Presented at the *3rd Summit on Gender Equality in Computing (GEC 2021)*, July 2, 2021, University of West Attica, https://acmw-gr.acm.org/en/gec21

The GrHarvard Corpus

A Greek sentence corpus for speech technology research and applications

Anna Sfakianaki
Computer Science Department
University of Crete
Heraklion Crete Greece
asfakianaki@csd.uoc.gr

George Kafentzis
Computer Science Department
University of Crete
Heraklion Crete Greece
kafentz@csd.uoc.gr

Yannis Stylianou
Computer Science Department
University of Crete
Heraklion Crete Greece
yannis@csd.uoc.gr

ABSTRACT

The present work concerns the construction of a phonemically-balanced sentence corpus in Modern Greek accompanied by a database of recordings, freely available for speech technology research and applications. Our motivation for the corpus construction was the lack of carefully designed sentence corpora in Greek, which can be used in speech intelligibility experiments. The GrHarvard Corpus has been based on the English Harvard/IEEE sentences and consists of 720 sentences containing five to nine carefully selected words of one to three syllables. A first version of the corpus has been published and is freely available in orthographic and phonetic transcription. Current work involves corpus validation and balancing.

KEYWORDS

Harvard sentences, Modern Greek corpus, speech intelligibility, speech technology

1 Introduction

Research in speech science and technology requires linguistic corpora as well as speech datasets that fulfill specific criteria in order to test the efficacy of algorithms and create successful applications. For example, enhancing intelligibility in communication devices is one of the main areas where the aforementioned corpora can be utilized. To the best of our knowledge, GrHarvard [1] is the first carefully designed sentence corpus in Greek, as previous published work concerns only word lists (e.g. [2]). The present corpus follows the format of the Harvard/IEEE sentences [3] which have been used extensively for intelligibility tests in the English language.

2 The GrHarvard Corpus

Recently, the Harvard/IEEE sentences have served as a basis for the construction of sentence corpora in other languages as well (i.e. Spanish [4] and French [5]). Although partly inspired by the Harvard/IEEE sentences, a direct translation into Greek was not possible in many cases due to grammatical differences between the two languages; hence the majority of the GrHarvard sentences are original. One- to three-syllable words were manually selected from the lexical database Greeklex 2 [6], and combined so that each sentence includes five keywords, while the total number of words in the sentence varies from five to nine. The sentences are meaningful, semi-predictable and resemble everyday conversation.

3 Corpus phoneme frequency and balancing

The corpus comprises a total of 3,600 keywords containing 20,230 phonemes and allophones. The phoneme frequency distribution of the present corpus is consistent with phoneme distribution reported for other Greek written and spoken corpora (e.g. [7]). An algorithm, similar to the one described in [4], is implemented for phonemic balancing between subsets of the corpus based on Euclidean distance.

4 Current work

Current work involves sentence validation and further balancing. Regarding validation, the sentences will be judged by native speakers of Greek as regards naturalness and difficulty. Sentences judged as unnatural will be replaced. In addition, difficulty regarding word familiarity and sentence comprehensibility will be rated, so as to include this extra factor in the balancing scheme. Listening experiments in noise will finally take place to validate the final sentence grouping.

ACKNOWLEDGMENTS

The first version of the GrHarvard Corpus was developed within ENRICH (Enriched Communication Across the Lifespan), http://www.enrich-etn.eu/, (H2020, MSCA GA 675324).

REFERENCES

[1] Anna Sfakianaki. 2019. Designing a Modern Greek sentence corpus for audiological and speech technology research. To appear in *Proc. 14th International Conference on Greek Linguistics (ICGL14)*. Available at https://elocus.lib.uoc.gr/dlib/0/b/f/metadata-dlib-user1612948003-17177.tkl?lang=en

[2] Vassiliki Iliadou, Marios Fourakis, Angelos Vakalos, John W. Hawks and George Kaprinis. 2006. Bi-syllabic, Modern Greek word lists for use in word recognition tests. *Int J Audiology* 45, 74–82. DOI: https://doi.org/10.1080/14992020500376529

[3] E. H. Rothauser et al. 1969. IEEE Recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17, 3, 225–246. DOI: 10.1109/IEEESTD.1969.7405210

[4] Vincent Aubanel, Maria Luisa García Lecumberri and Cooke Martin. 2014. The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology. *Int J Audiology*, 53, 633–638. DOI: https://doi.org/10.3109/14992027.2014.907507

[5] Vincent Aubanel, C. Bayard, Antje Strauß and J.-L. Schwartz. 2020. The Fharvard corpus: A phonemically-balanced French sentence resource for audiology and intelligibility research. *Speech Communication* 124, 68-74. DOI: https://doi.org/10.1016/j.specom.2020.07.004

[6] Antonios Kyparissiadis, Walter J.B. van Heuven, Nicola J. Pitchford and Timothy Ledgeway. 2017. GreekLex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information. *PLoS ONE* 12, 2, e0172493. DOI: 10.1371/journal.pone.0172493

[7] Athanassios Protopapas, Marina Tzakosta, Aimilios Chalamandaris and Pirros Tsiakoulis. 2012. IPLR: an online resource for Greek word-level and sublexical information. *Language Resources & Evaluation*, 46, 449–459. DOI: https://doi.org/10.1007/s10579-010-9130-z