Magdaléna Rysová, Filozofická fakulta Univerzity Karlovy v Praze

Způsoby vyjadřování diskurzních vztahů v češtině a jejich anotace v Pražském závislostním korpusu

Problematikou textu se zabývaly již mnohé studie. Jejich stěžejní otázkou je především to, co dělá text textem. Podle některých autorů (srov. např. Halliday a Hasanová 1976) je posluchač/čtenář (alespoň do určité míry) obdařen schopností posoudit, zda soubor po sobě jdoucích vět tvoří ucelený text, či zda se jedná pouze o náhodný shluk vět na sebe nenavazujících. Tato obecná lidská vlastnost je podle autorů důkazem toho, že existují základní atributy, které musí obsahovat každý soubor vět, aby mohl být považován za text. Hlavními atributy je podle nich koheze a koherence. Právě díky obsahové i jazykové soudržnosti lidé psaným textům rozumí. Je proto zřejmé, že je zapotřebí věnovat pozornost jazykovým prostředkům, které se na kohezi a koherenci textu podílejí.

Příspěvek [1] se tedy zabývá otázkou, jaké jazykové prostředky mají v češtině schopnost spojovat věty v ucelený text. V odborné literatuře jsou tyto výrazy nazývány textové či diskurzní konektory. Většinou se jimi rozumí prostředky patřící mezi určité slovní druhy, především spojky (*proto*, *ale*, *a*), příslovce (*potom*, *pak*) a částice (*také*, *například*). Tento příspěvek se ovšem zaměřuje na širší možnosti vyjadřování diskurzních vztahů, konkrétně na jazykové výrazy s konektivní funkcí, které nejsou lexikálně ani syntakticky nijak omezeny. Jedná se například o spojení typu *příčinou bylo*; *jinak řečeno* atd. Jako první se těmito prostředky podrobně zabývali autoři z Pensylvánské univerzity (srov. Prasadová a kol. 2010), kteří je nazvali alternativní lexikální vyjádření diskurzních konektorů (zkráceně altlexy). Tento příspěvek na studii o

[1] Výstup projektu Grantové agentury Univerzity Karlovy v Praze „Diskurzní konektory v češtině“ č. 36213 řešeného na Filozofické fakultě Univerzity Karlovy v Praze.

anglických altlexech, které jsou zpracovávány v rámci anotace textových vztahů v diskurzním korpusu Penn Discourse Treebank – PDTB (srov. Prasadová a kol. 2008), přímo navazuje a klade si za cíl udělat obdobnou analýzu těchto výrazů pro češtinu.

Autoři studie o anglických altlexech (srov. Prasadová a kol. 2010) tyto výrazy hodnotili po stránce sémantické, syntaktické a lexikální. Sémantická analýza českých altlexů již byla provedena v předchozím výzkumu (srov. Rysová 2012, 2013). Cílem tohoto příspěvku je proto dané výrazy detailně analyzovat po stránce syntaktické a lexikální.

1. Anotace alternativních vyjádření konektorů v Pražském závislostním korpusu

Analýza českých altlexů byla provedena na datech nejnovější verze Pražského závislostního korpusu (tj. PDT 2.5 – srov. Bejček a kol. 2012). Tento korpus obsahuje velké množství publicistických textů, které jsou zpracovávány na rovině morfologické (přibližně 2 miliony slov), syntaktické (přibližně 1,5 milionů slov) a sémantické (přibližně 0,8 milionů slov) (srov. Hajič a kol. 2006). Na rovině sémantické probíhá v současné době také zpracování textových vztahů (prvotní verze anotace diskurzu byla vydána samostatně jako Pražský diskurzní korpus 1.0 – srov. Poláková a kol. 2012).

V rámci prvotní anotace diskurzu byly zpracovávány pouze textové vztahy uvozené explicitními konektory – tj. např. spojkami typu *avšak*, *proto*, či příslovci typu *potom*, *následně* (srov. příručka k anotaci diskurzu – Mladová a kol. 2012). Diskurzní vztahy uvozené altlexy – tj. např. spojeními typu *důvodem*

je, příčinou bylo atd. – v této fázi anotace zachycovány nebyly. Dané výrazy byly ovšem opatřeny anotátorskými poznámkami „altlex“.

