Towards a learner corpus of Czech for Polish speaking students

Elżbieta Kaczmarska

University of Warsaw

Workshop on interoperability of L2 resources and tools
University of Gothenburg, Sweden
6–8 December 2017

Institute of Western and Southern Slavic Studies

- **Czech**
- Slovak
- Bulgarian
- Croatian
- Serbian
- Slovenian

From a teacher's point of view…

A learner corpus of Czech (for Polish speaking students)

Why Czech? (1)

- With more than 50 new students every year, Czech is the most popular language taught at the Institute of Western and Southern Slavic Studies of University of Warsaw

- It is also the second most popular Slavic language chosen by Polish students (at universities and private schools in Wrocław, Katowice, Sosnowiec, Opole, Racibórz, Poznań and elsewhere).

A learner corpus of Czech (for Polish speaking students)

Why Czech? (2)

- Polish L1 – Czech L2 → closely related languages; strong L1 interference at all levels (pronunciation, morphology, syntax, lexicon – false friends, including phraseology, metalinguistic communication)

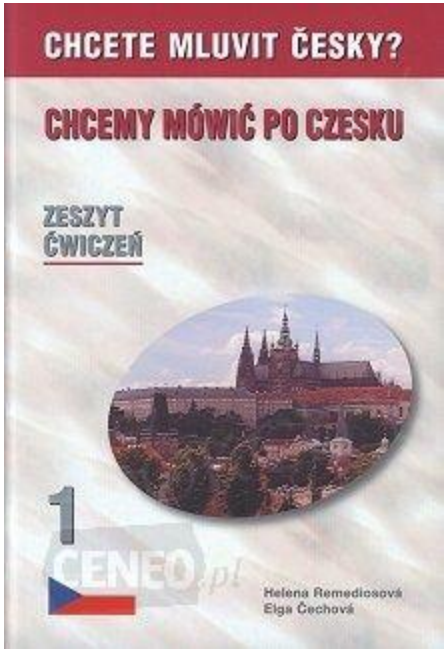
- to make teaching more efficient

* for teaching and learning – using Czech National Corpus, including its parallel section

A learner corpus of Czech (for Polish speaking students)

Why Czech? (3)

- no specialized materials dedicated for Polish learners of Czech



Facing the reality…

- **No tradition of building learner corpora in Poland**

* PELCRA Learner English Corpus (PLEC) by Piotr Pęzik (University of Łódź)

- **Lack of required infrastructure, specialists and sufficient financial resources**

* Team of Corpus Linguistics at Institute of Western and Southern Slavic Studies

(temporary) solution…

- We intend to build a corpus of Czech texts written by Polish learners by extending the L1 Polish – L2 Czech subcorpus of CzeSL (Czech as a Second Language), a learner corpus built at Charles University in Prague.

*More on that – Alexandr Rosen: *Trying make a learner corpus user happy: from annotation to search tools*

- The Polish – Czech subcorpus of CZeSL is quite small (77 texts, 15 th. words)
- It requires not only collecting new texts and cooperation with the CzeSL team, but also applying some subtle changes to the annotation system of CzeSL, considering the common mistakes made by Polish learners of Czech.

Common mistakes (1)

- Czech prepositions ***s*** and ***z*** **(in Polish – only *z*)**

cz	**Tom**	**pracuje**	**s**	**Michalem.**
pl	**Tom**	**pracuje**	**z**	**Michałem.**
en	Tom	works	with	Michal.

cz	**Robert**	**je**	**z**	**Polska.**
pl	**Robert**	**jest**	**z**	**Polski.**
en	Robert	is	from	Poland.

Common mistakes (2)

- Genetive after negation in Polish (Accusative in Czech):

cz	**(já)**	**Nemám**	**čas.**
pl	**(ja)**	**Nie mam**	**czasu .**
en	I	do not have	time.

- Polish: writing ***nie*** (*no*) with verbs separately – Czech: writing it together
- word order
- repetitive errors in declination and conjugation
- and more…

