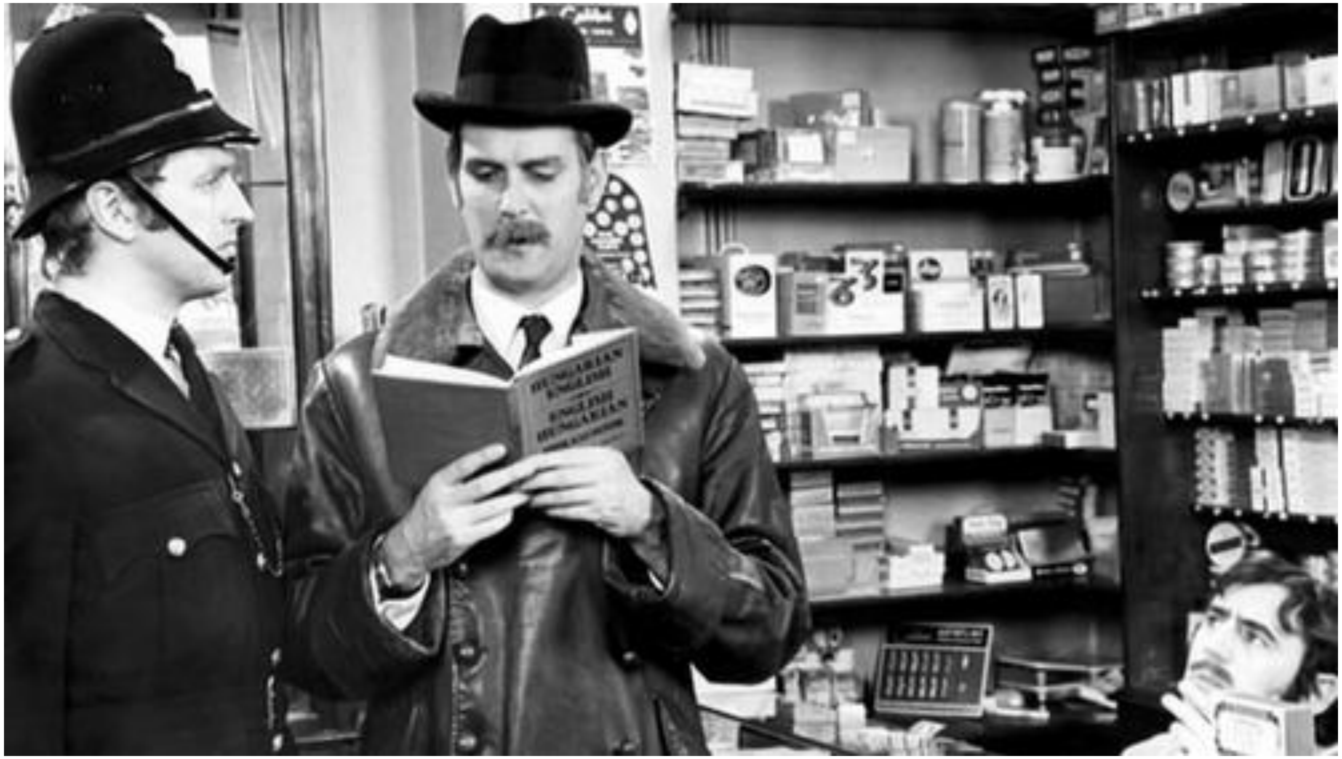
Language data for digital natives: old wine in a new bottle or...?

Simon Krek

Amebis, d.o.o., Kamnik, Slovenia

„Jožef Stefan“ Institute, Slovenia

… *„and now for something completely different“*



Vision 1: **A Language-Transparent Web and Media**

- Cross-lingual information access to the web and to media in all languages
    - 200 to 1000 languages, crosslingual queries, automated question-answering, natural language search, conversational agents, automatic translation of chats, tweets and e-mails.
- Multimedia multi-language subtitling
    - subtitles for television programmes in real time, audio and video translation
- Making documents understandable
    - rephrasing complicated documents, automatic summarisation, language generation