Z celkového počtu 43 955 vět na sémantické (tj. hloubkové, tzv. tektogramatické) rovině, které jsou přístupné veřejnému vyhledávání (srov. Hajič a kol., 2006), bylo nalezeno 261 relevantních výskytů českých altlexů (tento počet je ovšem zatím spíše orientační, předpokládáme, že po detailním zpracování výrazně vzroste[2]). Tyto výskyty jsme dále rozdělili do 94 typů (jako jeden typ označujeme např. sloveso *odůvodnit*, které se ve funkci altlexu objevilo v prvotní anotaci celkem ve 3 výskytech). Následně jsme provedli jejich lexikální a syntaktickou klasifikaci.

2. Syntaktická charakteristika českých altlexů

Ze syntaktického hlediska jsme dále zkoumali, zda jsou vyhledané české altlexy zapojeny do větné struktury, tj. zda plní funkci větného členu či nikoliv. Zjišťovali jsme tedy, zda altlexy ve větě působí jako výrazy rozvíjející jiné větné členy či celou větu (jako tzv. větné modifikátory). Poté jsme analyzovali jejich syntaktickou strukturu. Cílem bylo zjistit, zda české altlexy podléhají jistým syntaktickým vzorcům, příp. kterými z nich se řídí nejčastěji.

Z analýzy jazykového materiálu vyplynulo, že 78 typů (doposud nalezených) českých altlexů (tj. 83 %) je zapojeno do větné stavby, a plní tedy funkci větného členu, zatímco 16 typů (tj. 17 %) nikoli (tj. plní funkci větných modifikátorů). Těchto 16 typů zahrnuje altlexy, které komentují celou výpověď

[2] Pro ilustraci jsme vyhledali všechny výskyty jednoho typu altlexu, jehož jádro tvoří předložka *díky* a které se pojí s anaforickým vyjádřením odkazujícím k předchozímu argumentu (tj. např. *díky tomu*, *díky těmto skutečnostem* atd.). V celém PDT se předložka *díky* objevuje celkem ve 191 výskytech, přičemž jako altlex působí ve 14 případech. V první fázi anotace byla ovšem opatřena poznámkou „altlex“ pouze jednou.

jako tzv. disjunkty nebo slouží pouze ke strukturování textu, a tedy nezasahují do obsahu dané výpovědi. Disjunkty či větné modifikátory chápeme jako výrazy, které se vztahují „k obsahu celé věty či k způsobu jeho vyjádření“ (Dušková 2006: 474) – srov. např. altlexy *přeloženo*, *jednoduše řečeno*, *pravda* atd. Jako výrazy strukturující text označujeme altlexy typu *za prvé – za druhé*. Srov. tabulka 1:

	Alternativní vyjádření konektorů		**Celkem**
	Zapojená do větné stavby	**Nezapojená do větné stavby**	
Příklady	*Jiný*[3]	*Rozumějme*	
	Kvůli tomu	*Přeloženo*	
	Stejným dechem	*Jak je vidět*	
	Podobně	*Pravda*	
	I přes tato fakta	*Jednoduše řečeno*	
	Důsledkem tohoto kroku	*Za prvé – za druhé*	
	To je důvod, proč	*Dlužno dodat*	
Počet typů	78	16	94
%	83	17	100

Tabulka 1: Syntaktická charakteristika českých altlexů: Zapojení do větné stavby

Dalším kritériem, podle kterého jsme české altlexy hodnotili, je jejich syntaktická struktura, tj. zkoumali jsme typy syntaktických frází, ve kterých se altlexy objevují. Analýza jazykového materiálu ukázala, že vyhledané altlexy byly realizovány frázemi nominálními, adjektivními, číslovkovými, slovesnými, příslovečnými, předložkovými, částicovými nebo celými klauzemi (klauzemi

[3] *Jiný* ve spojení s anaforickým vyjádřením – např. *Do zástavy může přitom být vždy dána jen taková pohledávka, podle níž má být věřiteli poskytnuta nějaká věc, právo nebo jiná majetková hodnota – zpravidla bývá zastavována pohledávka peněžní.* ***Jiným*** *zvláštním druhem zástavního práva je zastavení cenných papírů.*

zde rozumíme větné a polovětné konstrukce, tj. – v pojetí PDT – takové stromy, jejichž řídícími uzly jsou slovesa v určitém či neurčitém tvaru).

