Univerzita Karlova v Praze

Filozofická fakulta

TEZE DISERTAČNÍ PRÁCE

Jiří Milička

Teorie komunikace jakožto explanatorní princip přirozené víceúrovňové segmentace textů

The Theory of Communication as an Explanatory Principle for the Natural Multilevel Text Segmentation

Vedoucí disertační práce: doc. PhDr. Petr Zemánek, CSc.

Základní součást: Ústav srovnávací jazykovědy

Studijní obor: Jazyky zemí Asie a Afriky

Praha 2015

Úvod

Na počátku této disertace stálo jednoduché tvrzení: „výběr slova je omezen jeho kontextem“, a otázka, proč tomu tak je. Skutečnost, že jedna výpověď ovlivňuje vhodnost nebo pravděpodobnost užití výpovědi jiné, je tou snadněji vysvětlitelnou částí tvrzení, neboť lingvistika nemá odpověď na otázku, proč je jazyk vůbec dělen do slov. Obecně řečeno, hledáme vysvětlení, proč je mluvená i psaná řeč segmentována do slov, vět, souvětí, odstavců, kapitol, svazků a snad i vyšších více či méně ohraničených jednotek. Slova jsou v teoriích často považována za předem dané kategorie, jakési významové celky, přičemž není jasné, v čem ona ucelenost spočívá. Dělení do slov nebo podobných segmentů považujeme za samozřejmé.

Proč považujeme řeč, která není takto segmentována, za nepřehlednou? Byla snad nějaká biologická, fyzikální omezení, která zabránila tomu, aby se náš mozek vyvinul takovým způsobem, aby byl schopen produkovat i zpracovat nesegmentovaný tok symbolů? Nebo můžeme najít nějaké jiné vysvětlení, které je dáno přirozenými vlastnostmi jakékoliv komunikace?

Tato studie ukazuje, že dostatečné vysvětlení tohoto fenoménu poskytuje samotná teorie informace (nebo spíše teorie komunikace). To samozřejmě neznamená, že toto vysvětlení je jediné možné, dokonce ani to, že je nejdůležitější. To, co budeme popisovat na následujících stránkách, je nejspíš pouze jedním z mnoha faktorů, které do celého procesu zasahují.

Předkládaná explanace je funkcionálního charakteru a předpokládá, že vývoj jazyka je extenzí evoluce, takže funkcionální vysvětlení je předpokladem pro vysvětlení kauzální, respektive pro některé účely je i hodnotnější (Wouters, 1999, 2007).

Tato práce ovšem může přispět i k formování obecné představy o tom, jak konkrétně funguje rozpoznání neúspěšných komunikačních aktů. Tím potažmo osvětluje konkrétní prostředky, jimiž k jazykové evoluci dochází, čímž přispívá k formování kauzální explanace nejen jevu, který je hlavním námětem této práce, ale i mnoha dalších.

První dvě kapitoly představované studie pojednávají o obecných principech, ze kterých studie vychází.

Následující část má přímočarou, téměř narativní strukturu, která je vyjádřena těmito body:

1. Na rozdíl od distinktivních rysů a morfémů nejsou hlásky, slova, věty ani souvětí logickou nezbytností jazyka.

2. Přesto je tento nebo podobný druh vnořené segmentace přítomen v různých jazycích a je pevně zakotven i v naší představě o jazyce.

3. Je tomu tak, neboť vnořená několikaúrovňová segmentace dovoluje vkládání

redundance na různých úrovních, což je efektivní způsob, jak přenést informaci přes kanál obsahující dávkový šum.

4. Existuje mnoho strategií, jak redundanci vložit, a druhů redundance, které vloženy být mohou.

5. Kvůli oddělování segmentů je do segmentů vkládáno určité množství informace navíc. Množství této informace je nezávislé na délce segmentu, který odděluje. Na tomto principu je možno založit úspěšný model pro Menzerathův vztah.

Každému z těchto bodů je věnována jedna kapitola. Pojďme se nyní na ony kapitoly podívat podrobněji. Následující shrnutí je pouze orientační, pro sledování argumentační struktury práce je nezbytné si ji skutečně přečíst, neboť samotná práce je popsána koncizně a není možné ji shrnout do útvaru omezeného na jednu desetinu jejího rozsahu; navíc práce používá termíny jinak, než jak jsou používány v současné lingvistice (zejména pojmy *komplexita* a *redundance* jsou používány tak, jak je definuje kolmogorovovská teorie komunikace). Čtenáři tohoto autoreferátu, kteří nemají přístup k vlastní práci, si ji mohou stáhnout na adrese www.milicka.cz/disertace.pdf. Při čtení je pak důležitá trpělivost, neboť smysl celé práce čtenář nejspíše plně docení teprve na konci 5. kapitoly, přičemž samotné přečtení závěru páté kapitoly bez znalosti kapitol předchozích ve čtenáři probudí spíše zmatek než porozumění.

