TIPA. Travaux interdisciplinaires sur la parole et le langage

38 | 2022
Numéro spécial : Panorama des recherches au Laboratoire Parole et Langage

Corpus - OMProDat

Daniel Hirst

Electronic version
URL: https://journals.openedition.org/tipa/6037
DOI: 10.4000/tipa.6037
ISSN: 2264-7082

Publisher
Laboratoire Parole et Langage

Electronic reference
Daniel Hirst, "Corpus - OMProDat", *TIPA. Travaux interdisciplinaires sur la parole et le langage* [Online], 38 | 2022, Online since 27 January 2023, connection on 29 January 2023. URL: http://journals.openedition.org/tipa/6037 ; DOI: https://doi.org/10.4000/tipa.6037

Corpus - OMProDat

Coordinateur : Daniel Hirst

The open prosody database **OMProDat** contains recordings of 40 five sentence passages, originally taken from the European **SAM** project, each read by 5 male and 5 female speakers of each language. The database contains both primary data, the recordings, and secondary data in the form of different annotation files. Currently the database contains recordings and annotations for five languages: *Korean, English, French and Chinese* plus a smaller subset for several languages. All the data are freely available on **Ortolang** (https://www.ortolang.fr/fr/accueil/).

Origin of the database

The Eurom1 corpus, *Chan et al* (1995)

This corpus was created as a deliverable of the European Esprit project **2589 SAM (Speech Assessments and Methodology)** and its follow-up project **SAM-A**.
Eurom1 contained, in particular, a series of 40 continuous and thematically connected fivesentence passages, intended to represent a *clean* version of the various types of speech which speech technology might be expected to deal with.
The passages were based on identical themes for the different languages, freely translated and adapted from the original English texts for the different languages. Two sample passages from the Eurom1-EN database are:

[**T02**] *I have a problem with my water softener.*
The water-level is too high and the overflow keeps dripping.
Could you arrange to send an engineer on Tuesday morning please?
It's the only day I can manage this week.
I'd be grateful if you could confirm the arrangement in writing.

[**T33**] *Hello, is that the telephone-order service?*
There seems to have been some mistake.
I ordered a teddy bear from the catalogue and was billed for an electric lawnmower.
And I don't even have a garden.

Would you put me through to the complaints department, please?

The passages were originally recorded in the 1980's for eleven European languages:
Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish.

The Babel corpus

A compatible speech database for East European languages was later recorded during the **Copernicus** project 1304, **Babel**, with similar recordings for *Bulgarian, Estonian, Hungarian, Polish, and Romanian* (*Roach et al* 1996).

The MULTEXT Prosodic Database

The continuous passages from the Eurom1 corpus in five languages: *English, French, Italian, German, and Spanish*, were re-used during the Esprit project **Multext**.
The recordings were provided with manually created annotation files for word labels and with automatically stylised f0 patterns using the Momel algorithm *Hirst & Espesser* (1993), *Hirst* (2007).
The database was published as the **MULTEXT Prosodic Database** *Campione & Véronis (1998)*.

A compatible version of the database for East European languages was produced as **Multext-East** *Erjavec 2004*.

Availability

The original Eurom1 recordings were protected by copyright assigned to the different laboratories that produced the recordings. For details see
http://www.phon.ucl.ac.uk/shop/eurom1.php. The database contained on 30 CDs is available for sale from the same address for £100.

The Babel corpora are available from ELRA (http://catalog.elra.info.product_info.php) at 600€ per language for researchers. ELRA members get a 50\% discount and ELRA membership costs 750€ for non-profit-making organisations.

The Multext Database is available from the same address for 100€ for academic researchers. The Multext-East recordings are freely available from http://nl.ijs.si/ME/.

Creating the *OMProDat* database

In order to provide a more solid basis for the analysis of prosodic metrics, we decided to build an open multilingual prosodic database **OMProDat**, to be archived and distributed by **Ortolang** under an open database license.

The aim of this database is to collect, archive and distribute recordings and annotations of directly comparable data from a representative sample of different languages representing different prosodic typological characteristics.

As mentioned above, the passages of the different versions of the original Eurom1 corpus were typically read by only two or three speakers each.
This makes the corpus of limited use for the study of speaker variability.

It was consequently decided to make new recordings of the corpus, with all 40 passages read by 10 speakers each.

Korean

The first language recorded under these conditions was **Korean** *Kim et al* (2008). The original English version of the Eurom1 text was translated and adapted for Korean.
The texts in Korean alphabet were Romanized and also transcribed in *SAMPA* and *IPA*. 10 Seoul speakers (5 male and 5 female) took part in the recording session, all were Korean native speakers in their twenties, either undergraduate or graduate students of Seoul National University. Each speaker read all 40 passages.

For prosodic annotation, the *Momel* algorithm was used (*Hirst* 2007) and the pitch targets obtained were manually corrected. The prosodic events were annotated in two ways: first, with the automatic annotation algorithm, **INTSINT** (*Hirst* 2007) and second, with manual labelling of prosodic units using just two tone labels (**H** and **L**).

