Investigationes Linguisticae, vol. VIII, Poznań, June 2002

The Corpus of Polish Intonational Database (*PoInt*)[1]

Maciej Karpiński

Institute of Linguistics, Adam Mickiewicz University
Ul. Międzychodzka 5, 60-371 Poznań, POLAND

maciejk@amu.edu.pl

Polish Intonational Database Project (*PoInt*), carried out at the Institute of Linguistics (Adam Mickiewicz University), is going to be finished in September 2002. One of its results is a corpus of digitally recorded Polish speech. While the primary purpose of the recordings was to provide material for intonational analysis, they can be used in any kind of linguistic research.

The corpus consists of three main parts:

a) *Read texts*. twenty male and twenty female speakers read three texts: two pieces of contemporary prose and a poem for children. The speakers were asked to read "naturally", "in their normal way". They read the texts once "for themselves" before reading them to the microphone.
b) *Semi-spontaneous monologues*. The same speakers were asked to answer some questions related to a number of artifacts (a painting, a piece of music, a comic strip) and items of natural origin (seeds, pieces of wood, a shell).
c) *Dialogues*. Each of fifteen pairs of speakers was asked to accomplish three tasks: (a) to guess who is in a presented picture by asking a limited number of polar questions; (b) to perform a special version of the "map task"; (c) to discuss a controversial topic in order to reach and present a common conclusion.

PoInt corpus includes a range broadcast recordings. This part consists of three types of monologues: "general" news, weather reports/forecasts, sports news, and of a small number of public discussions. Unfortunately, the availability of this part of the corpus is limited due to legal issues.

All the speakers came from the region of Wielkopolska (Western Poland) or had spent there at least some recent years of their life. They were between 18 and 40, with at least college level education and no reported speech or language deficiencies.

The sound material was recorded digitally on optical discs (CD-R), using Tascam CD-R700 recorders, Shure Beta Series microphones, AKG condenser microphone, and a Spirit Folio F1 mixing console. The sessions, recorded as cda files, were converted into Windows wav format (16-bit/44.1kHz) and copied to a hard disk. Monologues were recorded as mono files, while dialogues were recorded in stereo mode, although the speakers and their microphones were not completely acoustically isolated. The signals were normalized to ca. 97% of the available dynamics range.

Over 30 hours of speech have been recorded, but only a part of that amount will be publicly available. The published part of the corpus will be issued as compressed mp3 files (192 kbps).

The corpus will be freely available for education and research but, due to financial and organizational issues, we are only able to produce a limited number of copies. In the case of commercial applications, please contact the author of this report.

[1] The project was supported by KBN (grant no. H01 D 011 18)

I would like to express my gratitude to a number of persons involved in this project. First of all, to my excellent colleagues – the core of the project team: Prof. Wiktor Jassem and Dr. Janusz Kleśta. Emilia Baranowska and Katarzyna Francuzik (doctoral students) helped us immensely and infected us with their enthusiasm. Wojciech Klessa, a gifted programmer, agreed to write software for managing the intonational database. I cannot omit Prof. Grażyna Demenko and Prof. Piotra Łobacz (Institute of Linguistics, AMU) who supported us on many occasions. We are especially grateful to Dr. Esther Grabe (Oxford University), who was always ready to share her great knowledge and experience in the domain of prosody. Finally, there was a large group of people who offered their voices for recordings and were ready to spend one or two hours in our small anechoic chamber.