Jazyky a počítač: překážky a možnosti

Eva Hajičová

MFF UK Praha

Historický pohled

- 1947: Warren Weaver, dekódování
- 1949: memorandum Translation
- 1952: první konference o SP, USA
- 1954: Georgetown Univ., R → A, 250 slov,
  - Paul Garvin
- 1955: začátek výzkumu SP v Evropě i v Rusku

Warren Weaver v dopise Norbertovi Wienerovi:

When I look at an article in Russian I say: „This is really written in English but it has been coded in some strange symbols. I will now proceed to decode it.“

Březen 1947

Historický pohled

- 1949: memorandum Translation
- 1947: Warren Weaver, dekódování
- 1952: první konference o SP, USA
- 1954: Georgetown Univ., R → A, 250 slov,
  - Paul Garvin
- 1955: začátek výzkumu SP v Evropě i v Rusku

Historický pohled (pokr.)

- 1957: Noam Chomsky: Syntactic Structures
- 1959: vznik O(AL)TSP na FF UK, Praha
- 1960: A → Č, SAPO
- 1960: Yoshua Bar-Hillel: FAHQMT not possible
- 1963: Funkční generativní popis (Petr Sgall)
- 1965: ALPAC Report
- 1976: METEO, Montréal

SAPO přeložil první českou větu

- 1957: SAPO, A. Svoboda, VÚMS
- 1960:
  - *The consonants have not by far been investigated to the same extent as the vowels.*
  - Souhlásky zdaleka nebyly prozkoumány do stejné míry jako samohlásky.

Historický pohled (pokr.)

- 1957: Noam Chomsky, Syntactic Structures
- 1959: vznik O(AL)TSP na FF UK, Praha
- 1960: A → Č, SAPO
- 1960: Yoshua Bar-Hillel: FAHQMT not possible
- 1963: Funkční generativní popis (Petr Sgall)
- 1965: ALPAC Report
- 1976: METEO, Montréal