English and French : AixOx

New recordings were made for English and French read by native speakers, as well as for English read by native speakers of French and for French read by native speakers of English (*Herment et al* 2012). The speakers were all 20-30 years old. All speakers were from monolingual families. The English speakers were recorded in Oxford and spoke Southern British English; the French speakers were

recorded in Aix-en-Provence and spoke either a Southern or a standard variety of French, or something between the two.

The originality of this corpus is that it provides recordings for both natives and non-native speakers, so as to allow comparative studies on L1 and L2 productions.

Three groups of learners were recorded for each language, one group of native speakers on the one hand and two groups of non-native speakers, corresponding to the levels of the **Common European Framework of Reference for Languages*, (***CEFR*****): classified respectively as *independent users* (level B1/B2) and *proficient users* (level C1/C2).

The recordings are accompanied by *TextGrid* annotation files obtained semi-automatically from the sound and the orthographic transcription using the **SPPAS** alignment software (*Bigi & Hirst* 2012) using manual correction when necessary

Prosodic annotation was also obtained using the Momel and INTSINT automatic annotation algorithms (*Hirst* 2007).

Chinese

More recently, recordings for Standard Mandarin Chinese (*Ding & Hirst* 2012) were made. The speakers were 10 Chinese native speakers: 5 female and 5 male.
Their ages ranged from 21 to 31 years old, and they were all postgraduate students and speakers of standard Chinese.
Each speaker read all 40 passages. The annotation of the recordings using SPPAS and Momel/INTSINT is currently in progress.

The OpenProDat mutlilingual sample

A more limited set of data from a larger number of languages were also collected and distributed.

The two paragraphs from the English Eurom1 corpus given as examples above were translated and recorded. This shared corpus is hosted by Ortolang (https://hdl.handle.net/11403/openprodat/v2) under the name **OpenProDat** as a part of the more general **OMProDat** database described here.

These texts were transcribed in: *Dutch, French, German, Italian, Arabic, Spanish, Finnish, Hungarian, Japanese and Thai*. Each participant read both paragraphs, first in their mother tongue and then in each language that they felt able to read.

Some files were manually transcribed and annotated with **SPPAS** (*Bigi & Hirst* 2012). These annotations are also freely available in the **Ortolang** Platform**.**.

References

Bigi, B. and Hirst, D. J. 2012 SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In *Proceedings of the 6th International Conference on Speech Prosody*., May.

Bigi, B. & Hirst, D.J. (eds) 2013 *Proceedings of TRASP: Tools and Resources for the Analysis of Speech Prosody*, Aix-en-Provence, 30 August 2013.

Boersma, P. and Weenink, D. 1992 [2013] Praat, a system for doing phonetics by computer. http://www.praat.org [version 5.3.41, February 2013].

Campione, E. and Véronis, J. 1998. A multilingual prosodic database. In *Proceedings of ICSL'98,* Sidney, Australia.

Chan, D. Fourcin, A.; Gibbon, D.; Granstrom, B.; Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L.; Lindberg, B.; Moreno, A.; Mouropoulos, J.; Senia, F.; Trancoso, I.; Veld, C. and Zeiliger, J. Eurom - a spoken language resource for the EU. 1995. In *Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology*., 1, 867–870, Madrid., 18-21 September 1995.

Ding, D. and Hirst, D.J. 2012 A preliminary investigation of third-tone sandhi in Standard Chinese with a prosodic corpus. *8th International Symposium on Chinese Spoken Language Processing*, Hong Kong.

Erjavec, T. 2004 MULTEXT-East Version 3: Multilingual morphosyntactic specifications, lexicons and corpora. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal: 1535-1538. [available at http://nl.ijs.si/ME/]

Herment, S., Tortel, A., Bigi, B. Hirst, D., and Loukina, A. 2012 AixOx: A multi-layered learners corpus: automatic annotation. *4th International Conference on Corpus Linguistics*., Jaèn, Spain, (in Díaz Pérez, J. and Díaz Negrillo, A. (eds.) Specialisation and variation in language corpora, Peter Lang.) .

Hirst, D.J. 2007 A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. In *Proceedings of the XVIth International Conference of Phonetic Sciences*: 1233–1236, Saarbrucken.

Hirst, D.J.; Bigi,B; Cho, H.; Ding,H.; Herment, S.; Wang, T. 2013 Building OMProDat, an open multilingual prosodic database. in Bigi & Hirst (eds.) 2013.

Hirst, D.J. and Espesser, R. 1993 Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15: 75–85.

Kim, S-H.; Hirst, D.J.; Cho, H.-S.; Lee, H.-Y. and Chung, M.-H. 2008 Korean Multext: A Korean prosody corpus. In *Proceedings of the 4th International Conference on Speech Prosody*., Campinas, Brazil.

Roach, P.; Arnfield, S. and Hallum, E. 1996 BABEL: A multi-language speech database. In *Proceedings of SST-96: Speech and Science Technology Conference*., Adelaide: 351–4.

Véronis, J.; Hirst, D.J. and Ide, N. 1994 NL and speech in the MULTEXT project. In *Proceedings of AAAI Workshop on Integration of Natural Language and Speech*, Seattle, USA: 72–78.