Vision 2: **Natural and Inclusive Interaction**

- Natural interaction with agents and robots
    - self-learning, contextaware, personalised agents with speech, language and multi-modal input and output abilities; low-level tasks – processing e-mails, voice messages or telephone calls
- Assistive applications
    - personalized speech technology systems for persons with reduced motor control; sign language recognition, synthesis and translation
- Cross-lingual E-learning
- Cross-lingual meeting assistants
    - instant speech-to-speech translation; transform slides, presentations and handwritten notes into a preferred language; minutes automatically produced, video recordings automatically indexed to support voice searching, transcription and translation

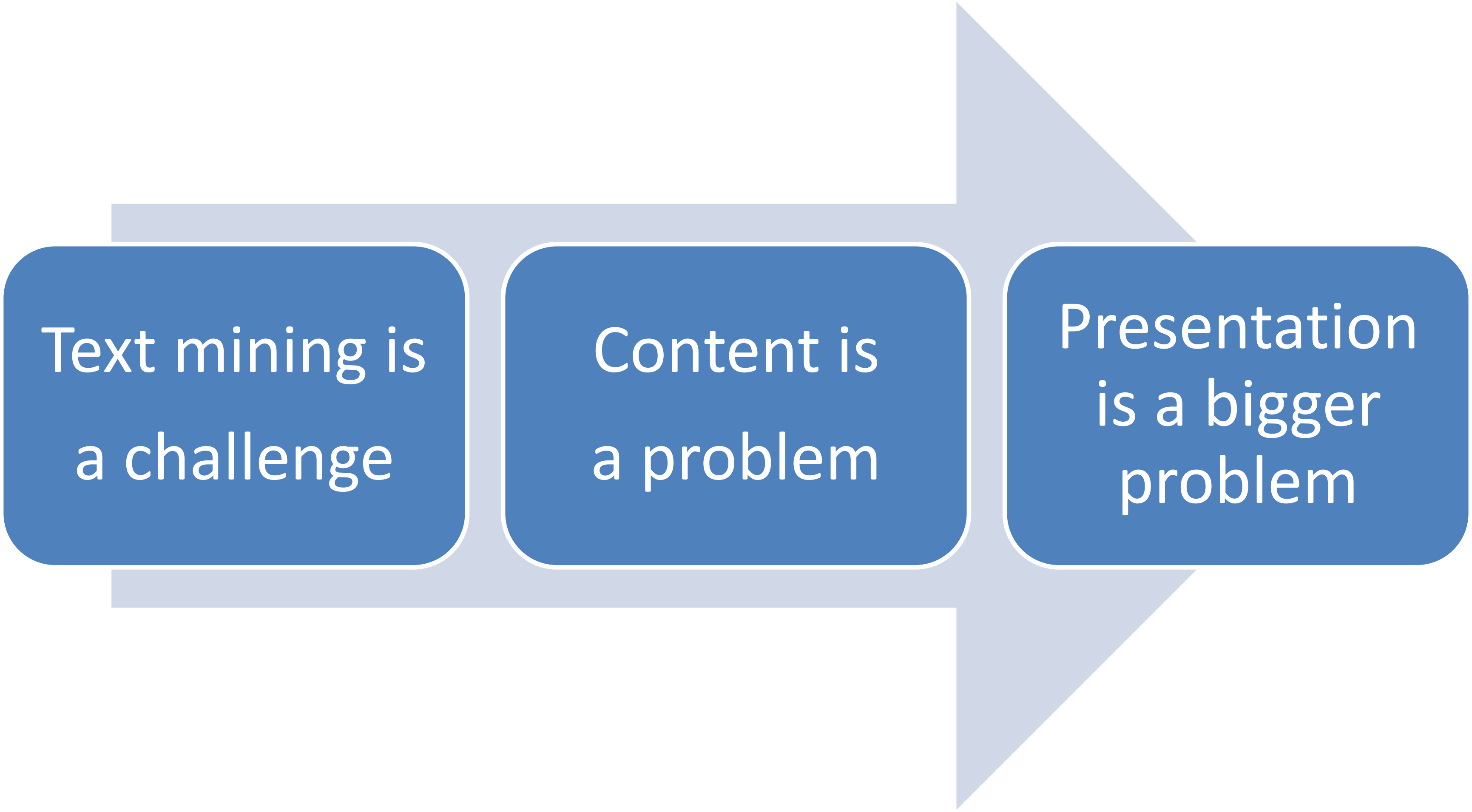
Vision 3: **Efficient Information Management**

