Phonological Reduction, Assimilation, Intra-Word Information Structure, and the Evolution of the Lexicon of English: Why Fast Speech isn't Confusing

Richard Shillcock
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place, Edinburgh
EH8 9LW, U.K.
[illegible]cogsci.ed.ac.uk

John Hicks
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place, Edinburgh
EH8 9LW, U.K.

Paul Cairns
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place, Edinburgh
EH8 9LW, U.K.

Nick Chater
Department of Experimental Psychology
University of Oxford
South Parks Road, Oxford OX1 3UD, U.K.
nick@psy.ox.ac.uk

Joseph P. Levy
Department of Psychology
Birkbeck College
Malet Street, London WC1E 7HX, U.K.
joe@cogsci.ed.ac.uk

Abstract

Phonological reduction and assimilation are intrinsic to speech. We report a statistical exploration of an idealised phonological version of the London-Lund Corpus and describe the computational consequences of phonological reduction and assimilation. In terms of intra-word information structure, the overall effect of these processes is to flatten out the redundancy curve calculated over consecutive segment-positions. We suggest that this effect represents a general principle of the presentation of information to the brain: information should be spread as evenly as possible over a representational surface or across time. We also demonstrate that the effect is partially due to the fact that when assimilation introduces phonological ambiguity, as in *fat man* coming to resemble *fap man*, then the ambiguity introduced is always in the direction of a less frequent segment: /p/ is less frequent than /t/. We show that this observation, the "Move to Markedness", is true across the board for changes in segment identity in English. This distribution of segments means that the number of erroneous lexical hypotheses introduced by segment-changing processes such as assimilation is minimised. We suggest that the Move to Markedness within the lexicon is the result of pressure from the requirements of a very efficient word recognition device that is sensitive to changes of individual phonological features.

Production of Fast, Continuous Speech

Normal speech is fast, casual, continuous and situationally located. These features mean that the pronunciation of any one word is likely to differ substantially from the way that it might be produced in "citation form" – the careful, ideal, isolated "dictionary" pronunciation of that word. First, the word may refer to a given in the discourse or the environment of the discourse, in which case even a poorly specified wordform may be sufficient for the listener to recognise the intended word. Second, the word may be a function word like *which*, *is* or *the*, whose identity is in part predictable from the syntactic context. In this case, top-down processing may augment the perception of the word (Shillcock & Bard, 1993). Third, the articulators must move between the physical positions required to produce consecutive speech segments and, particularly in faster speech, this may mean that the articulators do not reach the optimum position for any one segment before being required to move to new positions appropriate to the production of the next segment. This means that the pronunciation of any segment may be influenced by adjacent and nearby segments, even if those segments are on the other side of a word boundary. In this paper we develop, from an information processing perspective, a general principle concerning the production of fast speech and we reveal a striking generalisation about the distribution of segments in relation to these fast speech processes and the problem of word recognition[1].

The processes we describe, such as assimilation, are determined at least partly by the physical structure of the articulators. The lexicon of current English has developed against the backdrop of these physical constraints. We assume that the vocabulary of English has accommodated these physical constraints in terms of any implications they may have for information processing, intelligibility or ease of production.

Producing speech, at normal rates of perhaps 100–120 words a minute, is the most complex muscular activity in our repertoire. Similarly, listening to speech and deriving message-level interpretations virtually instantaneously, is arguably as complex a computational task as we undertake in such a timescale. In "fast" (that is, normal) speech processes, these pressures on speaker and listener are both present. We might expect the phonological content of speech to have been shaped so as to be readily speeded up with as little lack of intelligibility as possible. Below, we adopt a corpus-based approach to demonstrate the truth of

[1]Phoneticians contrast "citation form" pronunciation with "fast speech" pronunciation. The point we emphasise here is that normal, conversational speech is "fast", in these terms.

this claim. We will describe, in information theoretic terms, the results of building fast speech processes into a phonological transcription of a corpus of conversational speech. First, we describe the generation of the corpus and then we present statistical analyses of the implications for word recognition.