2.1 Předložkové fráze

Největší skupinu mezi vyhledanými altlexy tvoří předložkové fráze (33 typů z celkových 94). Tuto skupinu altlexů je dále možno členit na dvě podskupiny.

a) Jádro lexikálního významu altlexu neseno předložkou

První podskupina zahrnuje výrazy, u kterých jsou jádro lexikálního významu a schopnost působit jako altlex neseny předložkou. Tyto altlexy se skládají ze sekundární předložky a výrazu anaforické reference, která variuje. Pro ilustraci můžeme uvést altlex *na rozdíl od toho*. Neměnná část nesoucí význam diskurzního vztahu (tj. signalizující, o jaký diskurzní typ se jedná) je v tomto případě *na rozdíl od* klasifikovaná celkově jako sekundární předložka (klasifikuje ji tak např. Kroupová 1984). Druhou část altlexu zde tvoří anaforické vyjádření *toho*, které se může obměňovat (srov. např. *na rozdíl od toho / této skutečnosti / předchozího* atd.). Dalšími příklady této podskupiny mohou být altlexy *nemluvě o tom*, *na základě čehož*, *navzdory tomu*, *souběžně s tím*, *vzhledem k tomu*, *vinou toho*, *v rozporu s* atd.

b) Jádro lexikálního významu altlexu neseno jménem

Do druhé podskupiny patří altlexy, které jsou tvořeny primární předložkou a neměnným podstatným jménem, které signalizuje, že daný výraz se podílí na utváření diskurzního vztahu, a také naznačuje, o jaký typ diskurzního vztahu se jedná. Příkladem takového altlexu může být spojení

z tohoto důvodu. Jádro lexikálního významu je v tomto případě neseno slovem *důvod*, které také signalizuje, že mezi danými dvěma argumenty jde o vztah příčiny a důsledku – srov. výpovědi z PDT (altlex *z tohoto důvodu* je zde možné nahradit konektorem *a proto*):

(1) *S ohledem na toto ustanovení by se hrubé chování muselo týkat vaší osoby a nestačí pouze nevhodné zacházení s předmětem darovací smlouvy, to je darem.*
Z tohoto důvodu by byla vaše žaloba na vrácení daru u soudu zamítnuta.

2.2 Altlexy realizované celými klauzemi

Druhou nejobsáhlejší skupinu tvoří české altlexy realizované celou klauzí, které je opět možné rozdělit do dvou tříd.

a) Klauze obsahující slovesa s oslabeným lexikálním významem

První (a zároveň početnější) podskupinu tvoří klauze obsahující finitní slovesa s oslabeným lexikálním významem (např. *být*, *tvořit*, *sloužit*, *uvést*), přičemž jádro lexikálního významu je neseno jiným komponentem (zpravidla podstatným jménem) – srov. např. *důvodem je*, *rozdílem je*, *výjimku tvoří*, *jako příklad slouží*, *jako důvod uvádí*. Otázkou proto zůstává, do jaké skupiny altlexy tohoto typu řadit – zda pod celé klauze nebo je přehodnotit a přeřadit pod fráze, v nichž jádro lexikálního významu nese jméno (tedy např. pod fráze nominální).

