**Word-joins in real-life speech: a large corpus-based study**

John Coleman, Greg Kochanski, Ros Temple and Jiahong Yuan

## 1. Introduction

Spoken language, particularly ordinary conversation, has a special importance in the study of language (e.g. Saussure 1915 ch. 2, Abercrombie 1963). It predominates in all cultures and times, and is acquired naturally. The collection of large speech corpora such as the spoken part of the British National Corpus (Spoken BNC) and advances in speech technology now provide, for the first time, unprecedented opportunities to work with large quantities of real-life speech. To do this on a large scale, we will apply new methodologies drawn from speech technology research to answer linguistic questions which it has not been possible to address until now. The size of the Spoken BNC means that we can begin to study the pronunciation of combinations of words, with a sufficient number of tokens to obtain statistically valid results.

The differences between read and conversational speech (e.g. Fosler-Lussier 1999), between more or less formal situations (e.g. Coupland 1980), and between laboratory speech and natural speech, (e.g. Shockey 2003, Johnson 2004) have been at the heart of the variationist approach (e.g. Labov 1972, especially chapters 3 and 8). Such differences apply not only to pronunciation within words but also to the ways in which words are joined up. Corpus-based studies of French *liaison* (e.g. Ågren 1973, Durand and Lyche 2008, and references therein) have shown that the frequency and distribution of *liaison* patterns vary between formal and less formal speech. This suggests that English word-join phenomena, too, should be studied in everyday speech, not just in the laboratory.

Language is characterized by a hierarchy of improbability: larger corpora are required to find examples of more complex structures, or combinations of structural, sociolinguistic or contextual variables. For instance, most phonemes occur within a minute of speech, but finding pronunciations of common words requires about an hour of speech. However, to understand the phonological rules that glue words together, one needs not just single words but pairs. Finding an appropriate sample of pairs of words requires a corpus containing hundreds of hours of speech.

The Spoken BNC uniquely combines informality with size. It is the largest and most comprehensive snapshot of "language in the wild" ever collected: over 1200 hours. Volunteers carried a tape recorder around with them for several days, yielding a cross-section of everyday speech, including lunchtime conversations, discussions of boyfriends, and dog-directed speech. These conversations are unstructured, wide-ranging, often involve multiple people, and form about half of the Spoken BNC. The other half is more formal, but mostly unscripted speech, e.g. interviews and religious services. We will use both halves of the Spoken BNC. The recorded audio is transcribed at the word level, with rich linguistic annotation and metadata.

Because of its size and nature, the Spoken BNC will be an exceptionally valuable resource for studies of English phonetics and phonology. Sociolinguists and others have collected unscripted corpora (typically sociolinguistic interviews) but all are substantially smaller than the Spoken BNC[1]. The largest comparable transcribed and aligned dataset is the Switchboard corpus (Godfrey et al. 1992) which amounts to just 300 hours of audio. Among the largest British sociolinguistic corpora, the York Corpus (Tagliamonte 1998) contains c. 100 hours of transcribed sociolinguistic interviews, and the Newcastle Electronic Corpus of Tyneside English[2] contains c. 10 hours of transcribed audio.

---

1. Although not comprehensive, an extensive list of spoken corpora can be found at http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm.
2. http://www.ncl.ac.uk/necte/index.htm

## 2. Research Questions

We will investigate how words are joined together in fluent, colloquial English speech. We focus on consonantal assimilation and t/d deletion. Our study will be loosely parallel to Dilley and Pitt (2007) who studied American English to answer somewhat different research questions.

