Riley a kol. 2006). Anaforická reference může být přitom na povrchu vyjádřená či nevyjádřená (resp. podle terminologie Prasadové a kol. explicitní či implicitní). Prasadová a kol. uvádí jako příklad pro angličtinu výrazy *as a result of that* (*výsledkem toho*) a *as a result* (*výsledkem*).

Situace mezi českými altlexy se zdá být podobná. České altlexy také obsahují anaforickou referenci, která může být explicitní či implicitní (tj. na povrchu vyjádřená či nevyjádřená). Vyjádřit na povrchu anaforickou referenci může být u některých z nich dokonce obligatorní; srov. tabulka 2.

Kategorie „obligatorní a implicitní" v tabulce 2 znamená, že daný výraz nemá možnost anaforickou referenci na povrchu vyjádřit. Např. je nemožné říci **toto stručně řečeno*, ale pouze stručně řečeno. Kategorie „obligatorní a explicitní" zahrnuje altlexy, které jsou bez vyjádření anaforické reference negramatické – nelze např. říci samotné **kvůli*, ale je nutné referenci vyjádřit (*kvůli tomu*). Kategorie altlexů s „fakultativní" anaforickou referencí znamená, že u daného výrazu máme dvě možnosti – buď referenci na povrchu vyjádřit (tj. explicitně – *příkladem toho je*), nebo nevyjádřit (tj. altlex ji vyjadřuje implicitně - *příkladem je*).

Obecně je tedy možné říci, že u českých altlexů anaforická reference na povrchu být vyjádřena nesmí (*stručně řečeno*), musí (*kvůli tomu*) nebo může (*příkladem (toho) je*), viz tabulka 2.

Tabulka 2. Příklady implicitních a explicitních anaforických referencí

	Obligatorní	Fakultativní
Implicitní	Jednoduše řečeno	Dodal
	Přeloženo	První – druhý
	Jak je vidět	Důvodem je
	Jedním dechem	Příkladem je
Explicitní	Nemluvě o tom	Důvodem toho je
	Kvůli tomu	Důsledkem toho kroku je
	I přes tato fakta	Dodal k tomu
	S ním kontrastuje	Výsledkem toho je

Analýza jazykového materiálu ukázala, že altlexy s obligatorní implicitní referencí jsou lexikálně ustálené výrazy, které nejsou kombinovatelné s jinými lexikálními jednotkami, tj. ani s anaforickou referencí (*jak je vidět*). To je důvod, proč tyto výrazy na povrchu anaforickou referenci nikdy nevyjadřují. Další skupinou jsou altlexy vyjadřující anaforickou referenci obligatorně. Jsou to především slovesa, která vyžadují toto doplnění kvůli své valenci. Např. sloveso *kontrastovat* vyžaduje doplnění patientem, kterým je (v případě altlexu) právě anaforické vyjádření. Není proto např. možné říci **jiná skutečnost kontrastuje*, ale pouze *s tím kontrastuje jiná skutečnost*. Anaforická reference je vyjádřená obligatorně také u altlexů, jejichž jádrem je předložka, vyžadující doplnění v určitém tvaru (v případě altlexů jde opět právě o anaforické vyjádření) – srov. př. typu *i přes tato fakta*, *nemluvě o tom*. Altlexy, které vyjadřují anaforickou refe- renci fakultativně, jsou podobné jako v angličtině (srov. Prasadová a kol. 2010). Jedná se o výrazy typu *the result (of this) is – výsledkem (toho) je* a o výrazy s elipsou podstatného jména typu *the second (step) is – druhým (krokem) je*.

Z celkového počtu (tj. 94) typů českých altlexů dosud vyhledaných v PDT vyjadřuje anaforickou referenci 41% fakultativně, 31% obligatorně a 28% ji na povrchu vyjádřit nemůže; srov. tabulka 3.

Tabulka 3. Implicitní a explicitní anaforická reference – typy altlexů

Typy eltexů (z celkového počtu 94)		Počet		%	
Fakultativní typy (existují implicitní a explicitní varianty)[3]		39		41	
Obligatorní typy	**Implicitní**	26	55	28	59
	Ecplicitní	29		31	
Celkem		94		100	

Kromě jednotlivých typů altlexů jsme analyzovali také jejich konkrétní výskyty v PDT. Cílem bylo zjistit, zda altlexy s fakultativní anaforickou referencí mají tendenci tuto referenci na povrchu spíše vyjadřovat nebo nevyjadřovat, tj. zda se v konkrétních jazykových datech vyskytuje častěji např. vyjádření *příkladem toho je* nebo *pří- kladem je*. Jsme si zároveň vědomi toho, že počet vyhledaných altlexů v PDT není konečný a že některé altlexy mohou být v současné první fázi anotovány nekonzistentně (viz výrazy s *řečeno* v tabulce 1). Současná čísla jsou tedy spíše orientační a je nutno je přezkoumat po důkladnější anotaci altlexů v PDT; srov. tabulka 4.