b) Polovětné konstrukce

Druhá podskupina obsahuje polovětné konstrukce, tj. klauze, jejichž jádrem je sloveso v neurčitém tvaru – např. *jednoduše řečeno*, *přeloženo*, *dlužno dodat*. Všechny altlexy tohoto typu plní ve větě funkci větných modifikátorů. V této práci je řadíme pod klauze, protože se objevují v ustálených polovětných konstrukcích a slovesa sama nevystupují jako altlexy – srov. slovesa *říct* či *přeložit* sama o sobě ještě nesignalizují diskurzní vztah. Mohou se naopak vyskytovat v takovém kontextu, ve kterém diskurzní funkci postrádají, tj. ve kterém neslouží k usouvztažnění dvou slovesných argumentů v rámci textu – srov. např. výpovědi typu *Petr to řekl mamince* či *Jana přeložila větu do angličtiny*. Znamená to tedy, že daná slovesa nejsou altlexy inherentně, ale stávají se jimi až v jisté ustálené formě a příslušném kontextu (*přeloženo*) či v kombinaci s jinými výrazy (*jednoduše řečeno*). Srov. příklad (2) z PDT, ve kterém spojení *jednoduše řečeno* signalizuje diskurzní vztah generalizace, a příklad (3), ve kterém výraz *přeloženo* uvádí vztah ekvivalence:

(2) *Každý odklad nejenže přináší velké ztráty na dané investici, ale také se nepříznivě*

promítá do ekonomiky země i veřejného života.

Pokud budeme do vysokorychlostní železnice investovat v potřebném optimálním čase, můžeme využít všech jejich výhod.

Se zpožděním naopak žádné výhody nezískáme.

<u>Jednoduše řečeno</u>, čím déle budeme projekt odkládat, tím vyšší pak budou náklady.

(3) *Pořadatelem je EKULT – nadace pro ekologii a kulturu a v podtitulu Lipnice' 94 můžeme číst: Malý festival pro malou planetu.*

Přeloženo: Lipnice opět (promiňte ta nevhodná, leč přesná slova) ideově jde před svou dobou.

Na druhou stranu je jádro či hlavní uzel těchto altlexů tvořen slovesem. To může být argumentem pro to, aby byly dané výrazy řazeny pod verbální fráze. V takovém případě by bylo možné celou skupinu altlexů tvořených klauzemi zrušit, protože všechny typy altlexů, které obsahuje, mohou být přeřazeny jinam. Obě podskupiny ovšem vykazují oproti ostatním třídám, do kterých by mohly být zařazeny, své společné specifické znaky (srov. výše), a proto je prozatím necháváme ve zvláštní skupině.

2.3 Verbální fráze

Třetí největší skupinu tvoří verbální fráze. Jejich hlavními uzly jsou slovesa, která sama lexikálně signalizují určitý diskurzní vztah (alespoň v některém ze svých významů). K tomu, aby plnila funkci altlexů, se tedy nemusí pojit s jinými výrazy, ani nejsou vázané na určitou formu (jako např. *přeloženo* či *jednoduše řečeno*). Jedná se tedy o altlexy lexikálně i formálně volné (podle terminologie Prasadové a kol. 2010). Dané altlexy se v textu mohou vyskytovat v celém paradigmatu. Patří sem např. slovesa *předcházet*, *následovat*, *zdůvodnit* atd. – srov. příklady z PDT:

(4) *Gyula Horn se vyslovil pro možné zavedení majetkové daně.*
Zdůvodnil to tím, že utahování opasků se nemůže vztahovat pouze na lidi žijící ze mzdy.

(5) *Hranice jedné miliardy Kč by banka chtěla dosáhnout koncem roku 1996.*

Předcházet bude řada postupných kroků.

3. Lexikální charakteristika českých altlexů

Z lexikálního hlediska hodnotí Prasadová a kol. anglické altlexy podle toho, zda jsou lexikálně volné, či ustálené. Striktně pak anglické altlexy přiřazují do jedné z daných skupin. V tomto příspěvku ovšem nechápeme volná a ustálená lexikální vyjádření jako dvě uzavřené či oddělené entity, ale jako škálu se dvěma póly (srov. Howarth 2000).