- Federated multilingual audio-visual search
    - search for audio/video materials across languages; identification of objects, persons and actions; speech recognition of ordinary (untrained) voices; semantic analysis of audio and video content
- Personalised information assistants
    - filing documents, reformatting materials, copying information from one document to another, preparing standard letters and answering information requests
- Life logging
    - capture every utterance and conversation during the day; semantically structuring the information into meaningful bits and pieces

The world of digitization

- **December 6, 2010** – Today Google unveiled Google eBooks, a new service for buying and reading digital books. Google eBooks is cloud-based.
- **May 19, 2011** – Amazon announced today that it’s now selling more Kindle e-books than both hardcover and paperback books combined.
- **July 05, 2011** – South Korea has announced that it plans to replace textbooks and all paper in its schools with tablets by 2015.
- **October 11, 2011** – Redesigned Europeana launched: Europeana underwent a significant makeover. The new Europeana is now more visual, interactive and easier to use.

e-Lexicography 2011



Text mining is a challenge

Content is a problem

Presentation is a bigger problem

Text mining is a challenge

- Automatic – detection, extraction, retrieval, selection, etc.
- Types of information:
  - definitions
  - collocations
  - examples
  - synonyms
  - multiword expressions
  - phraseology
  - etymology etc. etc.

world wide web

digitized present

digitized past

Text mining issues

- Language technology is a lively field but narrow lexicographically relevant text mining activity is absent, at least on a large scale
- Apart from the investment required the reason may be the lack of consensus about the Content/Presentation (cf. Wikipedia)
- Double-edged sword: information is cheap, but meaning is expensive

Meaning is expensive

- Dealing with „meaning“ is an incredibly difficult task for computers (the field of AI)
- Basic task of lexicography has everything to do with meaning
- BUT: explaining it to human users, not to computers
- The question: if NLP is interested in meaning (cf. IBM Watson), is there a place for e-lexicography in this massive effort?

Jellyfish

Sinclair: Floating dictionary (2001)

- »A few years ago I felt that the time was ripe to plan a new kind of dictionary, one that would never exist on paper, but would be automatic or almost automatic in its selfupdating.
- It would, so to speak, float on top of a corpus, rather like a jellyfish, its tendrils constantly sensing the state of the language.
- As well as reporting on the settled usage and meanings of the words and phrases of a language, like a normal dictionary does, the floating dictionary, when interrogated, dips into the corpus and checks this information, offering instances that match its criteria for the senses; also it explores further to see if there are any instances that conflict with the criteria, and may signify a development of a sense or the emergence of a new usage altogether.
- Within the limits of its powers, it organises this evidence as a comment on the existing dictionary entry.«

Technologies involved

- information extraction
    - ... reporting on the settled usage ...
    - ... offering instances that match its criteria ...
- large scale text mining
    - ... tendrils constantly sensing the state of the language ...
- word sense disambiguation
    - ... match its criteria for the senses ...
    - ... signify a development of a sense ...
- text generation / visualization
    - ... a comment on the existing dictionary entry ...
- BUT: for a very specific purpose

Definition as a showcase

- Definition extraction
  - the world (wide web) is full of defining language: textbooks, Wikipedia, general, digitized texts etc.
- Definition generation
  - paradigm shift: from educating with difficult definitions to explaining with simple ones
  - identification of the ideal definition (whole sentence definitions?, semagrams?, lexical constellations etc.)
  - for individual users (or types of users)

Is it happening?

- **International Workshop On Definition Extraction** held in conjunction with the International Conference RANLP - 2009, 14-16 September 2009, Borovets, Bulgaria
- Kobyliński, L. and Przepiórkowski, A. 2008. **“Definition extraction with balanced random forests.”** In Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008, Springer Verlag, pp. 237-247.
- Roberto Navigli and Paola Velardi, **„Learning Word-Class Lattices for Definition and Hypernym Extraction“.** Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1318–1327, Uppsala, Sweden, 11-16 July 2010.
- **and more...**

But...