What we have…

If the Polish-Czech learner corpus was adequate…

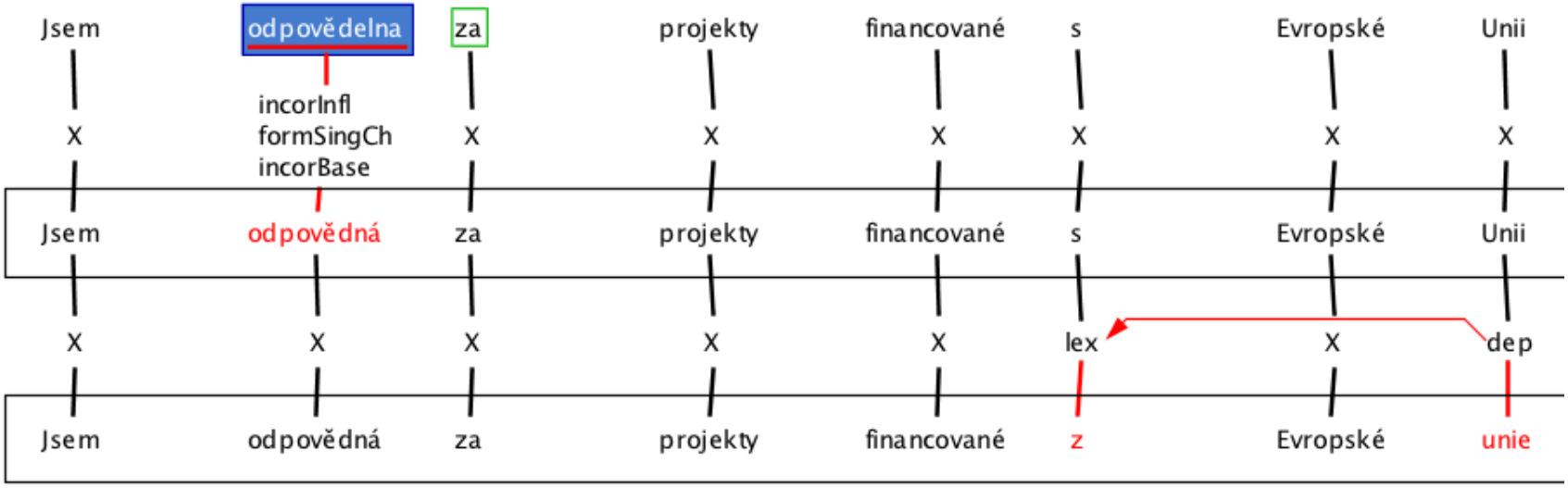
- Searching in **KonText –** a search engine available on CNC pages

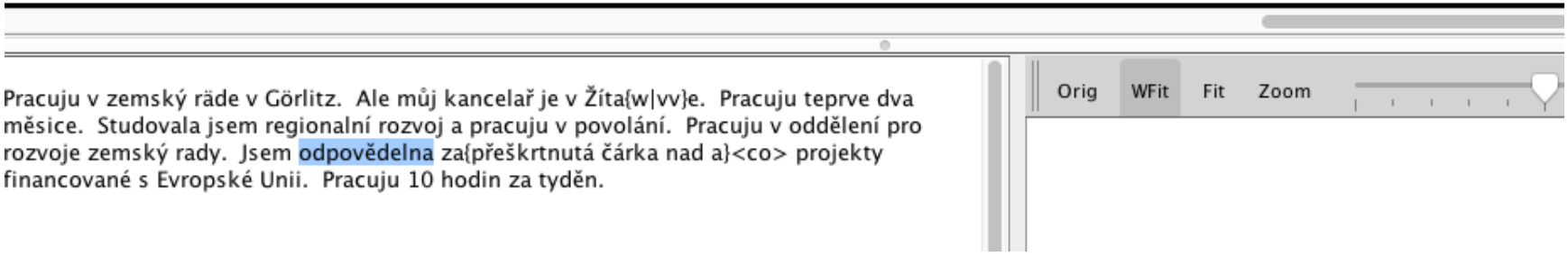
- **Feat –** anotation editor using by CzeSL with 2 level anotation

I am responsible for projects financed by the European Union.

