CZECH NATIONAL
CORPUS

CZECH NATIONAL CORPUS

ORATOR: A new corpus of spoken Czech

David Lukeš, Zuzana Laubeová

Presenting on behalf of…

- … colleagues who co-designed the corpus and coordinated data collection:
  - Zuzana Laubeová
  - Marie Kopřivová
  - Petra Poukarová



- … many external contributors who provided and transcribed data

OVERVIEW

1. WHAT KIND OF DATA?
2. A LOOK AT THE DATA
3. ORATOR FACTS & FIGURES
4. WHERE DOES ORATOR FIT IN?
5. CONCLUDING REMARKS

WHAT KIND OF DATA?

Spoken vs. written language

- or oral vs. literate
- the CNC has previously focused on prototypical spoken language:
  - spontaneous, multi-lateral, private
- ORATOR data is on the cusp between the two:
  - prepared in advance to some degree, but not read
  - mostly unilateral, but audience present
  - public, but often restricted (typically not nationwide)

Kopřivová M., Laubeová Z., Poukarová P., Lukeš D.: Relevant criteria for selection of spoken data: Theory meets practice. Jazykovedný časopis, 2019, no. 70, pp. 324-335.

A LOOK AT THE DATA

Comparison with ORTOFON

- See for yourself at https://kontext.korpus.cz
  - ORTOFON v1: https://kontext.korpus.cz/first_form?corpname=ortofon_v1
  - ORATOR v1: https://kontext.korpus.cz/first_form?corpname=orator
- (ORATOR v2 not publicly available yet but results presented below are based on it)

Comparison with ORTOFON

- hesitations – `[word="@+"]`:
  - ORTOFON v1: 9611.75 i.p.m.
  - ORATOR v2: 30,393.8 i.p.m.
  - ~3× more frequent in ORATOR
- demonstratives – `[tag="PD.*"]`:
  - ORTOFON v1: 79,517.48 i.p.m.
  - ORATOR v2: 65,508.86 i.p.m.
  - slightly more common in ORTOFON

Comparison with ORTOFON

- long demonstratives – `[tag="PD.*" & word=".{5,}"]`:
  - ORTOFON v1: 5,891.59 i.p.m.
  - ORATOR v2: 7,671.22 i.p.m.
  - slightly more common in *ORATOR*
- top of the frequency list of word forms:
  - *a/and* vs. *to/that* (1:2 in ORTOFON, 1:1 in ORATOR)
  - prominence of different discourse markers: *no/well* (ORTOFON) vs. *vlastně/actually* (ORATOR)

ORATOR FACTS & FIGURES

ORATOR at a glance

	ORATOR v1	ORATOR v2
available when?	already out	by the end of 2020
# tokens	736,407	1,535,609
# tokens w/o punctuation	578,398	1,207,255
| vocabulary |	60,952	97,816
# recordings	318	489
# speakers	332	468
total audio length	72 hours	149 hours

Composition and metadata

- gender:
  - male: 1,077,365 tokens
  - female: 440,415 tokens
- source:
  - freely available recording: 866,732
  - recorded specifically for ORATOR: 668,877
- genre annotation compatible with the CNC's written corpora

Composition and metadata

- 12 situation types, including:
  - lecture: 1,204,619 tokens (both easy to come by & long)
  - public assembly: 63,894 tokens
  - meeting: 49,442 tokens
  - tour: 43,074 tokens
  - …
  - ceremony: 12,658 tokens
  - sermon: 9,088 tokens
  - closing speech: 8,151 tokens

Composition and metadata

- framing / context:
  - popularization: 812,646 tokens
  - scientific: 361,754 tokens
  - professional: 178,224 tokens
  - official: 164,021 tokens
  - political: 18,964 tokens

WHERE DOES ORATOR FIT IN?

Czech National Corpus context

- ORAL series corpora
  - intimate discourse
- ORTOFON
  - intimate discourse
  - newer material
  - separate phonetic transcript layer
- DIALEKT
  - dialectological recordings
  - separate dialectological transcript layer

Czech context

- DIALOG
  - TV debates
- MONOLOG
  - Czech radio presenters
  - both institutionally supported by the Czech Language Institute of the Czech Academy of Sciences
- SCHOLA: school communication
- PMK & BMK: the inspiration for the ORAL series

International context

- spoken BNC1994
  - both intimate discourse and “context-governed encounters”
- Corpus Gesproken Nederlands (CGN)
  - a mix of informal, formal and even simulated speech
- Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK)
  - inspired by spoken BNC1994, similar to CGN
- Slovenský hovorený korpus (SHK)
  - long-running project, regular releases; also a mix
- (unnamed?) corpus of conversational Polish (Spokes interface by PELCRA)

CONCLUDING REMARKS

Future outlook

- see also https://wiki.korpus.cz/doku.php/cnk:orator
- a corpus of spontaneous speech on Czech radio (dialogues)
  - both guests and presenters
- a spoken supercorpus based on CNC data
  - cf. spoken BNC1994
  - more variation → more interesting comparisons
  - ORTOFON + ORATOR + radio + ...?

ACKNOWLEDGEMENTS

The design and compilation of CNC corpora is made possible by the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

Thank you for your attention!