# Towards the Realistic Natural Language Representations

## Examples and Experience

Petr Sojka

Laboratory of Electronic and Multimedia Applications and
Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

sojka@fi.muni.cz

RASLAN, December 7th, 2013

# Outline and Take-home Message

# An Example of Natural Language Processing

zuzana en-gram neverit vhlava mit

# Levels of Natural Language Processing

1. strings as information bearing chunks (Vítek)
2. words (morphology)
3. phrase and sentence parsing (syntax)
4. word usage in context, word sketches, collocations, named entity recognition, word meaning, phrase and sentence meaning, narrative representations (semantics)
5. natural language understanding and dialogue, information (knowledge) retrieval (pragmatics)

Mostly *discrete, non-distributional* representations: grammars, dictionaries, thema/rhema segmentations, word sketches.

# Failures of Natural Language Processing

- Machine translation
- Semantic web
- Turing test

# Successes of Natural Language Processing

- Watson, Siri: machine learning of continuous representations
- distributional semantics (LSI, LDA)
- random walking in texts (Page's rank), availability of computational resources (global, data intensive and driven computation)
- human computation with fun (Luis von Ahn)!

# Inspiration from Human Computers (Psycholinguistics Evidence)

- evidence for priming, semantic priming, "on-the-fly computation from distributed linked representation".
- Hoey's lexical priming theory: each occurrence of lexical item enforces 'priming' of it given a co-locational context
- "new encounter either reinforces the priming or loosens it": Hebbian learning principle of synaptic neural net learning
- locality principle (by magnetic resonance)
- subconsciousness processing (during sleep); one-frame a advertisment
- functionality, expressiveness, behavioural change depending on "spike trains" of internal (beer, marijuana, engagement) or external stimuli (conscious or unconscious learning): nothing carved forever as insect into amber – equilibrium dynamics
- language subjectivity (entity meaning)

# Towards Computational Lexical and Semantic Primings?

- what representations should be used?
- appropriate data structures for given algorithm and application is the key (c.f. OOP)
- empirical linguistics, natural language data (e.g. text or speech corpora) as valuable asset and data to deliver natural language application functionality

Doug Laney (Chief Data Officer): "We have only two assets: money and data. If we don't treat data as an asset, then it's an expense."

Peter Norvig (Google Research President): "We should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data."

# Unreasonable Effectiveness of Data

$$\text{Conscious} \rightarrow \text{Unconscious}$$

$$\text{Discrete} \rightarrow \text{Continuous}$$

# Unreasonable Effectiveness of Language Data
# Representations Computed from Corpora

- introvert/extrovert evaluation from tweets or facebook status changes
- opinion mining from texts by random walking in texts (Sebastiani)

from "scholastic" discrete language representations (word lists, word nets, word sketches, graphs, logics, parsing trees) into continuous representations: syntax 'only' encodes information structure, is secondary; meaning is primary

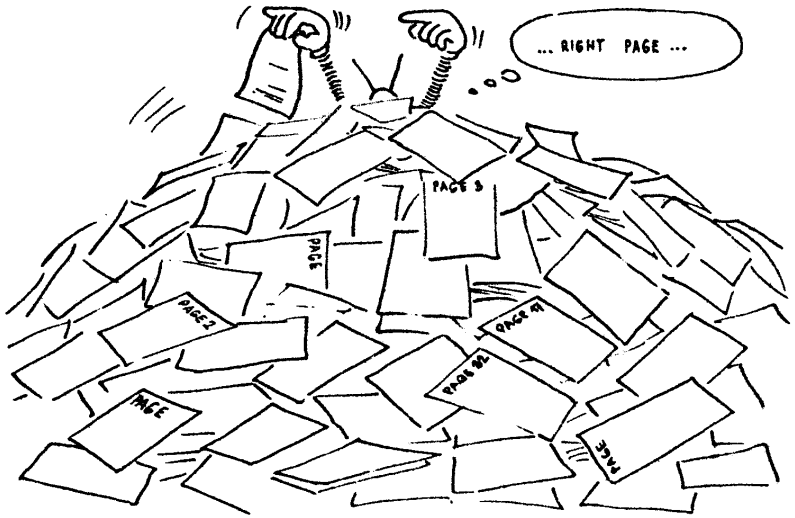
as natural language is subjective, so it make sense to use personal/thematic corpora for choosing and interpreting data

# Word Meaning Representation Computation from Corpora

- random walking (Pagerank): a document is supported when cited, and credit propagates (iterative computation over graph/matrix) ending in continuous credit evaluation computed from graph/matrix cluster
- a word meaning computed from corpora is supported when used, and word meaning propagates, ending in continuous word meaning evaluation from given (personal, domain) corpora

Although computationally extensive, use cases in [math] IR, word and formulae disambiguation,. . .

