York Papers in Linguistics
Series 2

Issue 13

December 2013

ISSN 1758-0315

Editors:
Natalie Fecher, Dima Al-Malahmeh, Mariam Dar, Ella Jeffries, Ania Kubisz, Man Ki Theodora Lee, Kaj Nyman, Rebecca Woods

Editorial Note

This is the thirteenth issue of *York Papers in Linguistics Series 2* (YPL2). It contains papers from staff, postgraduate students and associated colleagues of the Department of Language and Linguistic Science, University of York, United Kingdom.

The editors would like to thank the anonymous reviewers for their invaluable comments and advice during the preparation of this issue, and the editorial board and YPL members for their help and support.

Publication here does not prejudice future appearance elsewhere.

For more information on the current issue and past YPL publications visit http://www.york.ac.uk/language/ypl. Please address all further enquiries to ypl@york.ac.uk.

Natalie Fecher

York, 2013

Contents

WHY DON'T DEAF READERS GARDEN-PATH?

LINDA COULTER[1] & HELEN GOODLUCK[2]

[1]Ardkeen, Northern Ireland

[2]University of York

Abstract

Self-paced reading was used in three experiments to examine online sentence processing in sixteen deaf adults. Experiment 1 examined processing of simple declarative sentences. Experiment 2 examined processing of temporarily ambiguous sentences with filler-gap dependences, and experiment 3 explored sentences with a temporary ambiguity between direct object and subject analysis of an NP. The results from each experiment show that the deaf participants did not process the experimental items in the same way as the hearing participants. In experiment 1 we found no syntactic 'wrap up' effect (slowing of reaction times at the end of a sentence) for the deaf; in experiment 2, we found no tendency for the deaf to erroneously anticipate the position of a *wh*-word; and in experiment 3, we found that deaf readers do not follow a strategy of attempting to incorporate the most recently input material into the parse of an existing phrase. We suggest that lack of access to prosodic information in the language input during acquisition (because of auditory deprivation) results in atypical development of parsing strategies.

1. Introduction

It has long been documented that the reading ability of deaf individuals falls well below that of their hearing counterparts (Conrad 1979, 1977; King & Quigley 1985; Watters & Dowering 1990). More recent research indicates that the situation has not improved significantly over the years (Paul 2003, 1998; Traxler 2000). Although the past three decades have seen the development of a vigorous field of research in sentence processing in the hearing population, to our knowledge, much of this literature has not yet impacted on the study of language processing in the hearing impaired. Our aim in this paper is to begin to make a connection between sentence processing research in hearing adults and reading by the deaf.

1.1 Syntactic knowledge

For hearing children, syntactic structures are learned through daily interaction with parents and others, and not through any formal teaching. However, this is not the case for the deaf child. Most deaf children start school and begin learning to read without the syntactic knowledge that their hearing peers possess. Miller (2000) suggests that pre-lingual deafness may prevent readers from spontaneously internalising the syntactic rules of spoken language.

Successful reading involves accessing and using written information to construct meaning (Paul 2003). Although many deaf readers can access words and their semantic interpretation, this does not necessarily lead to comprehension of the overall meaning of the text (Lillo-Martin, Hanson & Smith, 1991; Kelly 1995; Miller 2000). In order to construct meaning from complex syntactic structures one needs to be able to assign thematic roles (such as agent or patient) to constituents and to link moved constituents (such as question words) to their

original position. Quigley, Power & Steinkamp (1977) and Hatcher & Robbins (1978) found that while deaf subjects could comprehend an idea expressed in a simple sentence, they could not understand the same idea when expressed in a sentence with a more complex syntactic structure. This indicates that lack of syntactic knowledge–or inability to apply such knowledge while reading–does affect comprehension. Even though deaf readers may have sufficient 'world knowledge' to enable them to understand an idea, lack of syntactic knowledge/processing ability hinders comprehension in a complex structure. Miller (2005) suggests that deaf individuals may not process text syntactically but rather try to derive meaning 'by applying their prior knowledge and experience in interpreting its content words'. Others claim that some deaf readers may simply rely on a limited set of syntactic rules (Hatcher & Robins 1978, Quigley 1982).

Quigley, Power, & Steinkamp (1977) tested a group of deaf students aged ten to nineteen using structures such as question formation, negation, conjunction, relative clauses, pronominalization and complementation. These syntactic structures were presented singly in sentences. Quigley et al.'s results revealed that the average eight-year-old hearing subjects performed better than the average eighteen-year-old deaf participants. The main differences occurred in constructions that involved holding information in short-term memory while waiting for semantic or syntactic resolution. It has been shown that many deaf readers who have difficulty comprehending complex syntactic structures such as subordination and conjunction tend to impose a subject–verb–object pattern on English sentences (Quigley 1982; Berent 1988). Miller (2000) suggests that for such individuals, subordinate clause markers function as co-ordinating conjunctions.

Lillo-Martin, Hanson and Smith (1992) studied reading English relative clauses by deaf persons who had learned sign language from birth. They found comprehension was significantly worse for relative clauses that modified the subject of the sentence than relative clauses that modified the object.

1.2 Automaticity

Kelly (2003) found that in less skilled deaf readers: (a) lexical access is not fast enough to allow word meanings to be combined with words read earlier before they have decayed from working memory; (b) poor performance in syntactic analysis affected their subjects' abilities to fully utilise vocabulary knowledge; and (c) sentences with complex syntactic structures were more difficult for less skilled readers to understand. These findings, together with the fact that average deaf readers read significantly more slowly than skilled deaf readers, led Kelly to claim that even among deaf college students, low automaticity is a significant obstacle to reading comprehension.