Generating a Realistic Speech Corpus

We describe in detail elsewhere the generation of an idealised phonological transcription of the London-Lund Corpus (LLC) (Svartvik & Quirk, 1980; Shillcock, Hicks, Cairns, Levy & Chater, 1995). The LLC is a corpus of orthographically transcribed conversational English speech, containing some 494,000 word tokens. The speech contains the false starts, filled pauses, repetitions and pausing that characterise normal speech. Some 19% of the word boundaries in the LLC are explicitly marked by pauses or speaker changeovers. The rest is continuous speech. We summarise below the procedure by which an idealised phonological transcription of this corpus was generated. The goal was to develop a phonological corpus that retained the psychologically important scale and representativeness of the original orthographic corpus.

First, the corpus was stripped of all annotations pertaining to speaker identity, prosody, pausing, and other paralinguistic information. Second, the orthographic word tokens were replaced by their respective citation forms, derived from the CELEX database[2]. A minority of tokens were transcribed as exceptions, either because they did not occur in CELEX, or they were fragments of words, or they had been given a phonological transcription in the original LLC. Third, phonological reductions and assimilations were introduced, as described below.

It is not possible to recreate the full richness of the phonological reduction of the original speech, which would differ between speakers, between utterances, and between different tokens of the same word. However, it is possible to create a version of the entire corpus that is rather more representative of normal speech than a simple concatenation of citation forms. One indicator of the validity of the transcription is the degree to which the resulting global statistics predict real processing behaviour. In this respect, there is reason to believe that the overall statistics we have extracted from the phonological version of the LLC are psychologically realistic. We have used these statistics successfully to model phoneme restoration and the acquisition of expressive phonology (see, *e.g.*, Chater, Shillcock, Cairns & Levy, 1995; Shillcock, Chater, Levy & Hicks, 1996).

We incorporated phonological reduction into the corpus for the function words only. In real speech, this class of words is disproportionately affected by phonological reduction, reflecting the fact that such words may be perceptually restored by their syntactic context. A single reduced version of each function word was taken from the reduced forms listed in the CELEX database, and was substituted for the citation-form version in all instances where this was appropriate to the phonological context. For instance, *had* was transcribed as /həd/, and *and* as /ənd/. Elsewhere we list all of the relevant function words and their reduced forms (Shillcock *et. al.*, 1995). These function words accounted for some 53% of the word tokens.

After the function word reductions had been made, we added to our phonological transcription an approximation of the effects of assimilation. These effects were applied across all words in the corpus, function and content. In real speech, consonants may be assimilated to the place of articulation of the following segment; for instance, the /t/ in *fat man* may be assimilated to the labial place of articulation of the following /m/ so that the /t/ acquires some of the characteristics of a /p/. In the extreme case the /t/ may become a /p/, giving *fap man*, but in other cases the /t/ becomes ambiguous between a /t/ and a /p/. In our phonological transcription of the LLC, we have adopted an idealised case in which the /t/ is changed into a /p/. Specifically, /d/ before a labial became /b/, /t/ before a labial became /p/, /d/ before a velar became /g/, /t/ before a velar became /k/. In addition, the six consonants were replaced by their unreleased versions when they occurred before another consonant. All of the rules regarding assimilation and the inclusion of the unreleased versions of the consonants were applied regardless of any word boundary information. Thus, assimilation could occur across the syllable boundaries within a polysyllabic word, and the first consonants in all consonant clusters became unreleased.

The procedure sketched above produced a phonological version of the LLC that is psychologically realistic in terms of the distributional statistics of its segments, in spite of the relative crudeness of such an automated procedure compared with a close transcription made by a trained phonetician.

The Information Structure of Words

We now consider each segment position in all the words of a particular length, and we calculate the predictability of the contents of that segment position. At one extreme is the case in which every possible segment occurs with exactly equal probability in the nth position in all words of length n or more: this segment position is therefore maximally informative and minimally redundant. At the other extreme is the case in which only one particular segment ever occurs in the nth position in words of length n or more: this segment position is minimally informative and maximally redundant. In reality, practically all cases fall between these two extremes. In the discussion below, we will be concerned with redundancy, so that the greater the redundancy, the more predictable is the processing of that particular segment position.