# Coping with Information Overload by Filtering of *Big Data*



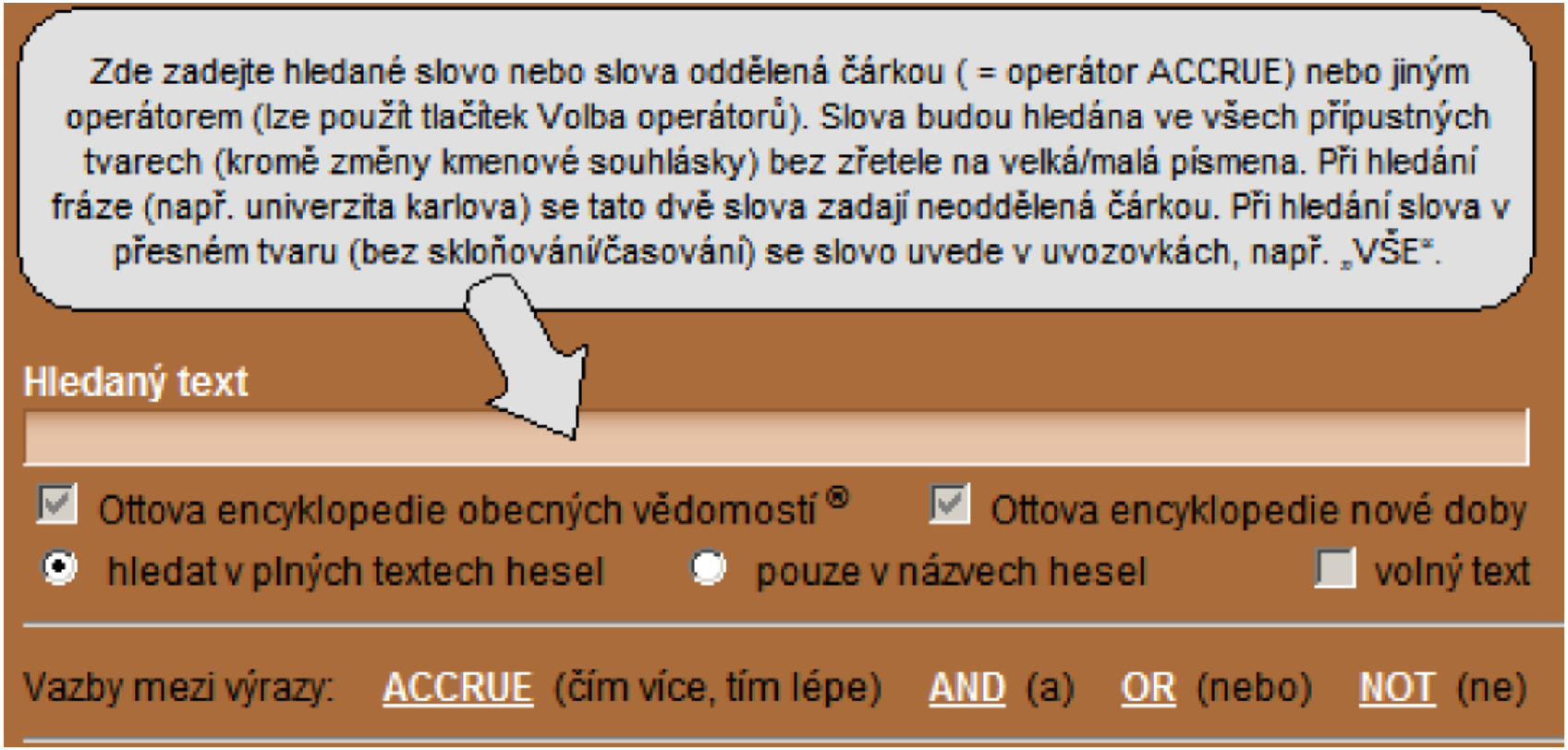

Life is searching, searching is *killer app* in information society: group similar and narrow focus of (faceted) search in [your] Big Data using natural language knowledge representation.

