

The analysis of speakers' errors in the Polish dialog corpus

The paper concerns the linguistic analysis of mistakes and errors found in the recordings of natural (spontaneous) spoken Polish dialogs. The basis for our research is the Polish corpus of dialogs collected within the LUNA – spoken Language UNDERstanding in multilingual communication systems – project, www.ist-luna.eu (IST 33549). The general assumptions of collecting all LUNA corpora are described in [Raymond et al. 2007]. The other source for our investigation are remarks of A. Dister, who considered the influence of grammar mistakes, slips of the tongue or repetitions in French for communication process [Dister 2008].

In the paper we briefly present the process of collecting and transcribing the Polish dialogs. The analyzed corpus contains spontaneous dialogs recorded at Warsaw City Transportation Information Center in spring 2007. The domain chosen for recording conversations is public transport in Warsaw. There have been 500 dialogs selected for the project. We will introduce general rules of transcription, which were agreed by all the project partners [Rodriguez et al. 2007], as well as rules especially invented to cover some phenomena significant for spoken Polish, e. g. syllabifying words, mostly proper names [Mykowiecka et al. 2007].

We distinguish several types of mistakes made by speakers of the recorded conversations:

- a. errors in pronunciation (e. g. confusion with Polish nasals; mispronunciation of proper names; incorrect pronunciation of foreign names);
- b. errors in inflection (e. g. confusion with the choice of appropriate inflectional suffix);
- c. errors in syntax (e. g. not realized, open clauses, mistakes in case government);
- d. repetitions and word truncations;
- e. neologisms.

All types of errors will be exemplified with the actual data taken from the corpus. The corpus has been annotated at the morpho-syntactic level, what also allow us to analyze the speakers' mistakes in Polish syntax and inflection.

The following table gives the primary statistics on the analyzed data concerning only the phenomena at the level of transcription:

	Number of occurrences
Words in the corpus	75572
Foreign words	288
Pronunciation errors	2830
Unintelligible fragments	1050
Babbled sections	4583
Word truncation	4630

Our analysis results in two sets of information. First, we gain the statistical overview of mistakes of the defined types and the most common mispronunciations. Comparing these data to the statistics of the entire dialog corpus we know exactly how high the rate of errors is in the spoken texts. Second, we can build a systematic description of frequent mistakes and get a basis for future automatic speech errors detection and recognition. We believe that the collection of errors made by actual speakers will be useful in training process of a dialog system for Polish.

References

- [Dister 2008] A. Dister. to appear. L'autocorrection immédiate en français parlé: le cas des déterminants. In *Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data, Lyon, France, 12–14 March 2008*.
- [Mykowiecka et al. 2007] A. Mykowiecka, K. Marasek, M. Marciniak, R. Gubrynowicz, and J. Rabięga-Wiśniewska. 2007. Annotation of Polish spoken dialogs in LUNA project. In *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of 3rd Language & Technology Conference. October 5-7, 2007, Poznań, Poland*.
- [Raymond et al. 2007] Ch. Raymond, G. Riccardi, K. J. Rodriguez, and J. Wisniewska. 2007. The Luna corpus: an annotation scheme for a multi-domain multi-lingual dialogue corpus. In R. Artstein and L. Vieu, editors, *Decalog 2007: Proceedings of the 11th Workshop on Semantics and Pragmatics of Dialogue, Trento, Italy, 30 May – 1 June 2007*, pages 185–186, Trento, Italy.
- [Rodriguez et al. 2007] K. J. Rodriguez, S. Dipper, M. Götze, M. Poesio, G. Riccardi, C. Raymond, and J. Rabięga-Wiśniewska. 2007. Standoff coordination for multi-tool annotation in a dialogue corpus. In *Proceedings of the Linguistic Annotation Workshop*, pages 148–155, Prague, Czech Republic. Association for Computational Linguistics.