

Project 1: Corpora for discovering language-specific and language-general patterns of FPD

Coordinators: Baayen (NL), Mattys (UK), Local (UK)

Team: Cenoz, Cooke, Cutugno, d'Imperio, van Dommelen, Ernestus, Ford, Frauenfelder, Giurgiu, Gussenhoven, Hawkins, Koreman, Lecumberri, Meunier, Moore, Nguyen, Ogden, Palková, Post, Svendsen, Volín, Wells

The S2S partners already use many corpora, ranging from spontaneous conversations between two or more people to tightly-controlled read phrases or isolated words, and from large, multi-speaker collections to just a few tokens of a few structural types from one or two speakers. We will use these corpora when appropriate. Project 1 will take S2S further by serving three related functions: infrastructure service provision, corpus-building in a range of languages, and hypothesis-testing.

Infrastructure. Project 1 will provide advice and resources for making corpora for the other S2S projects e.g. for a language for which there is no appropriate corpus (e.g. Basque), or for studying a particular type of FPD, such as a particular morphological distinction. The infrastructure service will be offered to all partners from month 1. It will be carried out in collaboration with the project concerned and needs no detailed description. Collaboration of experts in phonetics, psychology, computation, corpus design, and the main S2S languages should assure wise decisions.

Corpus-building. The aim is to discover salient FPD in as many languages as practicable (depending on who the Fellows are). S2S will encourage Fellows to build corpora for focused research in little-studied languages. Examples:

- Basque and Spanish for speech varieties and L2 learning.
- Romanian morphological and grammatical properties offer valuable comparisons with Italian, a much more studied Romance language; spread of palatalisation (involved in the singular/plural system) and connected speech processes at word boundaries, and to enhance Romanian ASR and TTS.
- Norwegian: the interaction of word accent (1 and 2) and global intonation contour, relevant to technology and L2 applications.
- Prague already has a large corpus of different varieties of Czech speech, and will begin to analyse for FPD guided by York and Sheffield and the needs of the other S2S projects Prague scientists undertake.

Hypothesis-testing by corpus comparisons.

(1) **Comparisons of speech style** using standard phonetic and automatic analyses will establish which types of FPD can be found in different types of speech. Many partners believe that FPD's most important perceptual role is in casual speech. However, it is extremely challenging to collect enough tokens of casual speech in the same linguistic context for statistical analysis or application of automatic machine methods such as HMMs. S2S partners have approached this challenge in different ways. Nijmegen uses large corpora (often of spontaneous speech) and tries to control for structural variation using appropriate statistics. But the choice of statistical controls involves many a priori assumptions, and there is debate about what insights are lost by lumping the same words together despite different contexts. At the other extreme, York and Sheffield phoneticians use close analysis of individual utterances in spontaneous speech, generalising across tokens grouped together by their interactive function in the conversation. This produces small sample sizes and problems or perceived problems of generality. In an intermediate approach, Cambridge phoneticians record specially-constructed read texts that have been practiced until they sound natural. This controls over the linguistic structures, but writing the texts is difficult, there may be limited numbers of repetitions, and, although the speech has many characteristics of casual speech, it is not spontaneous, and it is rarely in dialogue. Another intermediate solution, advocated by Bristol, is the

Map Task. Two people describe routes to each other using maps showing slightly different landmarks, while neither can see the other's map. This produces spontaneous speech, but although words are repeated, they are mainly content words reflecting where the maps differ, whereas much FPD reflects grammatical and discourse structure. Moreover, the syntactic/prosodic structures of the repetitions are uncontrolled, which means that there is limited scope to model the linguistic structures that mediate observed phonetic patterns.

(2) **Cross-language comparisons** will help to identify language-general vs. language-specific FPD. FPD discovery will be informed by (a) observations of expert phoneticians, (b) patterns of FPD in more-studied languages (Dutch, English, French), (c) needs and recommendations from other S2S projects. Examples include:

- morphological distinctions e.g. Dutch singular vs. plural nouns; English past tense forms compared with non-past-tense 'homonyms' (*massed-mast*) and similar patterns (e.g. *row-rowed-road*);
- segmental distinctions e.g. syllable-level properties associated with voicing in segments of a coda; long-domain reflections of particular segments, such as English /r/ and French vowel harmony;
- interactions between prosody and segments e.g. f0 alignment (timing of changes in the f0 contour) relative to the segmental structure of syllables, morphological structure of words, and higher-level prosodic boundaries; rhythmic (spectro-temporal) differences between productive vs non-productive morphemes (*mistimes-mistakes*).

When conversational and other forms of spontaneous speech are available, two topics will be studied in depth using CA, which offers a way to achieve comparability in spontaneous talk:

(i) **turn-taking**. The phonetic details of pausing in various languages will be examined, looking at the segmental and prosodic details of speech. Speakers sometimes use phonetic detail to project the next consonant type, e.g. by producing a velar closure when the word searched for starts with a velar; and in turn, recipients (listeners) can orient to this by providing a candidate word for the word-search which starts with the consonant projected in the prior turn.

(ii) **speech reduction in context**. This study will consider the detail, distribution and function of segmental reduction in conversational speech, by examining reduction in particular conversational structures. It unifies Nijmegen's reduction analyses with York's interactional and sequential analysis, and with lexis and intonation. Guiding questions include: What are the ranges of variability for any given lexical item or combination of lexical items? Does the range interact with conversational function? How do segmental and intonational features relate to one another? How do speakers and listeners orient to such features of talk as it progresses in time?

Indicative resources. Total of 3 fellows (2 ESRs and 1 ER to address fragmentation). For CA analyses, one ESR could be shared between Sheffield and York (close commuting distance) or two could start together. For less-studied languages, one ESR will be based at Trondheim and one ER in Prague. Visits: other countries for languages as appropriate (Basque Country, Cluj, Nijmegen), and York for CA and FPD. Backgrounds: phonetics with strong quantitative backgrounds, or possibly computation with a specialisation in language or speech. Morphology might be done by a psycholinguist especially if ER. Multilingualism an advantage.

Links. Underpins all other Themes, but especially other Theme I projects.

Will inform and be informed by the automated FPD discovery processes in Project 2.

FPD discoveries will be tested for perceptual salience when possible in Projects 3 & 4, & 5-8.