CZECH NATIONAL
CORPUS

Introduction to Text Corpora and Their Applications

# Corpora in language learning and teaching

Lucie Chlumská, Ph.D.

lucie.chlumska@korpus.cz

# OUTLINE:

1. **LECTURE**

- principles of corpus-based teaching

- applications

    - direct v. indirect

- open questions and criticism

    - appropriate models of language (> appropriate corpora?)

2. **SEMINAR**

- reading (O'Keeffe et al.): *From Corpus to Classroom: Introduction*

- How have corpora influenced language teaching?

# LECTURE

# Corpora in Language Teaching

# Corpora in language teaching

- beginnings: late eighties, early nineties

- scope
  - mostly focused on English
  - LSP (ESP), EFL/ESL...

- mainly tertiary level of education (primary and secondary level rather neglected...)

- main principles:
  - authenticity of language data
  - frequency-based information

# Corpora in language teaching

- CANCODE = Cambridge and Nottingham Corpus of Discourse in English

https://www.nottingham.ac.uk/research/groups/cral/projects/cancode.aspx

- ICLE = The International Corpus of Learner English

https://www.uclouvain.be/en-cecl-icle.html

CZECH NATIONAL CORPUS

# Corpora in language teaching

- Leech 1997: convergence between teaching and language corpora; three focuses

  1. the direct use of corpora in teaching

     - teaching about, teaching to exploit and exploiting to teach

  2. the indirect use of corpora in teaching

     - reference publishing, materials development, language testing etc.

  3. further teaching-oriented corpus development

     - LSP corpora, L1 developmental corpora, L2 learner corpora

CZECH NATIONAL CORPUS

# Applications: direct use

# Direct use

- *teaching about*
  - teaching corpus linguistics as an academic subject like other sub-disciplines of linguistics (both undergraduate and postgraduate)

- *teaching to exploit*
  - providing students with hands-on know how so they can exploit corpora for their own purposes

- *exploiting to teach*
  - using a corpus-based approach to teaching language and linguistics courses which would otherwise be taught using non-corpus methods

# Syllabus design

- Sinclair & Renouf (1988): *„lexical syllabus"*
  - designed around the most recurrent lexical patterns in language
  - however, grammar is not neglected

- lexical syllabus focuses on:
  a) the commonest word forms in a language
  b) the central patterns of usage
  c) the combinations which they usually form

- covers all aspects of language differing from a conventional syllabus only in that the central concept of organization is lexis

# How to teach?

- syllabus design and material development are closely associated with the principles of teaching
- traditional approach in language pedagogy > three Ps
    1. PRESENTATION
    2. PRACTISE
    3. PRODUCTION

- exploratory approach > three Is
- originally proposed by Carter and McCarthy 1995 to teach spoken grammar
    1. ILLUSTRATION = looking at real data
    2. INTERACTION = discussing and sharing opinions and observations
    3. INDUCTION = making one's own rule for a particular feature

# DDL

- Tim Johns 1991: DDL = data-driven learning
- student as a „language detective" or a „researcher"

  *„Research is too serious to be left to the researchers."*

- teacher-directed or learner-led (i.e. discovery learning)

- three stages of this inductive learning based on DDL principles:

  1. observation (of concordance evidence)
  2. classification (of salient features)
  3. generalization (of rules)

- bottom-up induction (authenticity and discovery) v. top-down deduction

CZECH NATIONAL CORPUS

# Applications: indirect use

# Developing reference materials

- DICTIONARIES

  - *Collins COBUILD English Language Dictionary* (1987)

  - …

- GRAMMAR BOOKS and TEXTBOOKS

  - *A Comprehensive Grammar of the English Language* (1985)

  - OUP: *Natural Grammar*

  - CUP: *Grammar in Use* series, *Vocabulary in Use* series

  - …

# Open questions and issues

# Resources

- What corpora/data to use?

1. **written language dominance** v. reality
   - spoken language features (discourse markers, elipsis, hedges, vague language…)

2. **native speaker models** (e.g. British English) for everyone?
   - lingua franca as a solution?

3. authenticity v. **teachability**
   - is it even possible to teach the language as it appears (especially in spoken register)?

4. **simplification** of language for certain purposes?
   - child language as a model for young learners?

# Thank you for your attention!

# Questions?

# SEMINAR

# Reading

common reading:

O'Keeffe et al. (2008): *From Corpus to Classroom*. Cambridge: Cambridge University Press.

suggested resources:
www.wordandphrase.info
(for wordlists, academic language and genre disctinction)

phrasesinenglish.org
http://phrasesinenglish.org/explore.html
(BNC-based n-gram generator)

CZECH NATIONAL CORPUS

# Discussion

- What are the main differences between traditional textbooks and corpus-based teaching materials?

- What does DDL stand for?

- What are the main principles used in exploratory learning?

- How can corpora be used in creating teaching material?

- What are the limitations of corpus-based approach in language teaching?

- In which areas of language teaching are corpus-based approaches and materials most widely used?

- What corpus resources are needed to support L1 teaching? Is there any difference from L2 teaching?

# Inspiration

- SkeLL for English, Czech and Russian

  - web-based resources

    - *the previous sentence does not relate to the following sentence*

    - *sentences of first concordances should be better than following ones in the point of less non-alphabet characters and interpunctions, more frequent words, etc*

  - provides basic information about the KWIC

    - sample of 40 concordance lines

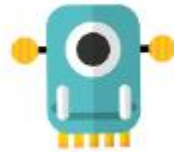    - basic word-sketches (syntactically informed collocations), distribution

https://cskell.sketchengine.co.uk/run.cgi/skell

CZECH NATIONAL CORPUS

# Inspiration

- **Umíme česky (We know Czech)**
  - a lot of interactive exercises for free

Zkuste to netradičně

| Vtipy | Roboti | Střílečka | Tetris | Příšerky |
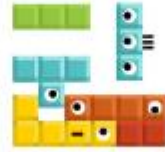
Pro více hráčů

| Závody | Týmovka | Územíčka |

https://www.umimecesky.cz/

# Inspiration

- Analyze your own text in Wordandphrase

  - frequency-based colouring of words in your text

  - academic English – uses lists of terms based on COCA

  - includes information from a corpus as well as WordNet

  - shows basic distribution in genres and concordances

  https://www.wordandphrase.info/

# Inspiration

- many other tools

  - especially for English:

    - VocabGrabber: https://www.visualthesaurus.com/vocabgrabber/

    - Visual Thesaurus: https://www.visualthesaurus.com/

      - similar principle as in wordandphrase

    - Englicious: http://www.englicious.org/

      - focused mainly on grammar

      - approved by the UK National Curriculum 2014

  - and many others...

CZECH NATIONAL CORPUS