„...Definition extraction is the task of automatically identifying definitional sentences within texts. The task has proven useful in many research areas including **ontology learning**, **relation extraction** and **question answering**. However, current approaches – mostly focused on lexicosyntactic patterns – suffer from both low recall and precision, as definitional sentences occur in highly variable syntactic structures.“

Text mining summary

- e-lexicography has not experienced a truly large scale text mining effort to create the whole range of lexicographic content (semi-)automatically (?)
- text mining wish list
  - what was done will be there (on the web, digitized)
  - what is there can be extracted
  - what is not there should be done ☺
  - what changes is interesting
  - what changes can be detected

Contents – Format

- Dictionaries as Language Resources
  - automatic acquisiton of lexical information from Machine Readable Dictionaries ('80s+)
  - parsing definitions (+other dictionary data) to produce „knowledge“ for Language Technology
  - EAGLES/ISLE, PAROLE, SIMPLE, many more ...
  - Lexical Markup Framework (ISO 24613:2008): morhpology, syntax, semantics, multi word patterns, multi-lingual notations, MRD etc.

WordNet-LMF format

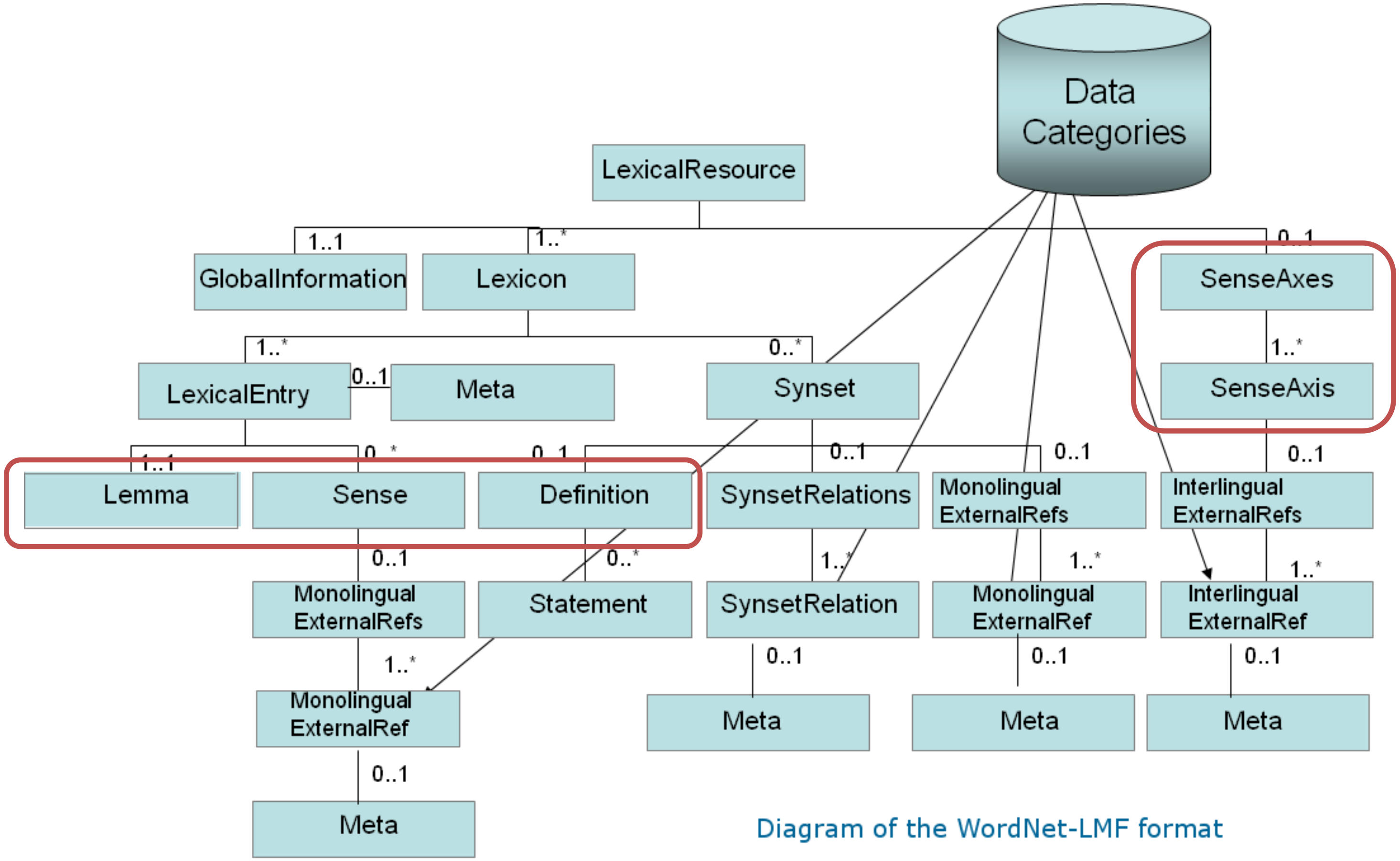


Diagram of the WordNet-LMF format

Text Encoding Initiative

- Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form.
- **9 Dictionaries:** This chapter defines a module for encoding lexical resources of all kinds, in particular human-oriented monolingual and multilingual dictionaries, glossaries, and similar documents.
- The elements described here may also be useful in the encoding of computational lexica and similar resources intended for use by language-processing software; they may also be used to provide a rich encoding for wordlists, lexica, glossaries, etc. included within other documents.

Now

- LT community now has a basic idea how to store various types of information
- also SW community: RDF, RDFa, RDFS, OWL, SKOS, and more
- standardization in human-oriented dictionary encoding was never really successful (XML, TEI?)
- the question is: if different types of lexicographic information intended for human users will have to know each other – will the format be dictated by LT standards? (Probably yes.)