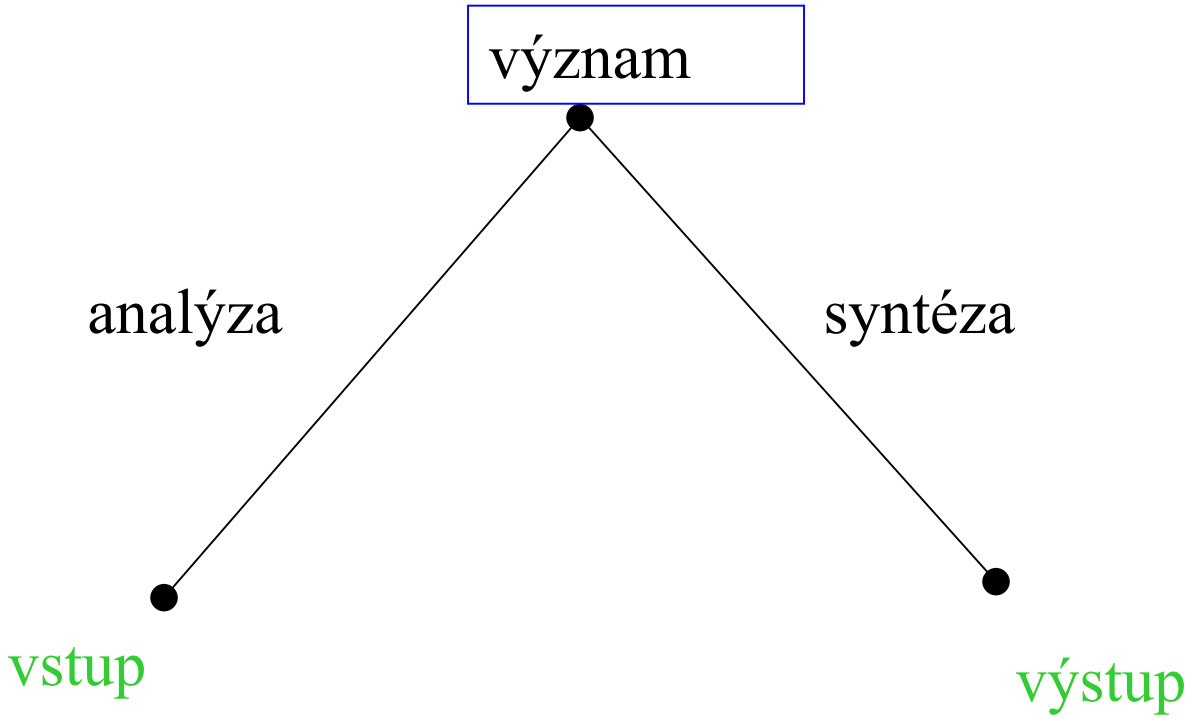
„Černá kniha“

- 1965: Automatic Language Processing Advisory Committee
- Doporučení:
  - Nesledovat utilitární přístup, ale soustředit se na vědecké otázky (computational linguistics), které mohou dát dobrý podklad pro budoucí „engineering enterprises“ (např. SP)

Historický pohled (pokr.)

- 1957: Noam Chomsky, Syntactic Structures
- 1959: vznik O(AL)TSP na FF UK, Praha
- 1960: A → Č, SAPO
- 1960: Yoshua Bar-Hillel: FAHQMT not possible
- 1963: Funkční generativní popis (Petr Sgall)
- 1965: ALPAC Report
- 1976: METEO, Montréal

Počítačové zpracování PJ



Zpracování mluvené řeči

- 1969: izolované rozpoznávání slov
  - „porovnávání vzorových obrazů“
- 1975: souvislá řeč, statistické metody - IBM
  - Tangora, Sphinx
- 1995: úspěšné komerční aplikace – Dragon
  - Naturally Speaking, technologie pro ViaVoice
- Po roce 2000: TU Liberec, ZČU Plzeň

Souvislá mluvená řeč a statistické metody

- Jelinek F., Bahl L.R., Mercer R.L.: Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech by Statistical Methods, IEEE Transactions 1975
- Baker J.K.: The DRAGON System – An Overview. IEEE Transactions, 1975
- Jelinek F.: Continuous Speech Recognition by Statistical Methods, Proc. of IEEE, 1976

Zpracování mluvené řeči

- 1969: izolované rozpoznávání slov
  - „porovnávání vzorových obrazů“
- 1975: souvislá řeč, statistické metody - IBM
  - Tangora, Sphinx
- Po roce 2000: TU Liberec, ZČU Plzeň
- 1995: úspěšné komerční aplikace – Dragon
  - Naturally Speaking, technologie pro ViaVoice

Potřebujeme lingvistiku?

- „*Whenever I fire a linguist our system performance improves*“
  - Fred Jelinek, 1988
- „*Some of my best friends are linguists*“
  - Fred Jelinek, Zampolli Award speech, 2004

Co je počítačová lingvistika?

- Martin Kay, schůzka v Rand Corporation 1965:
  - Computational linguistics
  - Mechanolinguistics
  - Automatic language data processing
  - Natural language processing

Co je počítačová lingvistika? (2)

- „Computational linguistics is trying to do what linguists do in a computational manner, not trying to process text, by whatever methods, for practical purposes“
  - Martin Kay, 2006
- Ale – vědní obory („komputační“) mají své aplikace
  - „good engineering requires good science“