Redundancy was calculated as follows (see, *e.g.*, Yannakoudakis & Hutton, 1992):

Let S = Set of allowed phonemes

$$= \{s_1, s_2, s_3, \ldots, s_n\}$$

$$P_i = \Pr(s_i)/\sum\Pr(s_j)$$

[2]CELEX Lexical Database of English (Version 2.5). Dutch Centre for Lexical Information, Nijmegen.

$H\text{-max} = \log_2 n$

$H = -\sum_{i=1}^{n} P_i \log_2 P_i$

Entropy $E = H/H\text{-max}$

Redundancy $R = 1 - E$

Redundancy ranges between 0 and 1. Figure 1 shows an example redundancy curve calculated over the four segment positions for all of the four-letter words in the final phonological version of the LLC. This shows that the general shape of the redundancy curves is one which rises over time: the early segments of spoken words are more informative than the later ones. This is in agreement with the upward slope of the curves reported by Yannakoudakis and Hutton for corpora derived from text, and follows their convention of plotting redundancy, rather than entropy, against segment position. Note that the measure of redundancy used here is concerned only with the distribution of segments within a particular segment position, and not with the phonological information accumulated up to that point in a particular word. In this analysis, redundancy at any one segment position is in no way conditional on what has occurred earlier; later segments are not redundant because the identity of a particular word is constrained by what has gone before.

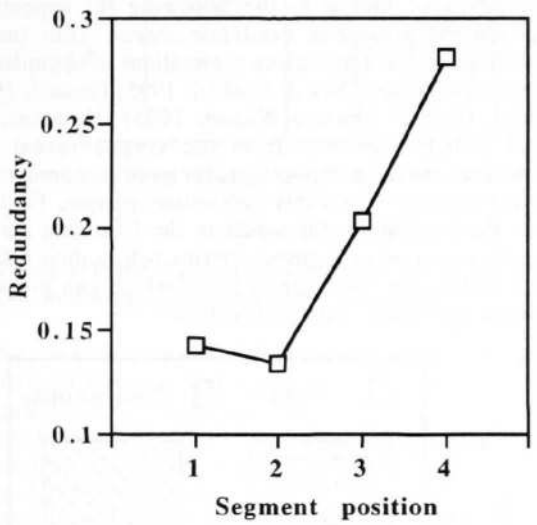

Figure 1: The redundancy curve across the four segment positions for all of the four-letter words in the LLC.

Figure 2 shows the aggregate redundancy curve for all of the words in the LLC of length 1–9 segments. The solid line is the redundancy curve for the words in citation form, before phonological reduction and assimilation. The dotted line is the redundancy curve for the words after the operation of these fast speech processes.

As Figure 2 shows, phonological reduction and assimilation increased the level of redundancy in the early parts of words and decreased it in the later parts of words. The overall effect is of a flattening of the curve, which we explore further, below. We have replicated this result with a different corpus of orthographically transcribed speech, which we have converted into a phonological transcription using the procedure described above. Figure 3 shows the corresponding data for a comparable amount of speech, the speech addressed to children aged 0–28 months in the CHILDES corpus (MacWhinney, 1991).

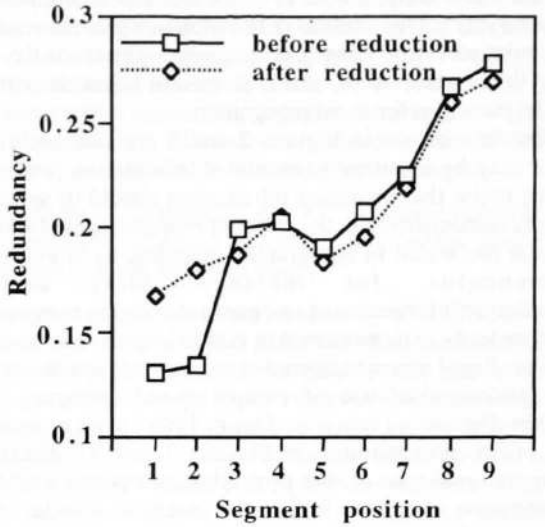

Figure 2: The effects of phonological reduction and assimilation on redundancy, for words of length 1–9 in the LLC.