Different types of representation useful: for plagiarism word/entity/citation *n*-grams, for narrative search trains of thought signatures (similarity of semantic, meaning representation in space), for topical exploratory search distributional semantics.

# Exploratory [Math] Search Trends (Daniel Tunkelang)

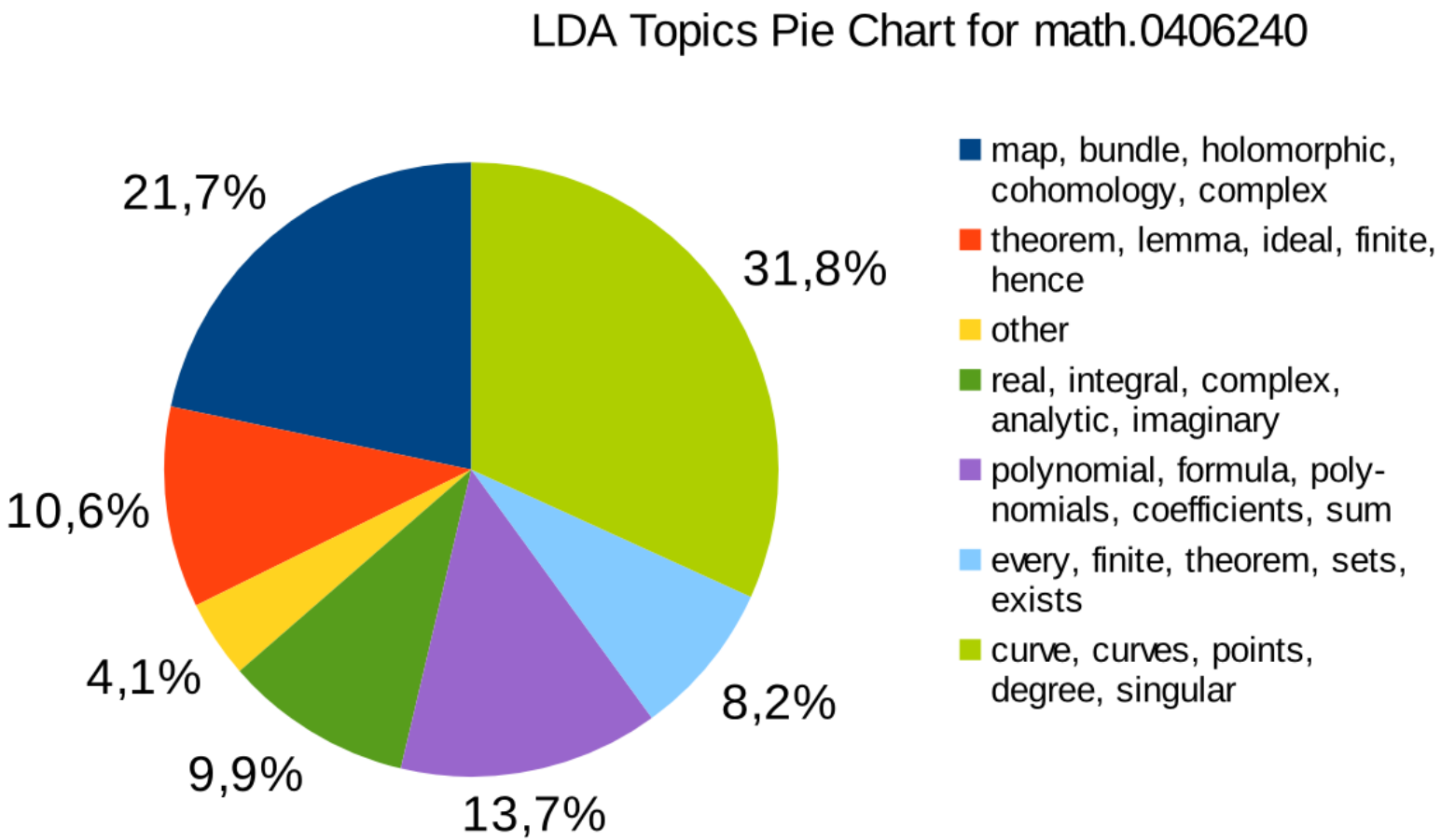
- Entity-oriented search (NER, word $\rightarrow$ entity)
- Knowledge graphs (KG (Google), Satori (Microsoft), Freebase), Linkedin, Facebook)
- Search assistance, suggestions, dialogue