První reprezentují spojení obsahující určité slovo, které je altlexem inherentně, tj. samo o sobě (v daném významu) signalizuje jistý diskurzní vztah a vstupuje do různých volných kombinací. Ty se vyznačují tím, že jsou gramaticky i lexikálně neomezené. Pro ilustraci můžeme uvést slovesné altlexy, které se v textu mohou vyskytovat v celém paradigmatu, tj. objevují se ve všech slovesných časech, v aktivu i pasivu, s modálními výrazy atd. – srov. nalezené příklady pro jeden typ altlexu, slovesa *dodat* (ve významu ‚dodatečně říci, poznamenat' /*SSČ pro školu a veřejnost* 2001: 62/): *k tomu je třeba dodat, dodal, dodává člen organizace, dodejme*.

Druhý pól tvoří ustálená víceslovná vyjádření, jejichž složky se stávají altlexy pouze v dané kombinaci. Tato spojení jsou lexikálně i gramaticky omezená, jejich podstatou je tedy jistá anomálie či nepravidelnost (srov. Čermák 2007). Vykazují přitom buď nepatrnou míru variability (tj. vyskytují se v omezeném počtu kombinací jako např. *jednoduše/krátce/obecně řečeno*), nebo jsou plně ustrnulá (tj. omezená na jedinou možnou kombinaci – např. *tím spíš*). Tyto altlexy jsou obvykle neúplné gramatické struktury, které patří mezi tzv. *lexical bundles* (lexikální svazky či celky) charakterizované jako nejčastěji se

vyskytující sekvence slov, které se podílejí na organizaci a strukturaci textu (Biber a Conradová 1999).

Ne všechny vyhledané altlexy ovšem bylo možné podrobit takto striktní kategorizaci (tj. nebylo možné stanovit, zda je daný výraz či spojení výrazů lexikálně zcela volné či ustálené) – srov. např. altlex *sloužit jako příklad*. Toto spojení nesplňuje všechny podmínky zcela ustrnulých (idiomatických) spojení (nejedná se např. o neúplnou gramatickou strukturu, sloveso může být časováno atd.), ale zároveň jej nejde zařadit mezi spojení volná, protože vykazuje jistou míru očekávání a prediktability, typickou pro ustálená spojení.

Z tohoto důvodu neuplatňujeme na české altlexy striktní lexikální kategorizaci, ale chápeme je jako škálu od plně volných kombinací po idiomatická spojení. Zároveň je třeba dodat, že zcela ustálených spojení by bylo mezi vyhledanými altlexy pouze několik a že většina z nich se vyskytuje blíže k hranici volných kombinací.

4. Závěr

Příspěvek přinesl syntaktickou a lexikální analýzu tzv. alternativních vyjádření diskurzních konektorů v češtině (zkráceně altlexů – srov. např. *zdůvodnit*) provedenou na datech Pražského závislostního korpusu.

Ze syntaktické charakteristiky českých altlexů vyplynulo, že tyto výrazy mají nejčastěji podobu předložkových frází (*z tohoto důvodu*), celých klauzí (*jako příklad slouží*) či verbálních frází (např. altlexy s lexémem *argumentovat*).

Z lexikálního hlediska se altlexy pohybují na škále od volných kombinací (srov. např. *k tomu je třeba dodat*, *dodal*, *dodává člen organizace*, *dodejme*) po

zcela ustálená spojení (kterých byla ve zkoumaném vzorku jazykových dat ovšem menšina – srov. např. *tím spíš*).

Literatura

Čermák František, 2007, *Frazeologie a idiomatika: česká a obecná*, Praha.

Biber, Douglas; Conrad, Susan, 1999, Lexical Bundles in Conversation and Academic Prose, *Out of corpora: studies in honour of Stig Johansson*. Atlanta, 181–190.

Dušková, Libuše a kol., 2006, *Mluvnice současné angličtiny na pozadí češtiny*, Praha.

Hajič, Jan et al., 2006, *Prague Dependency Treebank 2.0*, Philadelphia.

Halliday, Michael Alexander Kirkwood; Hasan, Ruqaiya, 1976, *Cohesion in English*, London.