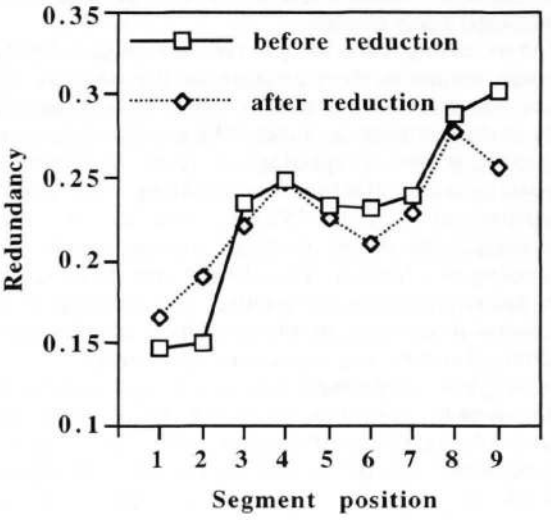

Figure 3: The effects of phonological reduction and assimilation on redundancy, for words of length 1–9 in the CHILDES Corpus.

The increase in redundancy in the earlier part of the curves in Figures 2 and 3 is principally due to the replacement of the closed-class words by their reduced forms. These words are typically short and their phonological reduction involves the replacement of citation form vowels by /ə/, as in *had* and *and*. This causes the distribution of segments in the earlier

positions to become more skewed, increasing redundancy. This increase in redundancy is reversed in the later part of the curve in both the LLC and the CHILDES corpora, with redundancy generally lower from the third segment on, following the application of fast speech rules (in Wilcoxon signed ranks tests, $z = -2.197$, $p < .03$; $z = 2.366$, $p < .02$, respectively). This reduction in redundancy is the result of assimilation introducing less frequent segments at the ends of syllables and words, and is discussed below in terms of the implications for word recognition.

The data shown in Figures 2 and 3 are consistent with what may be a general principle of information processing in the brain, that incoming information should be spread as evenly as possible over the relevant representational surface. This is illustrated in topographic mapping, as in Penfield's homunculus, for instance. More somatic stimulation/information at one particular region compared to another leads to more extensive cortical representation of the former. Equal representational space for both would result in an uneconomical use of computational resources. (For further discussion, see, *e.g.*, Smith, 1996.) We propose that the effect demonstrated in Figures 2 and 3 reflects the temporal analogue of this principle: the optimal profile of information over time is flat, the processor should not be subject to fluctuations in the informativeness of the speech signal at the phonological level. In information theoretic terms, the functional motivation for this phenomenon is that the frequency of occurrence of segments is evened out, thus allowing maximum transfer of information in an increasingly noisy channel.

The increasing redundancy curves in Figures 1–3 reflect a broader, communicative pressure on the structure of the lexicon, in that it is desirable for words to be distinguished early in their acoustic lifetimes. (We see this same pressure reflected at the morphological level in terms of a crosslinguistic preference for suffixing over prefixing (Hawkins & Cutler, 1985).) This is an abstract characterisation of the problem, appropriate to central processing in a lexicon of idealised forms. In contrast, the flattened redundancy curves reflect the exigencies of more peripheral processing, in which segment identity must be determined under time pressure in a noisy signal.

Although we only present data concerning across-the-board replacement by phonologically reduced forms for the closed-class words, the effects of reduction of the open-class words may be seen to be in line with the overall flattening of the redundancy curve described above. Open-class words predominantly have metrically strong initial syllables (Cutler & Carter, 1987), containing full vowels and not attracting the degree of phonological reduction that falls to the metrically weak later syllables. The loss of segments, in perception and production, from the later parts of words – as in the second syllable of *station* being replaced by a syllabic /n/, for instance – will remove from the statistics the very frequent, vulnerable segments such as /ə/ and /t/, and contribute to the further lowering of the high-redundancy part of the curve.

Conventionally, phonological reduction in the less informative, later parts of open-class words is seen as reflecting the fact that a particular word may be expected to have been recognised by this point. We present a complementary, alternative analysis here: these segments are more affected because they are more predictable simply on the basis of distance from the beginning of the word. Naturally there are other influences on intelligibility and informativeness, apart from those we have considered above: prosody is one such example. The tendency of fast speech processes to produce flatter redundancy curves is seen here as a general tendency that interacts with other factors.