Lingvistické aspekty počítačové lingvistiky

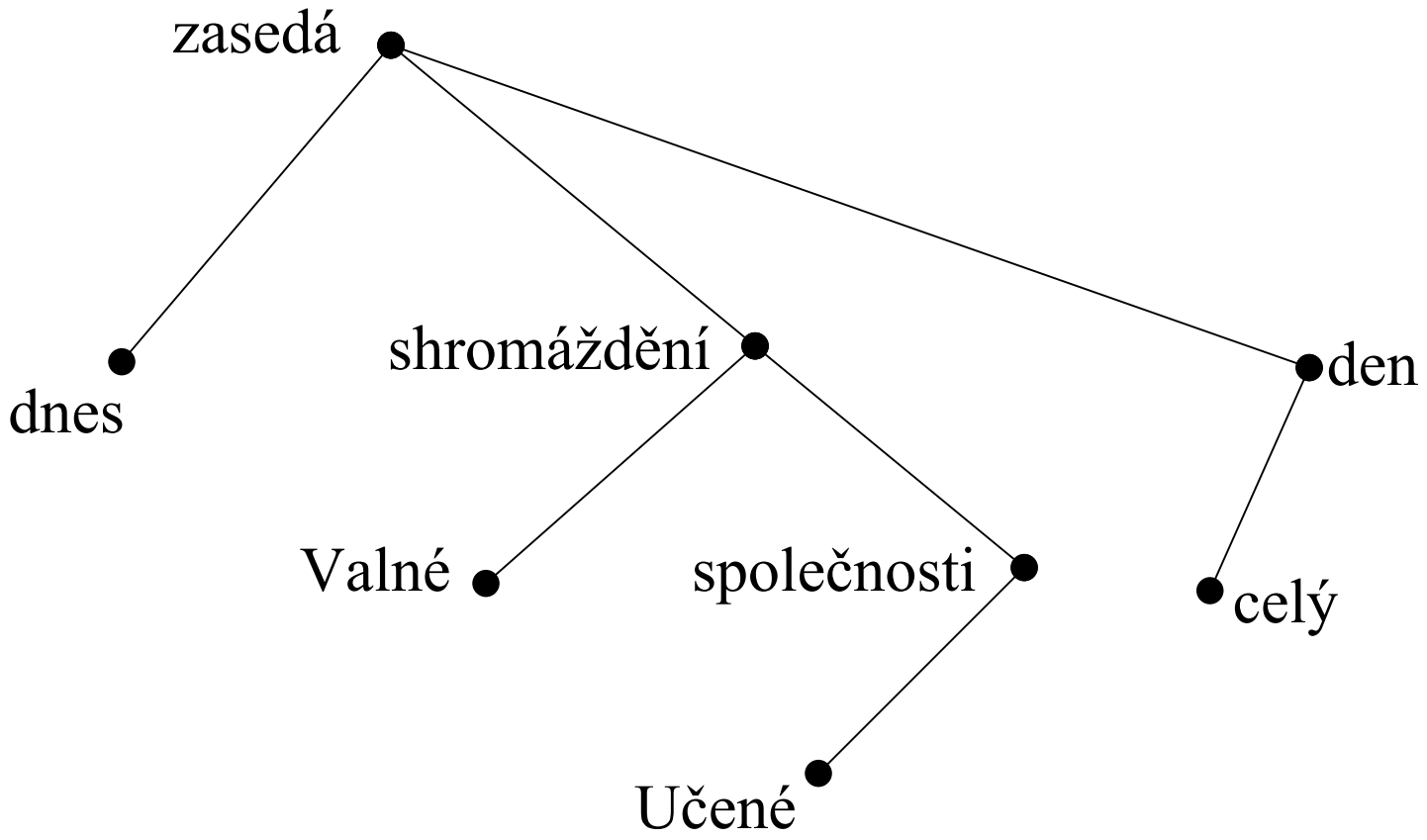
- Formální popis přirozeného jazyka
- Strukturní lingvistika nabízí dva modely:
  - Frázový (americký deskriptivismus)
    - N. Chomsky, generativní (transformační) gramatika
  - Závislostní („kontinentální“)
    - v Evropě už od 1836, Becker, Tesnière (PLK)
    - Funkční generativní popis (P. Sgall)

Porovnání základních syntaktických koncepcí

Dnes zasedá Valné shromáždění
Učené společnosti celý den.