Success stories

- Language Resources and Evaluation Conference, Malta 2010
  - Wikipedia: 16 papers
  - WordNet: 17 papers
  - FrameNet: 11 papers
- NLP: WordNet is for nouns, FrameNet is for verbs
- Wikipedia and WordNet (not so much FN) are used by both sides: human users and NLP

Contents – Information

- Do we know if e-dictionaries are „liked“?
- Studies on e-dictionary use
  - assessment of usability (Heid)
  - eye-tracking (Tono)
  - mouse movement, keystroke, gesture logging?
- Monitoring (web) log files
- These activities are performed on the existing types of (lexicographic) information

„And now for something...“

- Are users able to describe what kind of information they need about language?
- Do we have mechanisms to identify these needs?
  - with all the future LT machinery
  - in real time
  - on a large scale
- How about new media, social networks etc?

Example

- „Communication in Slovene“ project
- Internet „Style Guide“ portal
  - language information „hot spot“
  - explanations of more difficult parts of language (ortography, lexis, grammar) for lay public
  - links to corpora, dictionaries, lexical databases, lexicons etc.
- web crawling forums and web sites dedicated to language issues
- analysis of about 1,500 questions and discussions

Results

category	type of information	percent
A	ortography	33,4%
B	pronunciation	2,2%
C	morphology	16,3%
D	word formation	4,4%
E	lexis	26,3%
F	syntax	7,8%
G	text composition	2,9%
H	other	6,6%

Contents summary

- Can we expect that digital natives will have the patience to distinguish between different types of language information containers? No.
- Does that mean that it is time to think about a more universal information database providing different kind of language data? Yes, **all** of them.
- How much time do we have to provide this information? Not much, measured in seconds.

Presentation

- Codex format: 1,500-year tradition
  - alphabetization
  - thumb indexing
  - sense numbering
  - typography & layout
  - menus, signposts etc.
- User is left to his/her book selection and browsing skills to find information

E-dictionary

- Database:
  - – headword search
  - – full-text search
  - – advanced search
  - – contextualized search
  - – multiple choice interface etc.
- User is left to his/her database selection and searching skills to find information