Howarth, Peter, 2000, Describing diachronic change in English phraseology, *Las Lenguas de Europa: Estudios de fraseología, fraseografía y tradducción* Interlingua 12, Albolote, 213–230.

Knott, Alistair, 1996, *A Data-Driven Methodology for Motivating a Set of Coherence Relation*s, Edinburgh, Ph.D. thesis.

Kroupová, Libuše, 1984, Klasifikace sekundárních předložek z hlediska jejich tvoření, *Naše řeč 67 (3)*, Praha, 113–116.

Mladová, Lucie; Zikánová, Šárka; Bedřichová, Zuzanna; Mírovský, Jiří; Jínová, Pavlína; Zdeňková, Jana; Rysová, Magdaléna; Hajičová, Eva, 2012, *Příručka pro anotaci mezivýpovědních textových vztahů (diskurzu) v Pražském závislostním korpusu*, Praha, MFF UK.

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

Prasad, Rashmi et al., 2010, Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In *Coling 2010: Posters* (August 2010), 1023–1031.

Prasad, Rashmi et al., 2008, The Penn Discourse Treebank 2.0, *Proceedings of the 6th International Conference on Language Resources and Evaluation*, CD-ROM.

Rysová, Magdaléna, 2012, Alternative Lexicaliztions of Discourse Connectives in Czech. Calzolari, Nicoletta et al. (eds.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turecko: European Language Resources Association (ELRA), 2800–2807.

Rysová, Magdaléna, 2013, Jazykové prostředky vyjadřující textové vztahy v češtině a jejich zpracování v Pražském závislostním korpusu, *Bohemistyka 13 (1)*, 57–71.

Slovník spisovné češtiny pro školu a veřejnost, 2001, Praha.

Zdroj dat

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

Magdaléna Rysová

Vystudovala magisterské studium na Filozofické fakultě Univerzity Karlovy v Praze – obory Anglistika a amerikanistika, Učitelství českého jazyka a literatury pro střední školy. Nyní je studentkou doktorského studia na Filozofické fakultě Univerzity Karlovy – obor Český jazyk. Je interní zaměstnankyní Ústavu formální a aplikované lingvistiky na Matematicko-fyzikální fakultě v Praze.

Sie absolvierte das Masterstudium an der Philosophischen Fakultät der Karls-Universität in Prag – die Fächer Anglistik und Amerikanistik, Lehren der tschechischen Sprache und Literatur für Mittelschulen. Zurzeit ist sie eine Doktorandin an der Karls-Universität in Prag – im Fach Tschechische Sprache. Sie ist eine interne Mitarbeiterin des Instituts der formalen und angewandten Sprachwissenschaft an der Fakultät für Mathematik und Physik der Karls-Universität in Prag.

Abstract

The aim of the paper is to describe the possibilities of expressing textual relations in Czech – mainly the multiword expressions with connecting function like *s odůvodněním* (*with justification*), *výsledkem bylo* (*the result was*) etc. The analysis was done on the data of the Prague Dependency Treebank. The paper concentrates on the syntactic and lexical characteristics of these expressions and follows the similar analysis done also for English by the authors of the Penn Discourse Treebank.

Key words

connectives, discourse, Prague Dependency Treebank, text, textual relations

Abstrakt

Das Ziel der Arbeit ist, die Expressionsmöglichkeiten der Textbeziehungen im Tschechischen zu beschreiben – vor allem die aus mehreren Wörtern bestehenden Ausdrücke mit Verbindungsfunktion wie *s odůvodněním* (*mit Begründung*), *výsledkem bylo* (*das Ergebnis war*) usw. Die Analyse wurde auf den Daten des Prague Dependency Treebanks durchgeführt. Der Artikel konzentriert sich auf die syntaktischen und lexikalischen Merkmale dieser Ausdrücke und folgt eine ähnliche Analyse für Englisch, die von den Autoren des Penn Discourse Treebanks getan wurde.

Schlüsselwörter

Textbindewörter, Diskurs, Prague Dependency Treebank, Text, Textbeziehungen