Analýza jazykových dat ukázala, že PDT obsahuje 166 altlexů (resp. jejich konkrétních výskytů, ne typů) vyjadřujících anaforickou referenci fakultativně. Z tohoto počtu se 98 příkladů (tj. 59%) vyskytlo s referencí na povrchu nevyjádřenou, 68 (tj. 41%) s vyjádřenou. Zdá se tedy, že když je u daného vyjádření možnost výběru, je zde slabá tendence anaforickou referenci nevyjádřit. Jak již bylo ale řečeno výše, počet nalezených altlexů v PDT není zatím definitivní, a proto bychom tento postřeh měli brát spíše jako hypotézu, kterou je potřeba potvrdit či vyvrátit na větším množství jazykových dat.

[3] Fakultativní typy nejsou rozděleny na implicitní a explicitní, protože mohou existovat v obou variantách. Vyjádření anaforické reference na povrchu závisí na konkrétních výskytech altlexů, ne na jejich typech.

Tabulka 4. Implicitní a explicitní anaforická reference – výskyty altlexů

Konkrétní výskyty altlexů (z celkového počtu 261)	Obligatorní	Fakultativní	Celkem
Implicitní	35	98	133
Explicitní	60	68	128
Celkem	95	166	261

5. Další výhledy – propojení koreference a diskurzu

Sémantická charakteristika českých altlexů přispívá k jejich novému, detailnějšímu zpracování v PDT, které probíhá v současné době. Především jde o propojení koreferenčních vztahů se vztahy diskurzními. Jak již bylo řečeno, v češtině existuje skupina altlexů, které obligatorně vyjadřují anaforickou referenci odkazující k prvnímu argumentu (jedná se např. o výrazy *kvůli těmto skutečnostem*; *díky tomu*; *na základě zmíněných faktů* atd.). Koreferenční vztah mezi daným anaforickým vyjádřením a příslušným výrazem či úsekem předchozího textu je v PDT již značen (srov. Nedoluzhko 2010). Při vyhledávání altlexů této skupiny (tj. s obligatorní anaforickou referencí) je tedy možné anotaci koreferenčních vztahů využívat – dotazy lze omezovat např. na spojení příslušné předložky (např. *díky*, *kromě*, *vzhledem k*) s výrazem, od kterého je v anotaci veden koreferenční vztah k jinému výrazu či úseku předchozího textu. Vyhledávání a anotace altlexů v PDT je proto díky anotaci koreference značně usnadněna.

6. Závěr

Analýza jazykových anotovaných dat pro nadcházející verzi PDT ukázala, že značný počet diskurzních vztahů v češtině není realizován pomocí klasických konektorů, ale pomocí jiných jazykových prostředků, tzv. alternativními vyjádřeními konektorů. Pokud bychom tyto diskurzní prostředky nezohlednili, anotace diskurzu by byla značně ochuzená a neúplná (tj. některé diskurzní vztahy by v anotaci nebyly vůbec zachyceny). Analýza jazykového materiálu dále poukázala na překrývání hranic diskurzu a koreference, které je možné využít ve vyhledávání a zachycování nových altlexů v PDT.

Literatura

Bejček E. a kol., 2012, *Pražský závislostní korpus 2.5 – rozšířená verze PDT 2.0*. In: *Proceedings of the 24th International Conference on Computational Linguistics* (Coling 2012). Mumbai, India.

Blakemore D., 2002, *Relevance and Linguistic Meaning. The Semantics of Discourse Markers*. Cambridge: Cambridge University Press.

Forbes-Riley K., Webber B., Joshi A., 2006, *Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG*. „Journal of Semantics" 23, s. 55–106.

Hajič J. a kol., 2006, *Průvodce PDT 2.0*. http://ufal.mff.cuni.cz/pdt2.0/ [online 2013-03-7].

Hoffmannová J., 1983, *Sémantické a pragmatické aspekty koherence textu*. Praha: Ústav pro jazyk český.

Hrbáček J., 1994, *Nárys textové syntaxe spisovné češtiny*. Praha: Trizonia.

Irwin J. W., Pulver C. J., 1984, *Effects of explicitness, clause order and reversibility on children's comprehension of causal relationships*. „Journal of Educational Psychology" 76 (3), s. 399–407.

Mladová L., Zikánová Š., Bedřichová Z., Mírovský J., Jínová P., Zdeňkov J., Rysová M., Hajičová E., 2011, *Příručka pro anotaci mezivýpovědních textových vztahů (diskurzu) v Pražském závislostním korpusu*. Praha: ÚFAL MFF [nepublikováno].

Nedoluzhko A., 2010, *Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražeském závislostním korpusu)*. Disertační práce. Http://ufal.mff.cuni.cz/~nedoluzko/koref_anot/diser_text_last.pdf [online 2013-02-18].

Poláková L. a kol., 2012, *Pražský diskurzní korpus 1.0* [CD-ROM]. Praha: ÚFAL MFF UK.

Prasad R. et al., 2010, *Realization of Discourse Relations by Other Means: Alternative Lexicalizations*. In: *Coling 2010: Posters*, s. 1023–1031.

Pradad R. et al., 2008, *The Penn Discourse Treebank 2.0* [CD-ROM]. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Winter E., 1977, *A clause-relational approach to English texts: A study of some predictive lexical items in written discourse*. „Instructional Science" 6 (1), s. 1–91.