Markéta Lopatková, Martin Plátek

O zlomovém bodu introspekce a souvisejících problémech

Tento příspěvek je napsán u příležitosti významného životního jubilea Jana Hajiče a jako vzpomínka na den 11. září 2001. V článku postupně vysvětlujeme pojem zlomového bodu introspekce (ZBI), pojem statistického zlomového bodu introspekce (SZBI) a způsob, jak a proč jsme k těmto pojmům dospěli.

1 Historizující úvod

Tento článek autoři napsali u příležitosti významného životního jubilea, padesátin, Jana Hajiče. Místo pojmu *významné životní jubileum* či padesátiny však raději pracují s pojmem *statistický zlomový bod introspekce (SZBI).* Pojem statistický zlomový bod introspekce je odvozen z jednoduššího pojmu *zlomový bod introspekce (ZBI)*. Postupně se pokusíme tyto pojmy zavést. Přitom poukážeme na události a další důvody, které nás k těmto pojmům přivedly. Byli bychom rádi, kdyby byl čtenář po přečtení našeho příspěvku schopen nahlédnout, že jeden ze spoluautorů tohoto příspěvku *zlomovým bodem introspekce* ještě neprošel. Naopak by mělo být zřejmé, že druhý autor je již delší dobu o tuto zkušenost bohatší. Z tohoto pohledu považujeme složení autorského týmu za vyvážené.

1.1 Špičák na Šumavě, 10. září 2001

V této podsekci popíšeme událost, která se stala dne 10. září 2001 na kopci naproti Špičáku na Šumavě. Toho dne jsme byli na popsaném místě účastníky konference 'Text, Speech, Dialog', viz Kuboň and Plátek (2001); Skoumalová et al. (2001). Bylo pozdní odpoledne, pěkné podzimní počasí. Starší z autorů M.P. pozoroval přírodní úkazy nad Špičákem a pravil:

Podívejte se, z lesa vycházejí páry.

Přítomný V.K. po chvíli odpověděl:

Já vidím jenom jeden.

Tím uvedl staršího z autorů M.P. do zmatku pramenícího z neporozumění nečekanému obratu v konverzaci. Jeho pocit zmatku zvyšovala skutečnost, že zbytek skupiny žádné příznaky zmatení nejevil. Teprve později si M.P. uvědomil víceznačnost svého výroku i fakt, že řada přítomných se automaticky přiklonila ke stejnému významu zmíněného výroku jako V.K. Tenkrát se nad tím nikdo více nezamýšlel.

*Autoři děkují všem zmiňovaným za motivaci pro tento text.

1.2 Další den ráno

Jan Hajič oznámil účastníkům konference útok na mrakodrapy v New Yorku a jeho následky. Poznamenejme, že z těchto dvou po sobě jdoucích událostí si většina zúčastněných pamatuje spíše tu druhou.

1.3 Letošní rok na počátku jara

Jednoho dne tohoto roku na přelomu zimy a jara přišel domů ze školy student O.P. a vyprávěl přítomnému M.P. událost popsanou v podsekci 1.1. Studentův přednášející V.K. totiž tuto příhodu použil pro ilustraci možných nedorozumění pramenících z jazykové víceznačnosti a studenta O.P. tento příklad zaujal. M.P. jeho zaujetí začal sdílet a připomenul tuto příhodu na následujícím pravidelném semináři. Diskuse, která se rozvinula, je jádrem tohoto příspěvku. Hlavními diskutéry byli jeho autoři.

2 Diskuse

První věc, kterou jsme si během seminární diskuse uvědomili, byl fakt, že kdybychom na zmíněnou větu z 10. září narazili při značkování pro Český národní korpus[1] či pro Pražský závislostní treebank[2], viz Hajič (2006a), tak bychom – v závislosti na svých individuálních prožitcích – tuto větu analyzovali různě. Shledali jsme tedy, že značkování bývá v mnoha případech závislé na znalosti reálných událostí a na individuální introspekci.

Doposud byly introspektivní a statistické metody dávány spíše do protikladu a byly prezentovány jako víceméně nezávislé. Tak jsme také chápali habilitační a profesorskou přednášku Jana Hajiče. Připomeňme však v této souvislosti, že obvyklým podkladem pro statistické metody v lingvistice bývají označkované korpusy a treebanky. Událost z předvečeru 11. září 2001 nám dovoluje podívat se na vztah introspekce a statistiky novým pohledem.