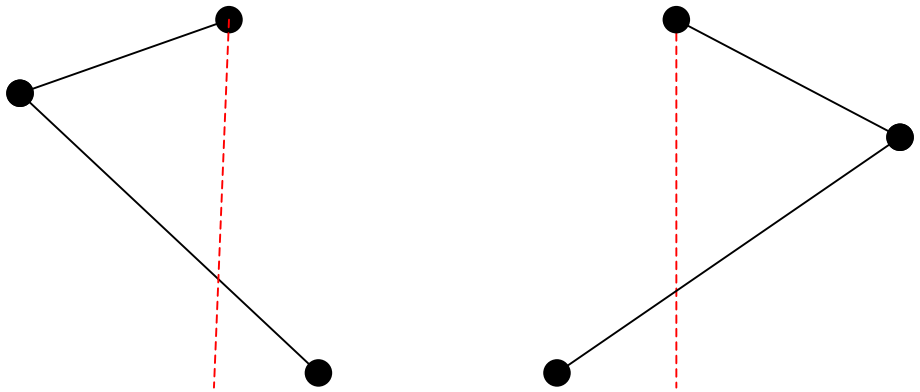
<Dnes {zasedá [(Valné shromáždění)
(Učené společnosti)] (celý den)}>

Závislostní struktura věty



Závislostní struktura věty

- vrcholový strom
- podmínka projektivity



Pražská škola: funkční pohled

- Aktuální členění věty je relevantní pro význam věty:
  - *Česky se mluví na Moravě. – Na Moravě se mluví česky.*
  - *O víkendech pracuji na své dizertaci. – Na své dizertaci pracuji o víkendech.*
  - *Dobrá zpráva: Češi udělali revoluci. – Špatná zpráva: Revoluci udělali Češi.*

Hloubková (podkladová) struktura věty

- Podobnosti mezi jazyky v hloubkové stavbě nápadnější i důležitější než jejich odlišnosti
  → studium hloubkové struktury

- N. Chomsky: language acquisition
  - Principy vrozené, parametry nastavené podle okolí

Jádro a periferie

- Jádro: projektivní strom s jedním vrcholem
  - Predikát a jeho argumenty
  - Blízko k lineárnímu řetězu
  - Struktura pro člověka zřejmě přirozená

Jádro a periférie (2)

- Periférie – složitá, velká oblast
- Jednotlivé vrstvy periférie – např. od různých vzorů ve skloňování apod. až po jednotlivé výjimky
  - *jít, jdu, šel*
  - víceslovná pojmenování (*vysoká škola*)
  - slovosledné varianty i porušení projektivity v povrchové podobě věty

Víceznačnost jazyka

(a) homonymie (jeden výrazový prostředek odpovídá více významům, funkcím)

(b) vágnost (jeden význam, ale vágní)

(a) *Nemocnice obviňuje ministerstvo z toho, že …*

(b) *Teď se už můžeme radovat*. (v tomto okamžiku? …, v naší době?)

Příklady homonymie

Nemocnice obviňuje ministerstvo, že …

– (i) … *nedostala včas příslušné finance.*

– (ii) … *čerpají prostředky nehospodárně.*

kdo koho obviňuje?

Pravopis:

Slepice honily holuby. – Slepice honili holubi.

kdo koho honil?

Příklady homonymie

… komentoval spisovatel Arnošt Lustig skutečnost, že mu Karel Gott odmítl schválit text už hotového knižního rozhovoru a navrhl, aby celou knihu napsali od začátku.

… komentoval spisovatel Arnošt Lustig skutečnost, že mu Karel Gott odmítl schválit text už hotového knižního rozhovoru, a navrhl, aby celou knihu napsali od začátku.

Víceznačnost v mluvené řeči

Víceznačnost v mluvené řeči

- Student: I didn't sleep all week, but I finished the paper for that conference in Italy …
- Prof. Smith: HUH??? What conference?
- Student: UM… I-I thought you said if I finished these texts early you would pay for fees, passage and room in Pisa for the meeting …
- Prof. Smith: Uh, no… I said I'd pay for a cheese, sausage and mushroom pizza for eating.

Možnosti

- Velké naděje na počátku (tisk!) 2001: Vesmírná Odysea, HAL: „*I'm sorry, Dave, I'm afraid I can't do that*."