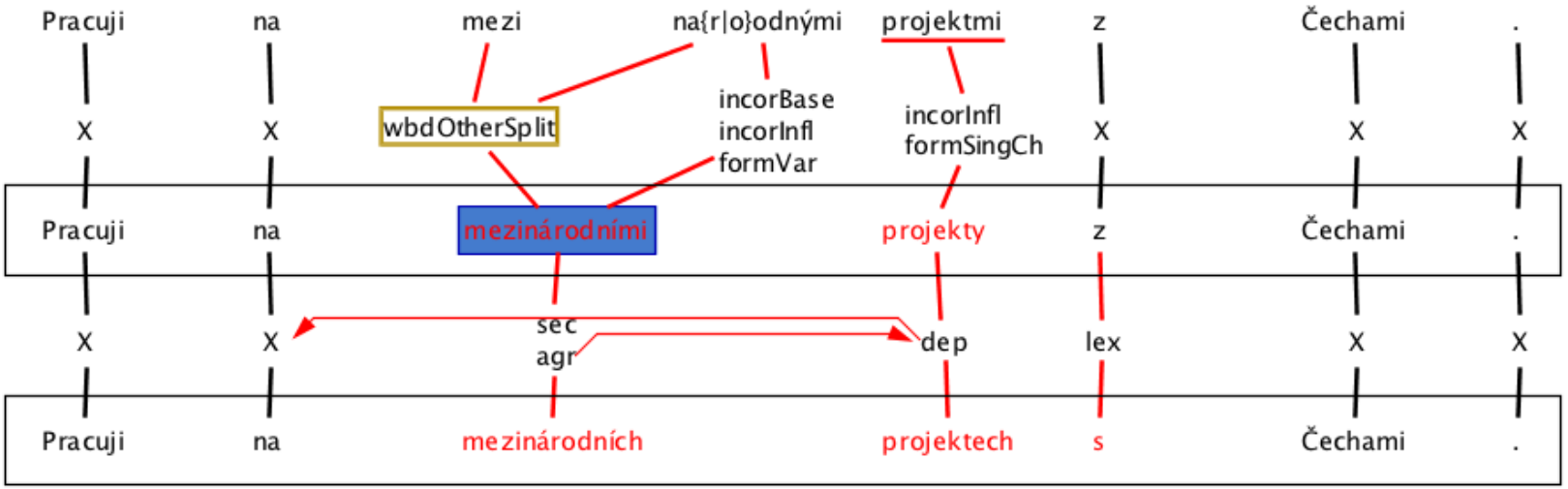
z, 25, pl, A2+

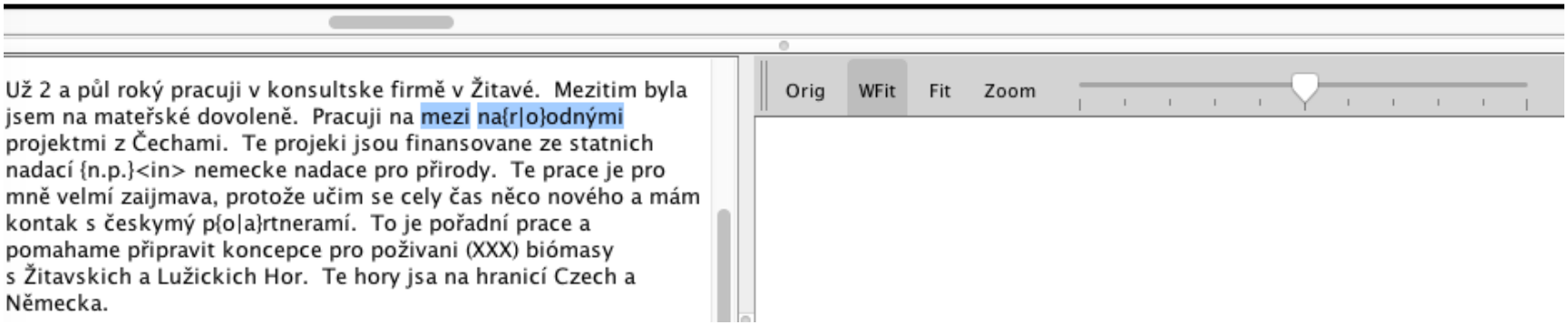
<s>Jsem [1:incorInfl|incorBase|formSingCh odpovědelna>odpovědná] za projekty financované [2:lex s>z] Evropské [2:dep Unii>unie] .</s>





I work on international projects with Czechs.





… and what we need

- To record texts and sentences after corrections (by a native speaker of Czech)

Free word order in Polish and Czech – slight difference of meaning and slightly different accentuation :

pl *Alena spała w domu sama.*

pl *Alena w domu spała sama.*

pl *Sama spała Alena w domu.*

pl *W domu Alena spała sama.*

- Lärka – such a tool possible also for Czech

First step(s)

- Negotiations with a foreign partner – CNC and CzeSL team

- We started a student project of collecting text in Czech language written by Polish students (Gabriela Gawrońska) + metadata

- The students will be taught how to rewrite collected texts and to anotate them
- the transcription should be done by a person with Polish L1

