Human Language Technologies – The Baltic Perspective
I. Skadiņa and R. Rozis (Eds.)

doi:10.3233/978-1-61499-701-6-46

Designing an Annotated Longitudinal Latvian Children's Speech Corpus

Ilze AUZIŅA [a,1], Kristīne LEVĀNE-PETROVA [a], Guna RĀBANTE-BUŠA [a], Roberts DARĢIS [a] and Antonio FÁBREGAS [b]
[a] *Institute of Mathematics and Computer Science, University of Latvia*
[b] *The Arctic University of Norway*

Abstract. The paper provides an overview of the ongoing development of the Annotated Longitudinal Latvian Children's Speech Corpus. The authors outline the design of this corpus and the layers of annotation (both orthographic and part-off-speech tagging) with which the speech signal is enriched.

Keywords. Speech corpus, children's speech, orthographic transcription, part-of-speech tagging

1. Introduction

The main purpose of this paper is to present a new Annotated Longitudinal Latvian Children's Speech Corpus. The corpus is a part of the project "Latvian Language in Monolingual and Bilingual Acquisition: tools, theories and applications"[2] which is running from 2015 March until April 2017. Data storage is already completed, but orthographic transcription still continues. The last stage of corpus development will also include part-of-speech tagging. Both orthographic annotation and part-of-speech tagging will be transformed according to the CHILDES (Child Language Data Exchange System) and MOR systems. This is the first such kind of language resource for Latvian.

Creation of the Annotated Longitudinal Children's Speech Corpus that represents the 17 to 43 month-old monolingual and bilingual (Latvian and Russian speaking) children speech is a very challenging task; therefore, the authors have structured the abstract so that it can be used as guidelines for creation of different child language corpora, an aspect that in related works on children's corpora has not been sufficiently addressed. The previous experience has been used also in our efforts, for example, [1] on Spanish/German, [2] on Spanish/German, [3] on Portuguese, etc.

2. Constitution of the Corpus

The aim of the project is to create both longitudinal sub-corpora of monolingual Latvian-speaking children, and longitudinal sub-corpora of simultaneous Latvian-Russian bilingual children. The data recording is completed, but the development of the

[1] Corresponding Author: Ilze Auziņa; e-mail: ilze.auzina@lumii.lv.

Annotated Longitudinal Children's Speech Corpus will be finalized until the end of 2016. Afterwards it will be possible to present the complete statistics; however, the statistics at the moment of writing the paper are given in Table 1. The corpus contains speech annotation in two levels: orthographic and morphological annotation.

Table 1. The statistic of the Annotated Longitudinal Children's Speech Corpus

Child	Age	Number of files	Total duration
Mo_B1	1;6 – 2;8	60	~24.00 h
Mo_G4	1;5 – 2;7	48	~31.00 h
Mo_G2	2;3 – 3;6	66	~37.00 h
Bi_B3	2;4 – 3;7	46	~26.00 h (Latvian) ~10.00 h (Russian)

2.1. Corpus size

The total size of the corpus is expected to reach 192 hours of child-directed and child-adult speech recordings that would be orthographically annotated. Monolingual children are recorded for 30 minutes per session, while bilingual child is recorded for 30 minutes per session in each of his languages (Russian and Latvian). Recordings have been carried out for 16 month. During this time, four recording sessions per month have been conducted at regular intervals.

Unfortunately, for various objective reasons the records have not been conducted with so much regularity, as it was originally planned. The largest break in the recording of one respondent is nine weeks. Currently amount of recordings is approximately 128 hours (more than 100 hours of speech).

2.2. Physical characteristics of speakers

When the speech recordings began, the youngest participant was 17 months old (1; 5), but the oldest was 27 (2; 3). The speech corpus is representative of speakers of both genders in equal proportions. The speech of youngest monolingual girl (Mo_G4) is being recorded from 17 months to 31 months of age. The speech of oldest monolingual girl (Mo_G2) is being recorded from 27 months to 42 months of age. The speech of monolingual boy (Mo_B1) is being recorded from 18 months to 32 months of age. A bilingual boy's speech data were recorded from 28 to 42 months of age. See Table 1.

2.3. Data collection and metadata

Recorded data correspond to child-adult interaction in a naturalistic setting: children were recorded at their home or other familiar environments interacting with their family (most often their mother). Obtained speech samples were uploaded to a special website for further data processing by the researchers. By uploading the file some metadata are added to the record, for instance, short description of the environment and activity, also date when interaction was recorded. It is also possible to upload pictures of the venue, for instance, room, where activities happened. Video recordings are not performed.

2.4. Recording devices and data formats

The data are collected with the different speech recording devices. The built-in microphone of the devices is used.

The corpus consists of speech audio files, multiple meta-data XML documents. The audio files are of one channel, with 16 bits allocated per sample. The frequency varies depending on the source audio data quality (with a minimum of 16 KHz). Since various recording devices (both various models of mobile phones and dictaphones) are used, the recordings of different audio formats and parameters are obtained. All audio files are converted to WAV format while retaining the original frequency.

3. Orthographic Annotation

Following earlier research on orthographically annotated speech corpora creation [4], [5], [6] and using previous experience in the development Latvian Speech Recognition corpus [7] the authors have created a set of rules for the orthographic annotation.

The rules specify how to annotate pauses, non-speech fragments (e.g., filled stops, babbling, etc.), abrupt words, unclear speech, words spoken in a different language (e.g., bilingual child in Latvian uses Russian words, etc.), physiological noise and processes (e.g., snuffling, smacking, coughing, crying, breathing, etc.), background noise and other types of information characterizing a speech fragment within an utterance. Unintelligible words with an unclear phonetic shape are transcribed as [xxx]. The phonological form of an incomplete or unintelligible phonological string is written out with an ampersand and the *correct* form is given the square brackets, as in &mā [māja] 'home'.