Implications for Word Recognition

The effects of assimilation are generally taken to be deleterious to word recognition: the identity of a segment is being compromised, either by making it ambiguous with another segment or, as in our idealisation of the effect, by changing its identity completely to that of another segment. Assimilation might be expected to impair word recognition by excluding the intended word from the cohort of lexical candidates and/or by introducing erroneous candidates into the cohort. For instance, in the phrase *street car* the assimilation of the /t/ to the following /k/ appears to introduce the erroneous candidate *streak*. (For further psycholinguistic and modelling explorations of assimilation, see Marslen-Wilson, Nix & Gaskell, 1995; Gaskell, 1993; Gaskell, Hare & Marslen-Wilson, 1995). However, the global statistics derived from the corpus reveal that assimilation can have implications for word recognition that are very different from this pessimistic picture. Figure 4 shows the percentages of words in the LLC that are not uniquely specified in segmental terms before their offsets. These words are like *mar* or *part*, which can go on to become longer words, *mark* and *partner*.

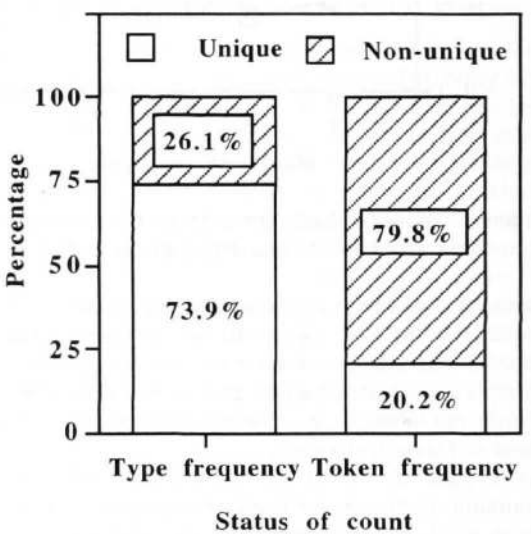

Figure 4: Percentages of uniquely specified words in the LLC, prior to application of the fast speech processes.

The 80% of non-unique tokens is double that reported by Luce (1986) in a similar, but dictionary-based, study, due to the inclusion in the current study of inflected forms, involving pairs such as *their* and *theirs*, and *walk* and *walked*. Compare Figure 4 with Figure 5, which shows the effects of applying phonological reduction and assimilation to the corpus. the result is a considerable reduction in the ambiguity introduced by the non-unique forms. These figures are based on the assumption that assimilated forms of words are lexicalised and count as lexical items, so that *last* expands into three lexical entries, ending with /t/, /k/ and /p/ respectively, the last two resulting from conjunctions like *last call* and *last place*.

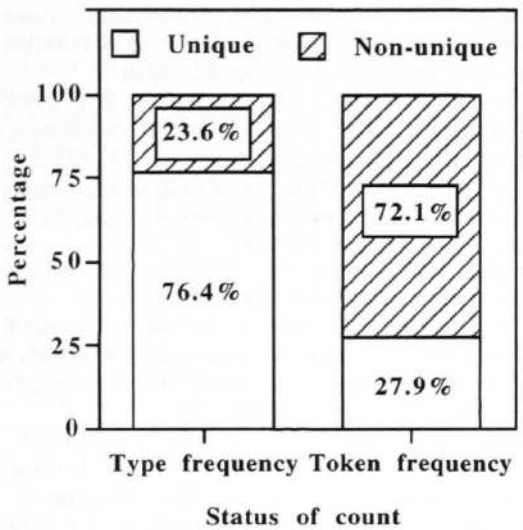

Figure 5: Percentages of uniquely specified words in the LLC, following application of the fast speech processes.