1 Teorie komunikace a jazyk

Komunikace mezi různými entitami si je vzájemně podobná a nezáleží, jestli těmito entitami jsou stroje nebo zvířata a jak konkrétně probíhá. Společné aspekty komunikace různých systémů jsou předmětem intenzivního studia informatiků minimálně od poloviny minulého století a lingvistika ji reflektuje (Shannon, 1948; Bar-Hillel, 1955; Ashby, 1962; MacKay, 2005). Teorie informace a komunikace není zajímavá jen z pohledu obecného lingvisty, ale i filolologa nebo překladatele, který díky ní může vysvětlit, proč dokáže odhadnout význam i z neúplného textu nebo proč pozná, když textu nerozumí.

Kapitola se snaží zejména osvětlit pojmy, se kterými se dále pracuje, a zbytek práce je na ní postaven, proto je psána se snahou o co největší srozumitelnost, pokud možno i pro ty, kteří nejsou zvyklí na matematická vyjádření. Hlavní úlohu v další práci bude hrát kolmogorovovská komplexita a redundance (Kolmogorov, 1965; Li – Vitányi, 1990).

Na jazyk je pohlíženo jako na jednu z mnoha metod komunikace a na text jako na výsledek užití této metody. Prostor je věnován dvojímu pohledu na jazyk, který je možno zkoumat jak jako systém abstrahovaný od mluvčích (Itkonen, 2003), tak jako zobecnění vlastností jednotlivých mluvčích (Yngve, 1986, 1996; Yngve – Wąsik, 2004).

2 Lingvistika a věda vůbec

Text této kapitoly je epistemologickým úkrokem stranou. Volně navazuje na tři nejvýraznější větve moderních evropských přístupů k vědě (Popper, 1979; Feyerabend, 1975; Kuhn, 1997). Poměrně podrobného vysvětlení teorie komunikace v předchozí kapitole je zde využito k ozřejmění, že počínání vědců se z pohledu teorie komunikace neliší od počínání ostatních lidí ani tak metodou či cílem, jako spíše důsledností a důrazem na některé aspekty. Ptáme se zde zejména po významu a smyslu explanace, neboť vědecké vysvětlení je leitmotivem této studie. Kapitola dále ukazuje, že chápání lingvistiky jako vědy nemusí být tak přímočaře jednoduché, jak by se mohlo zdát (což je odvislé od dvojího náhledu na jazyk nastíněného v závěru předchozí kapitoly).

3 Logická nezbytnost morfémů a segmentace na vyšší celky

Po nezbytných úvodních kapitolách seznamujících čtenáře se základními pojmy a metodami se konečně dostáváme k jádru disertace, k pěti bodům vytyčeným v Úvodu.

Tato kapitola má za úkol čtenáře přesvědčit, že dělení textu na hlásky, slova, věty, souvětí a tak dále není nic samozřejmého, natož logicky nezbytného, a že tedy ústřední otázka této disertace je něčím, co je hodno zamyšlení.

Kromě toho se tato kapitola věnuje nezbytnosti morfémů (respektive nějakých jednotek odpovídajících znakům) a distinktivním rysům.

4 Vnořená segmentace — její univerzálnost a variace

Tato kapitola čtenáře přivede k tomu, že přestože ona vnořená segmentace na slova, věty atd. není logickou nezbytností, rozhodně je jevem takřka univerzálním, vyskytujícím se v mnoha různých jazycích. Zároveň však kapitola ukáže, že univerzální je pouze vnořená segmentace jako taková, nikoli způsoby, jak je jí dosaženo, tedy že se jazyky velmi liší v tom, na jaké jednotky jsou segmentovány.

Od počátku je rezignováno na ideu přirozených kategorií (Haspelmath, 2011), nicméně kromě podkapitoly pojednávající o segmentaci, jak ji chápali a chápou lingvisté, kapitola obsahuje i podkapitolu o tom, jak je text segmentován jeho produktory, a to už od nejstarších dob.

5 Modely přenášení informace jazykem

Konečně pátá kapitola, těžiště této práce, se zabývá modely přenosu informací mezi lidmi, které onu vnořenou segmentaci mohou vysvětlit. Začíná od formalizovaného

popisu prostředí, které musela a musí překonávat mluvená řeč (zašuměný kanál), přičemž onen formalizovaný popis je přetaven v testovatelný model a ten je následně testován na reálných datech, která vznikla jako vedlejší výstup při tvorbě mluveného korpusu.

Kapitola dále přibližuje různé metody překonávání takovýchto prostředí, tak jak je známe z informatiky. Zde se ukazuje, že multidimenzionální vkládání redundance je nejen strukturně podobné tomu, jak jazyk vypadá a funguje, ale že je i efektivní při překonávání různých druhů zašuměných kanálů, které musí překonávat mluvený jazyk, což je ukázáno pomocí počítačových simulací.