# NER Search Example: Project Ottův Slovník naučný



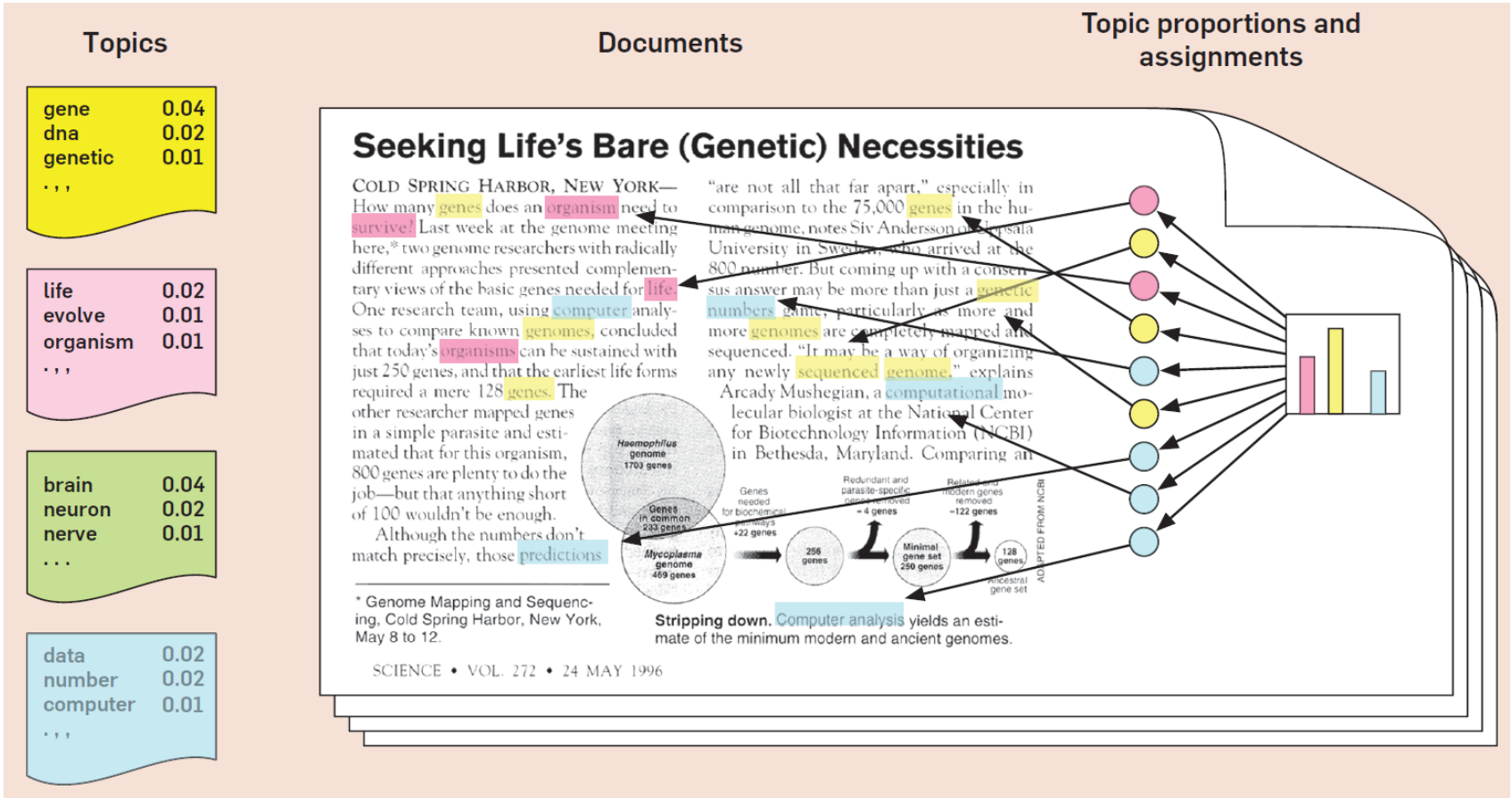

Grabbing the essence (content) of documents → topical modelling.