### 2.1. Consonant place assimilation

According to many handbooks and textbooks on English phonology (e.g. Kriedler 1989:257, Harris 1994: 72, Roca and Johnson 1999: 34-7, McMahon 2002: 45, Cruttenden 2008: 301), word-final alveolar consonants (i.e. /t/, /d/, /n/, /s/ and /z/) - and *only* alveolar consonants - change their place of articulation to match the consonant with which the next word begins. Dilley and Pitt (2007) studied the incidence only of theoretically *allowed* alveolar assimilations, like:

"that case" → "tha[k] case"

"bad case" → "ba[g] case"

"this shop" → "thi[ʃ] shop" (cf. "fish shop")

"ran quickly" → "ra[ŋ] quickly"

"his shop" → "hi[ʃ] shop

Such assimilation also occurs word-internally, e.g.:

"incompetent" → "i[ŋ]competent"

"input" → "i[m]put"

"infer" → "i[ʃ]fer" (somewhat like "imfer")

"ungodly" → "u[ŋ]godly"

"unbecoming" → "u[m]becoming"

"envy" → "e[ʃ]vy"

In English, assimilation of labial or velar consonants should *not* happen, according to the textbook rule. Thus, pronunciations such as "ki[m]pin" for "kingpin" or "alar[ŋ] clock" for "alarm clock" would be counterexamples to the general rule. If more counterexamples are found than can be explained as speech errors, the rule would need to be questioned or modified. Therefore, in contrast to Dilley and Pitt, we will also search for instances of assimilation in the theoretically *forbidden* cases. We must examine a large corpus, because violations of the rule will be less common than the allowed cases. Dilley and Pitt found that about 9% of the word-final alveolars assimilated to the place of articulation of the following consonant; we aim to uncover possible violations of this rule (i.e. *non*-alveolar assimilations) at the 1% - 3% level.

The allegedly forbidden assimilations are found in other languages (e.g. German: Zimmerer et al. 2009; Diola-Fogny, Korean: Jun 1995 ch. 2), so they are certainly physically possible. If they are not found in English, it can only be because some language-specific constraint(s) forbid them. Theory notwithstanding, there is some evidence that such violations occur. Barry (1985) has attested to "like that" → "li[t]e that", and Ogden (1999:74) attests to "I'm going" → "I'[ŋ] going", both involving final consonants that are not alveolars.[3] However, these are anecdotal observations, with no context, no audio available for detailed study, and no statistics on their frequency of occurrence. Thus it is important to establish whether non-alveolar assimilations occur in natural speech, their incidence, and (assuming they exist) the patterns of violation, the contexts in which they occur, and the sociolinguistic factors[4] governing them.

This is a case where phonological theory predicts unambiguous results, and we can directly translate the theory into an experiment that tests it. If we detect no violations or a rate small enough to be attributed to

---

3. There are also suggestive spelling mistakes like "sonetimes" (i.e. "sometimes") (200 ppm in Google) or "tinetable" (i.e. "timetable") (74 ppm in Google), in which a bilabial nasal seems to have assimilated to a following alveolar, or "Washinton" (for "Washington") (2093 ppm in Google), in which a canonically velar nasal is spelled as if it were alveolar. These might simply be spelling mistakes, but our study will reveal whether such assimilations actually occur in speech.

4. The BNC has metadata that allows data to be categorized by gender, dialect, and socioeconomic status.

speech errors[5], this would support the standard account. But finding substantial violations would overturn established theory, forcing us to take a probabilistic approach to phonology (e.g. Jun 2004) or supporting a gestural overlap account (e.g. Browman and Goldstein 1990, Ellis and Hardcastle 2002), and challenging theoretical proposals as to why, allegedly, only alveolars assimilate. Any of these would be an important revision of the established account.