Dále je popsáno, jakým způsobem tyto metody mohou být implementovány ve skutečném jazyce, a nakonec je nabídnut realistický formalizovaný model, z něhož vyplyne, že segmentace je pro efektivní přenos informací přes běžné typy kanálů výhodná. Což vede k funkcionálnímu vysvětlení toho, proč je text takto segmentovaný.

6 Příklady vkládání redundance na různých úrovních

Šestá kapitola představuje konkrétní implementace, tedy jakým způsobem je redundance vkládána v reálném jazyce. Text postupuje systematicky po jednotlivých úrovních a ukazuje příklady, jak je redundance vkládána. Ukazuje, že poezii můžeme chápat jako sekundární přenosový protokol nad jazykem, ovšem užívající obdobné metody jako jazyk. Tím se dotýká estetiky redundantních vyjádření, které je věnována celá jedna podkapitola.

7 Hranice segmentů

Hranice některých segmentů je schopen recipient textu často poznat bez jejich vyznačení. K vysvětlení tohoto jevu postulujeme delimitační informaci, která je vkládána společně s redundancí k přenášené informaci. Na předpokladu, že její množství nezávisí na délce segmentu, který delimituje, je postaven model pro Menzerathův vztah (Menzerath, 1928, 1954), který je přímou konkurencí modelu v lingvistice nazývaného Menzerath-Altmannův zákon (Altmann, 1978, 1980). Tento nový model je úspěšně empiricky testován na českém a arabském textu a uveden do souvislostí (kapitola je částečně založena na Milička (2014)).

Reference

ALTMANN, G. Towards a theory of language. *Glottometrika*. 1978, 1, s. 1–25.

ALTMANN, G. Prolegomena to Menzerath's law. *Glottometrika*. 1980, 2, s. 1–10.

ASHBY, W. R. Principles of the self-organizing system. In FOERSTER, H. – ZOPF, J. (Eds.) *Transactions of the University of Illinois Symposium*, s. 255–278. Pergamon Press: London, 1962.

BAR-HILLEL, Y. An Examination of Information Theory. *Philosophy of Science*. 1955, 22, 2, s. 86–105. ISSN 00318248.

FEYERABEND, P. *Against Method: Outline of an Anarchistic Theory of Knowledge*. Humanities Press, 1975.

HASPELMATH, M. The Indeterminacy of Word Segmentation and the Nature of Morphology and Syntax. *Folia Linguistica*. 2011, 45, 1, s. 31–80. doi: 10.1515/flin.2011.002.

ITKONEN, E. *What is Language? A Study in the Philosophy of Linguistics*. Publications in General Linguistics, 8. Turku: University of Turku, 2003. ISBN 951-29-2617-2.

KOLMOGOROV, A. Three Approaches to the Quantitative Definition of Information. *Problems Information Transmission*. 1965, 1, 1, s. 1–7.

KUHN, T. S. *Struktura vědeckých revolucí*. OIKOYMENH, 1997.

LI, M. – VITÁNYI, P. M. *Handbook of Theoretical Computer Science. Volume A: Algorithms and Complexity*, Kolmogorov Complexity and its Applications, s. 188–254. Elsevier; MIT Press, 1990. ISBN 0444880712.

MACKAY, D. J. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2005. Dostupné z: <www.inference.phy.cam.ac.uk/mackay/itila/>.

MENZERATH, P. Über einige phonetische Probleme. In *Actes du premier Congres international de linguistes*. Sijthoff Leiden, 1928.

MENZERATH, P. *Die Architektonik des deutschen Wortschatzes*. 3. F. Dümmler, 1954.

MILIČKA, J. Menzerath's Law: The Whole is Greater than the Sum of its Parts. *Journal of Quantitative Linguistics*. 2014, 21, 2, s. 85–99. doi: 10.1080/09296174.2014.882187. Dostupné z: <http://dx.doi.org/10.1080/09296174.2014.882187>.

POPPER, S. K. R. *Objective Knowledge*. Oxford university press, 1979.

SHANNON, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal*. 1948, 27, 3, s. 379–423. ISSN 1538-7305.

WOUTERS, A. G. Design Explanation: Determining the Constraints on What Can be Alive. *Erkenntnis*. 2007, 67, 1, s. 65–80. ISSN 1572-8420.

WOUTERS, A. G. *Explanation Without a Cause*. Disertace, Utrecht University, 1999.

YNGVE, V. H. *From Grammar to Science: New Foundations for General Linguistics*. John Benjamins Publishing Company, 1996. ISBN 90-272-21618.

YNGVE, V. H. – WĄSIK, Z. (Eds.). *Hard-Science Linguistics (Open Linguistics)*. Continuum, 2004. ISBN 08-264-6114-X.

YNGVE, V. H. *Linguistics as a Science*. A Midland book. Indiana University Press, 1986. ISBN 9780253334398.