Several of the acoustic event categories are listed in Table 2 and an example of orthographically annotated utterances and their translations is given in Figure 1.

To facilitate recognition of proper nouns only the proper names are written with capital letters. Some punctuation marks are use in orthographic transcription to mark a discourse, for example, question mark that indicates the end of interrogative utterance. Commas are usually used according to the Latvian grammar rules.

All adults and children utterances are intended to transcribe orthographically.

The orthographic transcription is easily transformable to CHAT (Codes for the Human Analysis of Transcripts) format, which is the standard transcription system utilized in Child Language Data Exchange System (CHILDES; [8]). Audio recordings and transcriptions are linked and synchronized.

Table 2. Categories of acoustic events in the orthographic annotation

	Type	Label
Pauses	Short pause	(.)
	Long pause	(*time in sec*), e.g., (0.5)
	Filled pause	e.g., (ā), (ē)
Physiological noises and processes	Laugh	@; <@> *text* </@>
	Babbling	(*time in sec*)
	Crying	<r/>; <r> *text* </r>
	Inhalation	(.h)
	Exhalation	(h.)
Word explanation, correct form		*daguns* [deguns] 'nose'

Mo_G4: pamodās no **tokšņa [trokšņa]** 'woke up from the noise'	Segmentation of audio into sentence-like chunks.
Mo_G4: **pēkšņi** pamodās 'suddenly woke up'	In cases of deviations from the norm, the correct form is given in square brackets.
Mo_G4: tie **i [ir] kalavīli [karavīri]** 'they are soldiers'	
Mo_B1: **&mā** [māja] 'home'	An incomplete or unintelligible phonological string is written out with an ampersand
Mo_G4_mother: tur **Brita** man liekas aizgāja gulēt 'I think there Brita went to bed'	Capital letters are used in proper names and acronyms only.

Figure 1. The examples of orthographically annotated utterances.

4. Morphological Annotation

In this section of the paper the last corpus development stage – morphological annotation – will be described.

Latvian part of the orthographically transcribed corpus is morphologically tagged. The previous morphological feature annotation standard for Latvian text corpora was used [9] and already existing morphological taggers [10] were adapted. The initial experiments with adapted tagger showed that it works sufficiently well. It was decided to use almost full morphological tag set excluding just some features that is hardly recognizable, for instance, some features for verbs. Some examples of this annotation layer are listed below:

kājas 'legs' – ncfpa4:
n – noun,
c – common noun,
f – feminine,
p – plural,
a – accusative,
4 – 4th declension;
tur 'there' – r – adverb.

This annotation is transformable to MOR – a program that provides a method for automatic tagging of corpora in the CHAT format, but for that reasons it is necessary to build MOR grammar for particular language – Latvian in this case.

5. Conclusions

In this paper the authors have presented the overall design of the Annotated Longitudinal Latvian Children's Speech Corpus. Work on the corpus is still ongoing: data is collected, but the annotation of the data continues. For experimental purposes a part of the orthographically annotated speech corpus (only monolingual children's utterances) will be provided also in the broad transcription (or phonemic transcription): transcription that relates the allophones produced by the speakers to the phonemes of Latvian.

Developing the Latvian Children's Speech corpus and making it freely available through specialized international databases will not only strengthen the empirical foundation of child language studies in Latvia and create a sustainable boost of interest in the national linguistic community; it will also facilitate scientific exchange on the European and global level, pave the way for new international collaboration and considerably increase the visibility of the Latvian language on the international arena of language studies.

Acknowledgements

The research project "Latvian Language in Monolingual and Bilingual Acquisition: tools, theories and applications" leading to these results has received funding from the Norwegian Financial Mechanism 2009-2014 under Project Contract No NFI/R/2014/053.

References

[1] M.S. Ulloa, C. Lleó, I.G. Sánchez, Corpora of spoken Spanish by simultaneous and successive German-Spanish bilingual and Spanish monolingul children, *Multilingual Corpora and Multilingual Corpus Analysis* (2012), 97-106.

[2] C. Lleó, Monolingual and bilingual phonoprosodic corpora of child German and child Spanish, *Multilingual Corpora and Multilingual Corpus Analysis* (2012), 107–122.

[3] A.L. Santos, M. Génereux, A. Cardoso, C. Agostinho, S. Abalada, A corpus of European Portuguese child and child-directed speech, *Proceedings of the 9th Conference on Language Resources and Evaluation – LREC 2014* (2014).

[4] W. Goedertier, S. Goddijn, J. Martens, Orthographic Transcription of the Spoken Dutch Corpus, *Proceedings of LREC-2000* (2000), 909–914.

[5] J.B. Johannessen, K. Hagen, J.J. Priestley, L. Nygaard, An Advanced Speech Corpus for Norwegian, *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*(2007), 29–36.

[6] N. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J. Martens, M. Moortgat, H. Baayen, Experiences from the Spoken Dutch Corpus Project, *Proceedings of LREC 2002* (2002), 340–347.

[7] M. Pinnis, I. Auziņa, K. Goba, Designing the Latvian Speech Recognition Corpus. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14)* (2014).

[8] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates (2000).

[9] K. Levāne, A. Spektors, Morphemic Analysis and Morphological Tagging of Latvian Corpus, *Proceedings of the Second International Conference on Language Resources and Evaluation. Athens, Greece* (2000), vol. 2, 1095–1098.

[10] P. Paikens, L. Rituma, L. Pretkalniņa, Morphological analysis with limited resources: Latvian example, *Proceedings of NODALIDA 2013* (2013), 267–278.