- Vědci, lingvisté – rezervovaní

- Hledání metod: metody založené na pravidlech nevedly rychle k cíli → statistické metody

Statistické metody jako východisko

- Warren Weaver v Memorandu 1949: ruský text vidí jako zakódovaný anglický text, překlad jako vyluštění kódu
- Claude Shannon (1948) – „engineering perspective“: kritickou charakteristikou sdělení je pravděpodobnost, s jakou je sdělení vybráno z různých množin alternativ, nikoliv vlastní obsah sdělení

Aplikace v automatickém rozpoznávání souvislé řeči

Najít posloupnost slov $w_{1,n}$, která maximalizuje součin

$$P(a|w_{1,n}) \cdot P(w_{1,n})$$

- $P(a|w_{1,n})$ = pravděpodobnost toho, že akustický signál je a, jestliže tou posloupností je $w_{1,n}$ (akustický model)
- $P(w_{1,n})$ = pravděpodobnost posloupnosti $w_{1,n}$ (jazykový model)

Pro a proti:

- Saffran et al., Science 1996: osmiměsíční dítě je schopno naučit se rozčleňovat plynulý tok řeči na slovní segmenty
- ACL 1990: 39 přednášek o parsingu, z toho 1 statistický → ACL 2003: 62 ku 48
- Problém nedostatečných dat (*sparse data*)
- Zkušenost lingvistů: přirozený jazyk je komplexní systém a realizace v textech nedává postačující informaci, abychom textům rozuměli

Názorový vývoj:

- Od bigramů a trigramů ke struktuře věty, dokonce i k hloubkové struktuře
- Lingvistická intuice spojená se schopností extrahovat informaci z textů určí strukturu modelů a jejich parametrizaci
- Spolehnutí na radu lingvistů při vytváření jazykových zdrojů

Vytváření jazykových korpusů

- 1967: Brown Corpus, Henry Kučera
- 1970: Lancaster-Oslo-Bergen Corpus
- 1982: morfologické značkování (Lancaster)
- 1983-1986: „treebank“ (syntaktické struktury), Lancaster
- 1992: UPenn treebank
- 1991-95: British National Corpus

Vytváření korpusu češtiny

- Od 2. pol. 80. let: Počítačový fond češtiny
- 1994: založen Ústav Českého národního korpusu (FF UK, Fr. Čermák)
- 1997: projekt Prague Dependency Treebank
- 2000: zveřejněn SYN2000 (100 mil.slov)
- Dnes:
  - k dispozici PDT 2.0 (3 úrovně, 3 168 dokumentů, 49 442 vět, 833 357 (výskytů) slov)
  - v „bance“ ČNK k dispozici texty o 300 mil. slov
  - Pražský mluvený korpus, Brněnský mluvený korpus aj.

Příklad na závěr:

- Multilingual Access to Large Spoken Archives (MALACH) – od r. 2000
- Řešitelé projektu:
  - IBM Thomas J. Watson Research Center
  - University of Maryland
  - Johns Hopkins University, Baltimore, USA
  - Karlova univerzita v Praze (ÚFAL MFF UK)
  - Západočeská univerzita v Plzni

Základ projektu:

- Shoa Visual History Foundation (dnes: Univ. of South California Shoa Foundation Institute for Visual History)
- Kulturní dědictví, Steven Spielberg
- 116.000 hodin digitalizovaných „*interviews*“
- 32 jazyků
- 52.000 mluvčích (přeživších, zachránců, svědků)
  - Česko 566, Slovensko 656, USA 19.841, Israel 8.504, Ukrajina 3.433, Polsko 1.438, …

Cíl projektu:

- (a) automatické rozpoznávání řeči
- (b) počítačem podporovaný překlad (tezaurus)
- (c) prostředky pro zpracování přirozeného jazyka k automatické tvorbě metadat
- (d) podpora pro účinnou katalogizaci
- (e) podpora pro vyhledávání a využití informace

Dobrý příklad spojení vědeckého výzkumu s konkrétními úkoly!

- Doklad, že V. Jamek (Vesmír, únor 2006) nemá pravdu, když píše, že:

  … názorovou absencí, alibismem a hodnotovou rozbředlostí se vyznačují vědy, které mají člověka a společnost přímo v popis práce, obory společenské a „humanitní“.

- Lingvistika (nejen počítačová) si je vědoma „své odpovědnosti za budoucnost země a lidstva“