O takový pohled jsme v pokračující diskusi usilovali. Povšimli jsme si věkového složení skupiny s ohledem na interpretaci věty z podsekce 1.1. Starší členové se přikláněli výrazně k interpretaci M.P., mladší jasně preferovali interpretaci V.K. Začali jsme pracovat s hypotézou, že pro jisté věty se jejich interpretace mění s věkem. Existuje tedy teoretický věkový bod, po jehož překročení potenciální anotátor spontánně mění interpretaci. Tento bod nazýváme *zlomovým bodem introspekce (ZBI)*. Vidíme, že zlomový bod introspekce je důležitý věkový předěl, který souvisí jak s lingvistickou introspekcí, tak se statistickými metodami používanými v lingvistice. To je důležité pozorování, se kterým budeme dále pracovat.

3 Závěry

Naší snahou je prezentovat významné výročí Jana Hajiče jako bod, který by měl být i pracovně přelomový. Původně jsme se domnívali, že pro tento účel je vhodný koncept

[1] `http://ucnk.ff.cuni.cz/`
[2] `http://ufal.mff.cuni.cz/pdt2.0/`

zlomového bodu introspekce. Uvědomili jsme si však, že hodnota zlomového bodu introspekce bývá velice individuální. Neměli jsme žádná data (a ani mít nemohli), která by nám potvrdila existenci zlomového bodu introspekce Jana Hajiče a jeho blízkost významnému jubileu. Proto jsme zformulovali hypotézu, že průměr zlomového bodu introspekce vhodného vzorku obyvatelstva je blízký číslu 50. Průměr zlomového bodu introspekce vhodného vzorku obyvatelstva jsme pro zjednodušení nazvali *statistickým zlomovým bodem introspekce*. O ověření hypotézy, že statistický zlomový bod introspekce je blízký padesátce, jsme požádali Doc. Petra Boschka, CSc., statistika katedry psychologie FFUK, viz Boschek (2001). K naší radosti Doc. Boschek naši hypotézu po jisté testovací době potvrdil. Máme tedy možnost předložit své závěry.

Dá se očekávat, že bude-li Jan Hajič používat ve své práci v budoucnu metodu introspekce, může zjistit, že dochází k odlišným závěrům než v letech předchozích. Naopak, bude-li používat v práci svých týmů metod statistických, může výsledky značně ovlivnit věkovým složením týmu anotátorů. S ohledem na poslední bod autoři po jisté diskusi doporučují věk anotátorů pravidelně rozložit kolem statistického zlomového bodu introspekce.

4 Otevřené otázky

Čtenář si jistě povšiml, že autoři otevřeli otázku, jak může věkové rozložení anotátorského týmu ovlivnit výsledky lingvistických anotací. Je možné si představit i jiná hlediska pro zkoumání vlivu složení anotátorských týmů na výsledky jejich práce. Na těch se však autoři zatím neshodli a ponechávají tak některá zajímavá témata otevřená.

Literatura

Petr Boschek. *Tabulky norem a psychometrické vlastnosti české verze testu WISC-IIIUK*. Praha, IPPP ČR, 2001.

Jan Hajič. Statistické metody v počítačové lingvistice. Habilitační přednáška, MFF UK, 2002.

Jan Hajič. Complex Corpus Annotation: The Prague Dependency Treebank. In M. Šimková', editor, *Insight into Slovak and Czech Corpus Linguistic*, pages 54–73. Veda, Slovakia, 2006a.

Jan Hajič. Počítačová lingvistika jako experimentální věda. Profesorská přednáška, MFF UK, 2006b.

Vladislav Kuboň. Úvod do počítačové lingvistiky. Přednáška, MFF UK, 2009/10.

Vladislav Kuboň and Martin Plátek. A Method of Accurate Robust Parsing of Czech. In V. Matoušek, P. Mautner P., R. Mouček R., and K. Tašer, editors, *Proceedings of the 4th International Conference 'Text, Speech and Dialogue', TSD 2001*, pages 99–99. LNAI 2166, Springer-Verlag, Berlin Heidelberg, 2001.

Hana Skoumalová, Markéta Straňáková-Lopatková, and Zdeněk Žabokrtský. Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. In V. Matoušek, P. Mautner P., R. Mouček R., and K. Tašer, editors, *Proceedings of*

the 4th International Conference 'Text, Speech and Dialogue', TSD 2001, pages 142–149. LNAI 2166, Springer-Verlag, Berlin Heidelberg, 2001.