**2.2. T/d deletion**

The project will also enable us to test on a large scale a very well-known sociolinguistic variable[6] which occurs at word endings. Deletion of a final /t/ or /d/ in a consonant cluster (first studied as a sociolinguistic variable by e.g. Wolfram 1969) is a very common phenomenon, especially in rapid or informal conversational speech. The following phonological context has consistently been found to be the strongest factor in determining the probability of deletion. Guy (1991) described the process in terms of the theory of Lexical Phonology. In his account, the past tense of a word is constructed in up to three stages and at each stage, a probabilistic deletion rule is applied. According to this account, monomorphemic words ending in t/d (e.g. “mist") have three chances at deletion, strong verbs where the final t/d is generated as the past-tense is formed (e.g. “kept”) will have the rule applied only twice, or just once if they are weak verbs (e.g. “stopped”). Tagliamonte and Temple (2005) found this pattern did not hold in a sample of 40 speakers from the York Corpus, and further qualitative investigations by Temple (2009) suggest that it may be more appropriate to treat so-called t/d deletion not as a (variable) phonological rule, but as a phonetic (continuous) speech process. The size and nature of the Spoken BNC and the techniques we propose to develop will allow us to reinvestigate the process on a large scale and to investigate commonalities between t/d deletion and assimilation, and if a unifying descriptive model can be found.

Guy's (1991) account has been supported by subsequent analyses based on VARBRUL/GOLDVARB variable rule models using stepwise logistic regression (e.g. Sankoff et al. 2005). However, since this approach was originally developed, related analysis techniques have been shown to be biassed towards false positives (e.g. Wilkinson and Dallal 1981, Copas 1983, Derksen and Keselman 1992). In the course of the analysis of our data, we will investigate the statistical limitations of VARBRUL-type algorithms. We shall use Bayesian statistical techniques, which have the potential to significantly improve upon the variable rule methodology.

**2.3. Secondary Research Goals**

To address the research questions, we will convert the text of the Spoken BNC to phonemic transcriptions and align every word and phoneme with the corresponding portion of the audio recording. This will allow easy access to desired words and phones. The Spoken BNC is currently not available to most researchers as accessing the tapes requires permission from the BNC Consortium. Furthermore, searching the tapes for samples of interest is prohibitively time-consuming. As members of the BNC Consortium, we and the British Library have privileged access to the audio recordings. A major goal is to publish them through the British Library Sound Archive, where copyright permits and with the appropriate license.

To align the transcriptions with the audio, we will further develop an automatic speech-to-phoneme alignment (ASPA) system, the Penn Phonetics Lab Forced Aligner[7], based on the HTK speech recognition toolkit (Young et al. 2009). (To transcribe and align 1,200 hours of speech manually would take approximately 240,000 hours of researcher time, c. 120 person-years.) We are the only BNC Consortium members with the experience and technical know-how to carry out the text-to-audio alignment. We will document and distribute our computer scripts, making our automatic techniques accessible to language researchers of whatever discipline. The final result will be a multi-million-word transcribed audio corpus

---

5. We will estimate the overall speech error rate on a speaker-by-speaker basis from the transcriptions. In cases of doubt, we will look for evidence of correction at each site. Obviously, if the speaker begins a correction immediately after a possible rule violation, it should be regarded as a speech error.
6. See e.g. references at http://privatewww.essex.ac.uk/~patrickp/TDbiblio.html.
7. http://www.ling.upenn.edu/phonetics/p2fa/

where researchers can locate the speech corresponding to points in the text. The Spoken BNC will then be a world-class resource for sociophonetics, phonetics and phonology research, or any area of study where ordinary spoken language is of interest (e.g. ethnomethodology).

## 3. Methods and Procedures

### 3.1. Digitization and Distribution

Digitization will be subcontracted to the British Library. The digitization effort is substantial, involving 1,213 90-minute tapes. Through our collaboration, the Library has digitized 43% of the Spoken BNC, a portion of which we have used to confirm that the tapes are in good condition and the audio quality is acceptable. They will produce archive-quality recordings, curate them, and handle the licensing and permissions. The resulting audio will be released under terms similar to the existing text BNC corpus (i.e. easily available to researchers). The Library will serve the audio over the web, and we will add the alignment information to the XML representation of the Spoken BNC and publish it electronically through the BNC website.

