# Hierarchical Graph for Machine Translation

**Vít Baisa**

DTEDI seminar

spring 2011

# Motivation

- Language resources for MT questionable (WordNet, VerbaLex, PDT).
- Context is crucial but we cannot handle it properly.
- Much data + simple algorithms vs. sparse data + complex algorithms.
- Even fundamental concepts are vague: what is word, meaning, (well-formed) sentence, good translation?
- People can talk and even translate without any linguistic knowledge.
- Everything we know we have learned.
- Meaning is not intrinsic: structure of a sentence does not suffice: *Žvoulal si krákornul mášnou kulínou.*

# State of the Art general purpose MT

MT systems:

- Rule-based (morphology, syntax, semantics),
- statistical (Google),
- hybrid (Yahoo!),
- neural networks, . . .

# Known issues of MT

- Tokenisation: *doesn't* $\rightarrow$ does + n't or doesn + 't?
- Lemmatisation: *neměl* $\rightarrow$ mít or nemít?
- Morphologic analysis: *ženu* $\rightarrow$ hnát or žena?
- Syntactic analysis: *Karel mluvil o sexu s Marií*.
- Word Sense Disambiguation: *silný čaj* $\rightarrow$ *powerful tea*.
- Named entities: *Včera jsem viděl Královu řeč.*
- Multiword expressions: *vysoká škola* $\rightarrow$ *high school*.
- Metaphors, idioms: *Bez práce nejsou koláče* $\rightarrow$ *No pain no gain*.
- Anaphora: *Dej ji do vázy.* $\rightarrow$ *Put her in a vase.*

# Language as hierarchy of language units

- No distinction between morphology, syntax, semantics and reasoning.
- Elementary units are characters or phonems.
- Only one relation: $(s, t) \in R$ if $s$ is *said together* with $t$.
- Inductive definition of language:
  - *n* $\rightarrow$ *a*
  - *má* $\rightarrow$ *ma*
  - *mám* $\rightarrow$ *hlad*
  - *odešel* $\rightarrow$ (*protože* $\rightarrow$ *musel*)
  - *když se nenamažeš krémem* $\rightarrow$ *spálíš se*
- Very robust: $(3 \rightarrow (+ \rightarrow 7)) \rightarrow (1 \rightarrow 0)$
- Equivalent with lambda calculus?

# Hierarchical graph

- Nodes are language units.
- If $s$ and $t$ are LU then $s \to t$ is LU.
- Several types of edges:
  - Constituency: *mám* $\to$ *hlad*
  - Equivalency: *bych* $\rightleftharpoons$ *would*, *ps* $\to$ *dog*
  - Partial forward constituency: *m* $\rightsquigarrow$ *hlad*
- Meaning: $\mathrm{meaning}(s) =$ set of neighbours of $s$ in graph.
- Synonymy: $\mathrm{synonymous}(s, t) \Leftrightarrow \mathrm{meaning}(s) \cap \mathrm{meaning}(t) \neq \emptyset$

# Properties I

- Formal approach.
- Absolute majority of words (and all sentences) can be divided into two parts.
- Easy knowledge representation: *Johann Sebastian Bach se narodil v roce 1685.*
- Upper levels are equivalent for various languages (we all think in very similar ways) – interlingua.
- Simple treatment of complex grammatical constraints and constructions:
  - *Chci (aby to věděl)* $\rightarrow$ *I want (him to know)*
  - *Nic nemám* $\rightarrow$ *I have nothing*
- Dictionary and grammar within single data structure.
- There is linear increase between neighbouring levels.
- The more we know the better we memorize (latin, music, math).

# Properties II

- Passive vs. active knowledge of vocabularies.
- Cimrmans theory of externalism.
- We understand if we know context: *XYZ s r. o.*
- Synonymy on all levels:
  - ý – ej,
  - pěkný – hezký,
  - Odejdi – Běž pryč.
- Meaning on all levels:
  - *í*
  - *eji*
  - *Karel*
  - *nejím maso*
- Discreteness on all levels:
  - *mš*
  - *bát*
  - *Petr vyřešil*

# Goals

- Get learning data: we need simple phrases.
- Building and tuning of hierarchical graph for Czech and English.
- Implement simple algorithms for learning, understanding and translating.
- Standard evaluation + manual evaluation by comparing with state-of-the-art MT systems.
- MT between very different languages (Hungarian, Japanese . . . ).
- Derive standard relations and rules from the graph (synonymy, hyperonymy, subject predicate agreement, . . . ).