# Leading Edge Example: Automated Meaning Picking from Texts



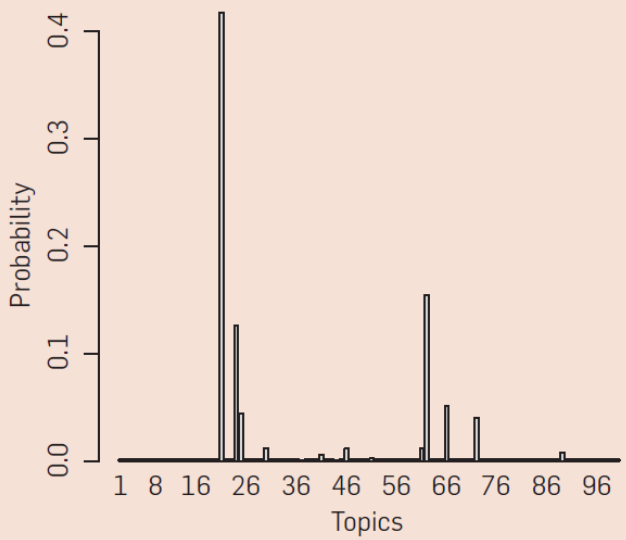
# Probabilistic Topical Modeling: Latent Dirichlet Allocation

- topic: weighted list of words
- document: weighted list of topics



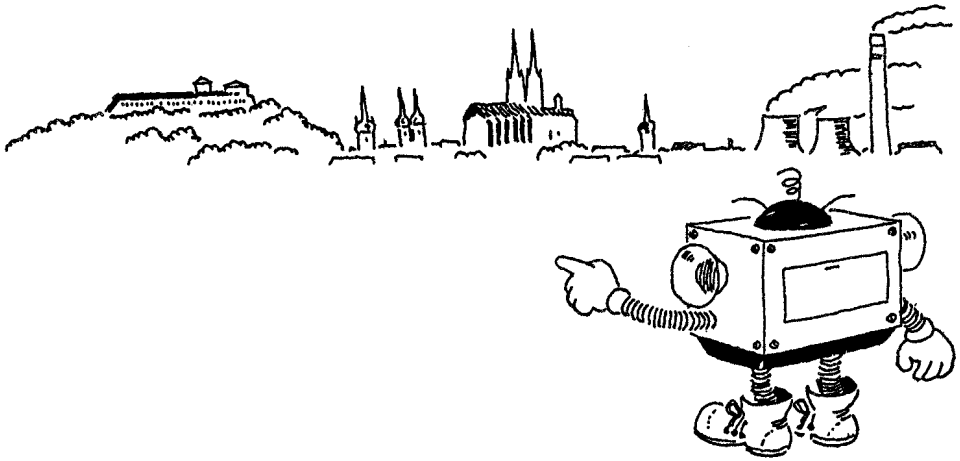
# Topical Modeling: Latent Dirichlet Allocation II

- all topics computed automatically from document corpora




| **"Genetics"** | **"Evolution"** | **"Disease"** | **"Computers"** |
|---|---|---|---|
| human | evolution | disease | computer |
| genome | evolutionary | host | models |
| dna | species | bacteria | information |
| genetic | organisms | diseases | data |
| genes | life | resistance | computers |
| sequence | origin | bacterial | system |
| gene | biology | new | network |
| molecular | groups | strains | systems |
| sequencing | phylogenetic | control | model |
| map | living | infectious | parallel |
| information | diversity | malaria | methods |
| genetics | group | parasite | networks |
| mapping | new | parasites | software |
| project | two | united | new |
| sequences | common | tuberculosis | simulations |

# Conclusions, Summary



- evidence for better NL representation from psycholinguistics
- from discrete to continuous language representations
- data-driven, empirical linguistics: the unreasonable effectiveness of distributive, continuous, language representations computed from linguistic data like corpora
- use cases for exploratory search, disambiguation in information retrieval

Credits: Jiří Franek, (illustrations); Daniel Tunkelang, David Blei (pictures)