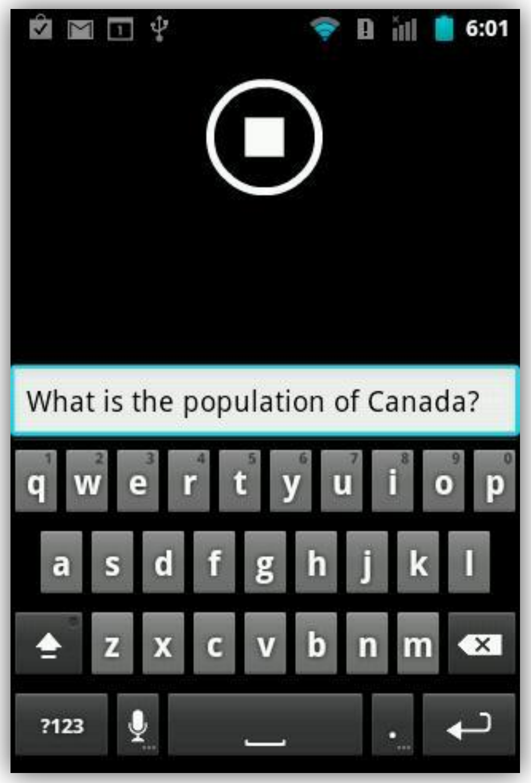
„And now for something...“

- What if we reversed the situation?
- The user just has to express the need for information about a language problem
- Contextaware(ness)
  - what am I doing? travelling, studying, browsing web etc.
  - what am I reading? sports, history, finance, physics textbook etc.
  - who am I? Slovene speaker, dentist, student, learner of French etc.

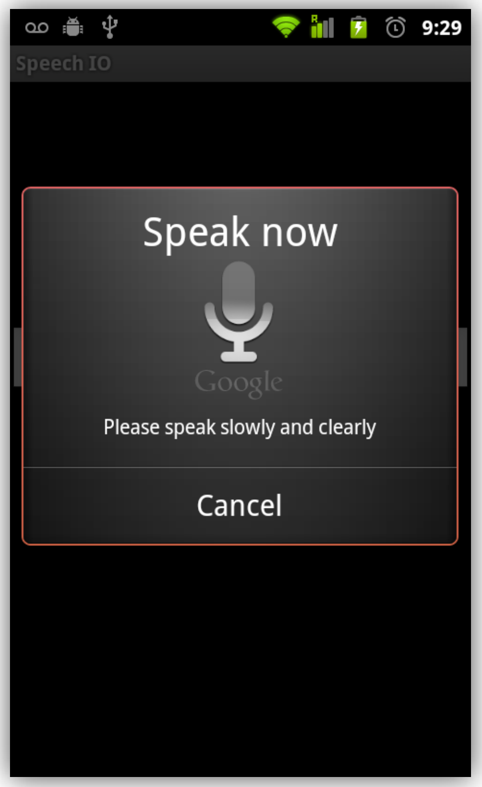
Expressing the need

- input type
  - question answering
  - conversational agents etc.
- input mode
  - mouse, keyboard (physical, touchscreen)
  - voice recognition (Apple, Android)
  - OCR (Abbyy)
  - gesture (Kinect) etc.

„Question answering“ mode

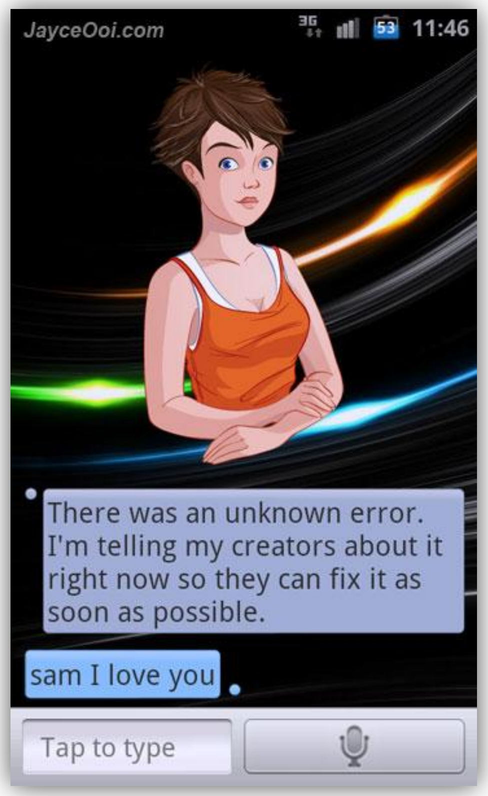


What is the population of Canada?



Speak now

Please speak slowly and clearly



There was an unknown error. I'm telling my creators about it right now so they can fix it as soon as possible.

sam I love you

„Browsing & visualization“ mode

- Three-dimensional model of related information about languge (WordNet visualizations?)
- combination of textual and all other kind of information
- like browsing Google Earth: Language(s) as Earth and hapax legomena as appartments in a building

Conclusions

- A new interest in language technology is emerging, related to the requirements of the information society
- EU seems to be preparing a coordinated action to get back on track after the success of North American companies
- Is it possible to jump on the bandwagon, with a conceptual break with tradition in lexicography?