The "lexicalisation solution" of simply storing the assimilated forms as legitimate words is unlikely to be an adequate complete solution to the processing of assimilated forms; indeed, Gaskell *et al.* provide evidence against this solution and in favour of online inferencing to "undo" the assimilation, for certain stimulus materials. However, lexicalisation does not emerge, in the current study, as such a clumsy or expensive option as might have been thought. In this instance, it causes an overall increase from 18,671 word-types to 20,630, an increase of 9.5%, to achieve a 7.7% increase in unique tokens[3]. Lexicalisation of frequently encountered assimilated wordforms and online inferencing as a parallel/back-up process may be an attractive compromise.

The Distribution of Segments

We have seen in Figures 2 and 3 that fast speech processes increase redundancy in early segment positions and decrease it in later segment positions. The phonological reduction of the, predominantly short, function words is primarily responsible for the early increase in redundancy. Although assimilation was applied across all the words in the corpus irrespective of word boundaries, it has its effect at the ends of syllables and words and is therefore primarily responsible for the decrease in redundancy in the later parts of the curves. We have already briefly alluded to the fact that this effect is due to assimilation replacing frequently occurring segments by less frequently occurring ones. We now elaborate upon this observation concerning segment distribution, which we term the "Move to Markedness", which is true of English and which we predict will hold for other languages:

> "If segment x is influenced by the following segment to resemble segment y, then x will occur more frequently in the language than y."

For example, /t/ may be influenced by a following /p/ to resemble /p/ itself, as in *last place*, and statistical analysis of our phonological version of the LLC reveals that /t/ occurs more frequently than /p/. These same overall frequencies (see Shillcock *et al.*, 1995) show that this relation holds for all of the segment changes we list below, which include some assimilations not instantiated in the version of the LLC we have described above, as well as changes involving unreleased consonants: /d/ → /b/, /t/ → /p/, /d/ → /g/, /t/ → /k/, /n/ → /m/, /n/ → /ŋ/, /s/ → /ʃ/, /b/ → /b`/, /d/ → /d`/, /k/ → /k`/, /g/ → /g`/, /p/ → /p`/, /t/→ /t`/. The fact that all 13 of these changes are in the direction of a less frequent segment – towards markedness – is significant by the binomial test, $p < .001$; for the seven assimilations alone, $p < .008$[4].

The functional motivation that we claim for this distributional constraint is that it offsets interference with spoken word recognition from the effects of assimilation. Assimilation causes segments to become more ambiguous or to change their identity. These effects are potentially disruptive to word recognition, given that the processor seems to be sensitive to single changes at the feature level (see, *e.g.*, Marslen-Wilson, 1993): a change from /t/ to /d/ at the end of the word *apricot* is sufficient to switch off priming by that word. The processor needs to prevent a change in segment from causing the activation of a large new cohort of erroneous lexical candidates at the same time as the correct candidate is apparently disqualified. For instance, if the correct candidate is *batman*, it may be excluded from the cohort of active lexical candidates by assimilation producing *bap-*. This candidate can be rescued either by the lexicalisation account discussed above, in

[3]Note that there is marginal evidence for the lexicalisation of fast speech forms in those individuals who miswrite *could have* as *could of*, or *handbag* as *hambag*.

[4]Note that this is a conservative test of the observation, in that the judgements were made on a version of the corpus in which the fast speech processes had already been modelled, and in which the numbers of the less frequent segments had therefore been increased. That is, the disparity between the frequencies of /t/ and /p/ is somewhat offset by assimilation, in which the former are turned into the latter. If any aspect of word recognition employs pre-assiilation, citation-form templates, then the observation is true in even more striking form, with greater frequency disparities.

which *bapman* is stored, or by an online process of inference once the following /m/ is encountered. Both of these solutions may exist in human listeners, and may be determined by the frequency of the assimilated wordform. The other aspect of the problem is the possible inclusion of erroneous candidates in the cohort. The Move to Markedness ensures that the number of erroneous candidates is minimised; this is the motivation for the constraint. Its efficacy is ensured by the fact that the new segment identity is lower in frequency than the segment it replaces. This is illustrated by our example: if the /t/ in *batman* becomes a /p/, then *bap*, *baptism*, *baptistery* and *baptise* are erroneously added to the words already activated, but *batch*, *batchy*, *bateau*, *bateleur*, *batik*, *batiste*, *baton*, *batsman*, *battels*, *batten*, *batter*, *battery*, *batting*, *battle*, *battledore*, *battlement*, *battue*, *batty* and *batwoman* will be deactivated (words taken from the Concise Oxford Dictionary). It is impossible to ensure that no erroneous candidates at all are activated, but the fact that the new segment identity is lower in frequency than the segment it replaces reduces the number of new and erroneous candidates. This argument applies even if assimilation only makes the segment ambiguous between two segments as opposed to completely switching its identity; in this case, the Move to Markedness limits the confusion sown rather than actually reducing it.