We have located the original agreements under which the recordings were made. With the assistance of an academic visitor, Dr. Céline Poudat, we examined the paperwork relating to the recordings, discovered how the recording numbers relate to the XML database, and resolved the IP/licensing conditions for use and eventual release of this data. This confirmed the legal and technical feasibility of releasing the digitized recordings with the transcriptions.

Public distribution requires blanking out names and other speaker-specific information from the recordings, because the speakers were promised anonymity. The necessary deletions are already marked in the text part of the Spoken BNC using XML <gap> tags; we will simply mute the audio corresponding to the <gap> tags[8]. Manual checking will ensure that deletions are successful. We have surveyed the literature on voice identification (e.g. Campbell et al. 2004, Kenny et al. 2008) and have detemined that release of the audio (with appropriate deletions) will not allow identification of the Spoken BNC volunteers, especially since comparison recordings should not be easily available after 17 years.

### 3.2. Aligning the Speech with the Text

In a recent pilot project, we established the feasibility of aligning the entire Spoken BNC. The British Library donated a six-hour sample of digitized audio from the BNC and we used it to establish that the audio quality was adequate and to evaluate the Penn Phonetics Lab Forced Aligner as a tool to align the audio with the corresponding transcriptions. This pilot project also provided estimates of the computer resources needed to align the Spoken BNC recordings with the text.

This pilot was very successful: in our sample, 83% of the computed phoneme boundaries were within 2 seconds of their correct position. Therefore, we should be able to select almost any word in the Spoken BNC, look up its position, and automatically display it on a screen. This capability will make over 8 million spoken words available for analysis.

Despite these successes, the Spoken BNC remains a challenging corpus for speech alignment. It has background noises ranging from televisions to birdsong and traffic, it covers a wide range of dialects, and the environmental acoustics varies from reverberant rooms to outdoors. Additionally, although the transcriptions are accurate, they do not capture all filled pauses, repairs, and other conversational phenomena, and do not precisely represent simultaneous speech by several people. Our pilot project has given us a good understanding of these problems and some strategies to solve them, but this project will include some work to improve the alignment technologies, which may yield publishable results.

An early technical goal is to determine a running confidence measure for the accuracy of alignment across

8. The fraction of the BNC already published as the COLT corpus (Stenström et al. 2002) was anonymized in a similar way.

the whole corpus. The pilot project established that we can automatically locate individual sounds within the Spoken BNC, since aligning the text yields a phoneme-level segmentation of the speech. We found that 24% of the phoneme boundaries were within 20 ms of expert human labels. This is not a large fraction, but when one boundary is accurate, almost all the boundaries within 2 seconds are also accurate. This fact should allow us to tell whether each given region is aligned well.

### 3.3. Phonetics/Phonology Research

To be able to make a strong statement about violations of a rule, one wants to find multiple instances in various words or word-pairs. A single exception could be disregarded on the assumption that it was just a speech error. If rule violations occur with <10% probability, and speech errors with <1% probability, we need at least 30 instances of each word-pair to allow a good chance of detecting an exception and determining whether it is a rule violation or a speech error. Consequently, we need a database on the scale of the Spoken BNC.

Where we have more than 30 instances per dialect of a word or word pair, we will be able to search for violations on a dialect-by-dialect basis, and cases with 1000 or more instances will allow stronger statistical tests and may permit quantitative comparisons of violation probabilities between dialects.

We have already identified 20 word pairs where non-alveolar assimilation might be found that occur with sufficient frequency, and 16 words where the same consonant combinations are found word-internally. There are 53 distinct words that are sufficiently frequent to study t/d deletion. Overall, in each experiment we expect to have about 5,000 tokens of acceptable recording quality. We will also include c. 3,000 controls. For instance, if we are looking at the possibility that “I'm trying” might be assimilated into “I'[n] trying”, we can determine the acoustic properties of [n] in that context by finding instances of “been trying”. Similarly, we can get canonical instances of /m/ in a similar context by searching for “I'm prying”.