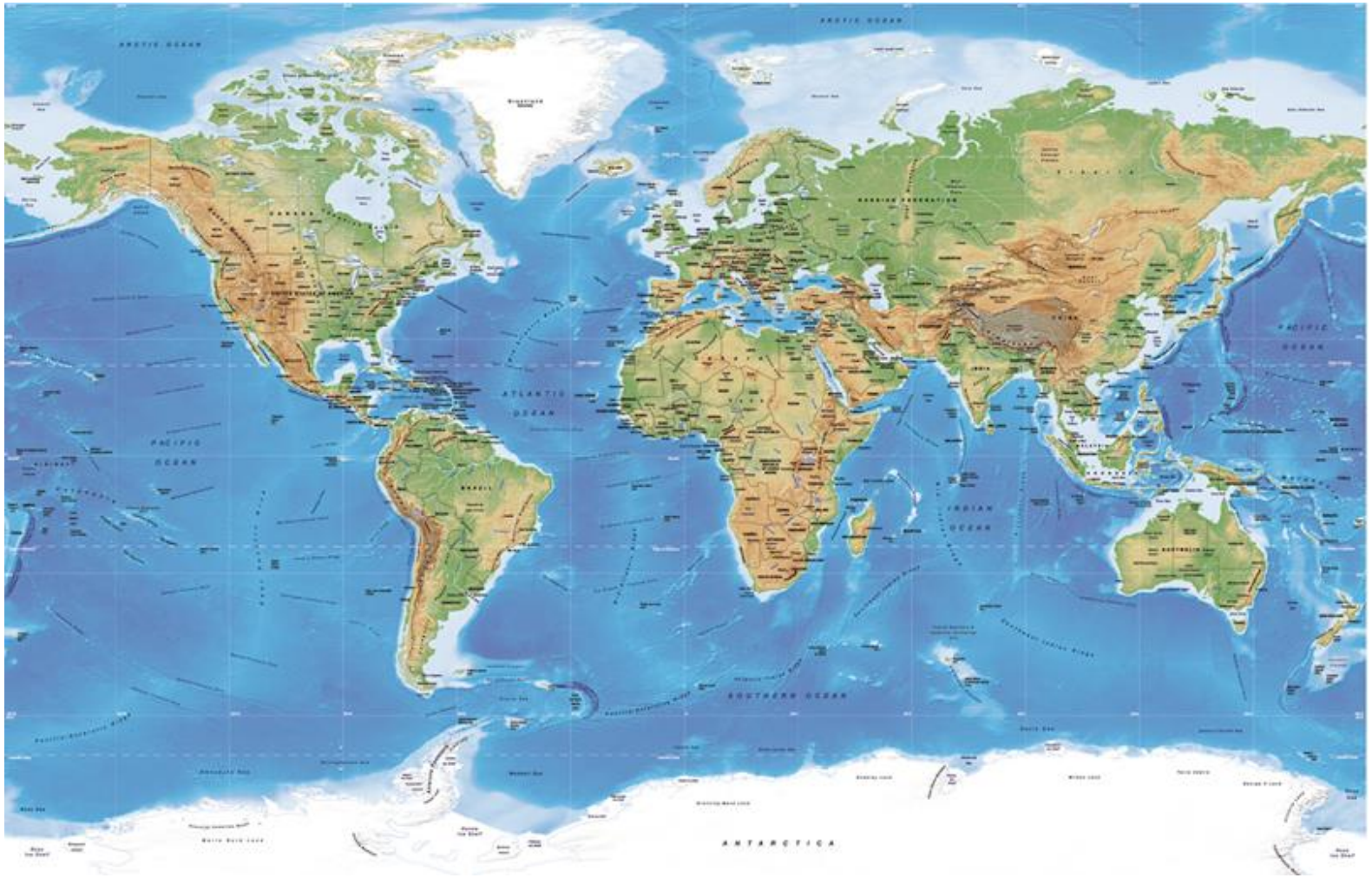
z, 22, pl, B1

<s>Vždycky [2:odd ,>] když se [2:dep mně>mě] někdo zeptá , co je pro [2:dep mně>mě] nejdůležitější v životě , [1:incorBase|formQuant0 vzpominám> [2:use vzpomínám>vzpomenu]] si na rozhovor [2:dep ze>se] známým polským filozofem [1:incorBase|formCaron1 Leszkěm>Leszkem] Kotakowskim , který jsem četla asi před dvěma lety .</s>

Our hopes

- motivating scientists to conduct research and scientific analyses (Polish L1 - Czech L2) and to publish them
- inspiring teachers to use the corpus and new publications in their didactic offer
- significant improvement in the quality of learning Czech as a foreign language

In the world of learners corpora there is one missing…



Towards a learner corpus of Polish

- Polish as L2 (Ukraine, Belarus, Georgia, Chechnya, Turkmenistan, Kazakhstan, Azerbaijan, Armenia, Vietnam)

- Special focus on the language of Polish repatriates, including the children of emigrants (USA, Germany, England, France, Argentina, Brazil, Sweden, Australia, Ireland)

- Possible partners: Polonicum UW, Foreign Language Teaching Foundation Linguae Mundi, all universities and schools teaching Polish as L2