Finally, the Move to Markedness has implications for speech segmentation. The probability of the transition between consecutive segments is potentially valuable information for deciding that two segments straddle a word boundary (see, *e.g.*, Cairns, Shillcock, Chater & Levy, 1994); more unlikely transitions are better candidates for boundaries. Movements towards markedness make such segmentation decisions more likely, in that they will tend to produce less frequent transitions.

Conclusions

A corpus-based approach can reveal novel aspects of language structure and processing. On the basis of empirical exploration of the London-Lund Corpus, we suggest a general goal of fast speech processes: producing a flatter redundancy curve over segment positions in words. We also report a functional constraint on the distribution of phonological segments, the Move to Markedness. This constraint is motivated by the requirements of the word recognition process, in which there is a virtually instantaneous sensitivity to individual feature changes. We claim that the Move to Markedness is the result of ease of intelligibility selecting for a particular distribution in the lexicon, given the presence of physically determined effects such as assimilation. This analysis reveals the information theoretic consequences of these effects.

Acknowledgements

This research was supported by ESRC (UK) grants R000 23 3649 and R000 22 1435. We would also like to acknowledge valuable discussions with Mark Ellison.

References

Cairns, P., Shillcock, R.C., Chater, N., Levy, J. (1994). Lexical segmentation: the role of sequential statistics in supervised and unsupervised models. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, Georgia Institute of Technology.

Chater, N., Shillcock, R., Cairns, P. & Levy, J. (1995). Bottom-up explanation of phoneme restoration: Comment on Elman and McClelland (1988). Submitted to Journal *of Memory and Language*.

Cutler, A. & Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, **2**, 133-142.

Gaskell, G. (1993). *Spoken word recognition: A combined computational and experimental approach*. PhD thesis, Birkbeck College, University of London.

Gaskell, G. Hare, M., & Marslen-Wilson, W.D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive Science*, **19**, 407–439.

Luce, P. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics*, **39**, 155–158.

Macwhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ, Erlbaum.

Marslen-Wilson, W.D., Nix, A. & Gaskell, G. (1995). Phonological variation in lexical access: Abstractness, inference and English place assimilation. *Language and Cognitive Processes*, **10**, 285–308.

Marslen-Wilson, W.D. (1993). Issues of process and representation in lexical access. In G.T.M. Altmann and R.C. Shillcock (Eds.) *Cognitive Models of Speech Processing: The Second Sperlonga Meeting*, Erlbaum.

Shillcock, R.C. & E.G. Bard. (1993). Modularity and the processing of closed class words. In Altmann, G.T.M. & Shillcock, R.C. (Eds.) *Cognitive models of speech processing*. The Second Sperlonga Meeting. Erlbaum.

Shillcock, R.C., Chater, N., Levy, J.P., & Hicks, J. (1996). Order of acquisition of expressive phonology in English: the effect of phonotactic range. *Manuscript*.

Shillcock, R.C., Hicks, J., Cairns, P., Levy, J., & Chater, N. (1995). A statistical analysis of an idealised phonological transcription of the London- Lund corpus. *Submitted to Computer Speech and Language*.

Smith, J. (1996). *Neural Networks, Information Theory and Knowledge Representation*. PhD dissertation, University of Edinburgh.

Svartvik, J., & Quirk, R. (eds.) (1980). *A Corpus of English Conversation*. Lund Studies in English 56. Lund: Lund University Press.

Yannakoudakis, E.J. & Hutton, P.J. (1992). An Assessment of *N*-phoneme statistics in phoneme guessing algorithms which aim to incorporate phonotactic constraints. *Speech Communication*, **11**, 581–602.