Językowe bazy danych: eksploracja, interpretacja, analiza

Kurs wprowadzający, edycja druga (ilustrowana nowymi przykładami).

Prowadzący: dr Katarzyna Klessa, UAM Poznań (klessa@amu.edu.pl)

W czasach, gdy pozyskiwanie różnego rodzaju danych, w tym także danych językowych, staje się coraz łatwiejsze dzięki nowym technologiom, rodzą się nowe wyzwania związane z komputerowym przetwarzaniem i analizą tych danych. Rosnąca dostępność urządzeń rejestrujących audio i wideo, wielkie przestrzenie dyskowe, architektura sieciowa i dostęp do Internetu prawie z każdego miejsca na ziemi dają z jednej strony niespotykane wcześniej możliwości, ale z drugiej sprawiają, że rosną także potrzeby w zakresie rozwoju metodologii i tworzenia nowych narzędzi badawczych.

Podczas warsztatu omówione zostaną wybrane metody „radzenia sobie" z dużymi i małymi zbiorami danych językowych uzyskanych (przede wszystkim, choć nie tylko) w wyniku nagrywania dźwięku mowy i jego opisywania. Spróbujemy choć w części odpowiedzieć na pytania:

- Jakie czynniki warto uwzględnić na etapie projektowania baz danych?
- W jaki sposób narzędzia komputerowe mogą wspomóc opracowanie danych językowych?
- Jak efektywnie opisywać nagrania mowy?
- Jak przyspieszyć swoją pracę z danymi?
- Jak przeszukiwać i analizować zebrane informacje?

Warsztat składa się z części teoretycznej i praktycznej. Obejmuje prezentację wybranych zagadnień z dziedziny projektowania językowych baz danych oraz przykładowych działających już komputerowych systemów anotacji dużych językowych baz danych. Ćwiczenia mają charakter podstawowy i związane są z wykorzystaniem komputerowych narzędzi na potrzeby opracowania i eksploracji kolekcji plików, odpowiedniego przygotowania danych i metadanych, elementami automatycznego przeszukiwania i przetwarzania anotacji nagrań oraz jej konwersji na formaty arkuszy kalkulacyjnych.

Jako materiał ćwiczeniowy wykorzystane zostaną dane z istniejących zasobów językowych, zawierających realizacje wypowiedzi swobodnych, quasi-spontanicznych, cechujących się zróżnicowanym nacechowaniem emocjonalnym, tempem mowy czy barwą głosu, np. *Paralingua*. (Klessa et al., 2013), elementy multimodalnej (audio-wideo) bazy danych polsko-niemieckiego projektu *Borderland* (http://borderland.amu.edu.pl/), przykłady ze stron: http://languagesindanger.eu/ oraz http://inne-jezyki.amu.edu.pl/ oraz wybrane opcje programu *Annotation Pro.*

Wybrane referencje:

Bavarian Archive for Speech Signals [oprogramowanie]: http://www.phonetik.uni-muenchen.de/forschung/Bas/
Bigi, B. SPPAS [oprogramowanie]: http://sldr.org/SLDR_data/Disk0/preview/000800/?lang=en
Czoska, A., Klessa, K., Karpiński, M. Polish Infant Directed vs. Adult Directed Speech: Selected Acoustic-Phonetic Differences (2015), *Proceedings of the 18th International Congress of Phonetic Scences*, Glasgow, UK.
Jung, D., Klessa, K., Duray, Z., Oszkó, B., Sipos, M., Szeverényi, S., Várnai, Z., Trilsbeek, P. & Váradi, T. Languagesindanger.eu – including multimedia language resources to disseminate knowledge and create educational material on less-resourced languages, *Proceedings of the 9th LREC*, Reykjavik, Iceland. ISBN 978-2-9517408-8-4.
Klessa, K. (2014). Dokumentacja języków, w: Nau N., Hornsby M., Karpiński M., Klessa K., Wójtowicz, R. (red.), *Księga wiedzy.* On-line: http://pl.languagesindanger.eu/book-of-knowledge/ oraz http://pl.languagesindanger.eu/book-of-knowledge/language-documentation/
Klessa, K. Anotation Pro [oprogramowanie]: http://annotationpro.org/.
Klessa, K., Demenko, G. (2009). Structure and Annotation of Polish LVCSR Speech Database. *Proceedings of Interspeech Conference*, September 6-10, 2009, Brighton, UK.
Klessa, K., Karpiński, M., Wagner, A. (2013). Annotation Pro – a new software tool for annotation of linguistic and paralinguistic features. In D. Hirst & B. Bigi (red.) *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, pp. 51-54.
Klessa, K., Karpinski, M., Czoska, A. (2105). Design, structure, and preliminary analyses of a speech corpus of Infant Directed Speech (IDS) & Adult Directed Speech (ADS). *Proceedings of 48th Annual Meeting of the Societas Linguistica Europaea, 2-5 September 2015, Leiden, the Netherlands.*
Klessa, K., Wagner, A., Oleśkowicz-Popiel, M., Karpiński, M. (2013). "Paralingua" – a new speech corpus for the studies of paralinguistic features In Chelo Vargas-Sierra (red.), *Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics. Procedia – Social and Behavioral Science*. Volume 95. pp. 48-58.
Klessa, K., Wicherkiewicz, T. (2015). Design and Implementation of an On-line Database for Endangered Languages: Multilingual Legacy of Poland, *Input a Word, Analyse the World: Selected Approaches to Corpus Linguistics*, F. A. Almeida, I. O. Barrera, E. Q. Toledo and M. Sánchez Cuervo (red.), Newcastle upon Tyne: Cambridge Scholars Publishing.
Stanford on-line: Introduction to databases https://www.coursera.org/course/db