We are then faced with the problem of determining whether deletion or assimilation has occurred in c. 10,000 instances. This would be difficult and costly to do manually, so we will use a semi-automated system to detect deletions or assimilations. The strategy is to train (for example) one HMM speech recognition model for /n/ and another for /m/ using the tokens in the spoken BNC which are unambiguous, for example training /m/ on "I’m prying" and /n/ on "been trying". We then use these "canonical" models to identify whether an assimilation occurs in tokens such as "I'm trying" where assimilation is possible. (This approach uses a variant of the technology we use for alignment.) We can then compute the relative likelihood of the two pronunciations in each case, and thus the HMMs will tell us whether deletion or assimilation has taken place in each case, or whether the likelihoods of deletion and/or assimilation are so close to chance that an automatic classfication is unreliable, because where both pronunciations are roughly equally likely. For these ambiguous cases, we will obtain human judgements from phoneticians (project team members). We will use each batch of expert judgments to improve the training of the automated system so that it better approximates human judgments. Aside from a few test cases, we should not need to spend human effort on the cases where the HMM makes a clear identification. Overall, we expect to need human intervention for only about 20% of the instances (c. 2000 maximum). This approach is known as "discriminative training". To ensure that the results cannot be attributed to any individual bias on the part of our team members, we will confirm any rule violations we find by presenting them to a set of naive listeners in a small perceptual experiment (a forced-choice identification test with naive listeners).

## 4. Facilities

In Oxford, the computations will be done on the Phonetics Lab's 48-core Condor cluster. Condor7 is a system for knitting together a group of Linux workstations, allowing large multi-processor jobs to be queued onto available processors when workstations are not otherwise fully utilized. Data and intermediate results will be stored on a dedicated disk array. The Phonetics Lab's Fibrenetix controller can store the necessary 6 Terabytes of storage and workspace and has sufficient data transfer rate to support the Condor cluster for our anticipated usage pattern. Oxford University Computing Services will provide data back-up.

**Summary of Timetable**

| Dates | Activities | Carried out mainly by |
|---|---|---|
| Oct 2010 - Mar 2011 | Installation of hardware and software | Pybus |
| Oct - Dec 2010 | Work to tune and improve the alignment techniques | Ravary, Yuan, Kochanski |
| Jan - Mar 2011 | Scripts for discriminative training | Kochanski, Ravary, Yuan |
| Apr - Jul 2011 | Aligning the demographic subset | Ravary, Yuan, Kochanski |
| Jul - Oct 2011 | Aligning the context-governed subset; writing | Ravary, Yuan, Kochanski |
| Aug - Oct 2011 | Prepare release of anonymized audio; writing | Kochanski |
| Nov - Dec 2011 | Release the aligned data; start running the discriminative training alignments | Kochanski, Ravary, Burnard |
| Jan 2012 | Website 1 - alignment techniques | Kochanski, Pybus |
| Jan - Apr 2012 | Running the discriminative training alignments; complete alignment paper(s) | Kochanski, Ravary, Yuan |
| May - Sept 2012 | Finding and classifying assimilation instances | Kochanski, Ravary |
| Oct 2012 - Apr 2013 | Writing assimilation paper; finding and classifying t/d deletion instances | Coleman, Temple, Kochanski |
| Apr 2013 | Website 2 - assimilation and t/d deletion results | Kochanski, Pybus, Coleman, Temple |
| May - Sept 2013 | Writing t/d deletion paper | Temple, Kochanski, Coleman |
| May - Sept 2013 | Final report | Coleman |

**Acknowledgements**
We thank Prof. Mark Liberman, Dr. Ladan Baghai-Ravary, Dr. Anastassia Loukina, Prof. Burton Rosner, Dr Jonnie Robinson, Dr Joanne Sweeney and Dr Victoria Drew for comments.