Hearing children pass through developmental stages, initially focusing on decoding print into sound and recognising word shapes (usually between 6-8 years of age). This is usually fairly automatic by age nine, enabling the child to move on from word level processing to processing at sentence level, then text level and eventually to fluent reading. When decoding is automatic as in fluent reading, attention can focus on comprehension. According to Snyder & Downey (1991), if a child cannot reach this level of automaticity then he/she may be unable to advance to the stage of reading for meaning. Given that the average deaf school leaver has a reading age comparable to that of a nine year old hearing reader it is possible that many do not reach the level of automaticity required in order to progress to fluent reading (Traxler 2000).

1.3 Modality independent comprehension skills

Recent research has focussed on whether cognitive factors over and above language comprehension in general might play a more vital role in the comprehension process than previously thought (Marschark, Sapere & Convertino 2009). Their findings are based on the fact that 'the weaknesses exhibited by deaf students in many of the subskills involved in reading are paralleled by similar weaknesses in understanding sign language' (Marschark et al. 2009:359). Marschark et al. found that while deaf readers had difficulty monitoring their own comprehension of written material, deaf signers also appeared to have poor comprehension monitoring skills during communication in sign language or lip reading. They suggest that rather than continue to focus attention on the areas mentioned above, it might be advantageous to investigate the interaction of cognitive processing, language comprehension and learning, regardless of which mode of communication is used (print or sign language).

2. Goals of the present study

The key subskills of reading described in the previous section have all been examined in an effort to find an explanation for the difference in reading ability between deaf and hearing readers. The technique used in our study was self-paced reading (see below for a description of the technique), which has widely been used in studies of reading in hearing adults. Kelly (1995), Wauters, Telling, van Bon & Mak (2007), Van Hoogmoed, Verhoeven, Schreeder & Knoors (2011) and others have used this technique with deaf subjects; the focus of these studies has been the overall comprehension of written passages and the ease of processing different classes of content word vocabulary and morphology. In a recent self-paced reading study closer to the intent of this study, Traxler, Corina & Morford (2010) found that deaf signers reading English had patterns of relative clause and passive sentence comprehension similar to those for hearing adults, particularly for subjects who were native signers.[1]

Our study examined whether syntactic reading patterns and the parsing strategies used by hearing adults are also used by deaf individuals. There were some clear predictions from the studies reviewed in the previous section, although it was not always the case that previous findings could be applied to the present testing situation. Thus the documented difficulty deaf readers have with structures that require holding an element in storage (Quigley et al. 1977) led to the prediction that deaf readers would be more prone to errors in interpreting *wh*-questions, and the shorter digit spans of deaf individuals (King & Quigley 1985) led to the prediction that subjects would have more difficulty with sentences that require revision of the syntactic analysis.

Our first experiment examined the online processing of simple declarative sentences which could be read quite easily by most hearing children. In the hearing population, function words such as *a* and *the* are read more quickly than content words, creating a rise and fall pattern in declarative sentences (Aaronson & Scarborough 1976). This experiment thus addressed the issues of automaticity in lexical retrieval and the construction of basic syntactic structures.

Experiments 2 and 3 examined whether deaf readers employ the same strategies for syntactic analysis that hearing readers do**.** Experiment 2 looked at the processing of questions**.** In questions, the reader needs to link the sentence initial question phrase to its original position in order to arrive at the correct interpretation. For the hearing population, the processor has been shown to actively seek out a site for the *wh*-phrase during online processing, sometimes erroneously anticipating a position for the *wh*-word where one does not exist (Crain & Fodor,

[1] We have only the abstract of Traxler et al.'s (2010) study.

1985; Frazier & Clifton 1989, Stowe, Tanenhaus & Carlson 1991, 1986). This active search for a position for the question word plausibly follows from pressure to reduce load on working memory (holding the question word in memory until it has been integrated into the incoming sentence). The aim of this experiment was to determine whether or not deaf individuals process questions in the same way as the hearing population and whether pressure on working memory for the deaf would exacerbate patterns that are found in the hearing population.

In experiment 3 we attempted to replicate with deaf readers the findings of Frazier & Rayner (1982). Hearing readers prefer to interpret the phrase *his horse* in a sentence such as *Before the King rides his horse is always groomed* as object of the first verb (*rides*) – an incorrect analysis, since the phrase is in fact subject of the main verb *is*. This is attributed to a processing principle that dictates that incoming material is associated with the most recently processed phrase.

3. Method

3.1 Procedure

Before the experiment began, written instructions were given on the computer screen used for the presentation of the words. These instructions were also given to the deaf participants in sign language. At the beginning of each sentence the instruction 'PRESS TO BEGIN NEW SENTENCE' appeared on the screen. Each word appeared in the centre of the screen and the participant pressed a button on a response box to remove the current word and bring a new word into view. On the screen that appeared after the last word in a sentence, a '$' sign was shown to indicate the end of the sentence (a full stop in the centre of the screen was considered to be too small for comfortable viewing). Participants familiarised themselves with the procedure with practice items before beginning the experiments. The materials for the three experiments were intermingled in blocks of one token of each sentence type. Forty filler sentences were also included.

A secondary comprehension task was used to ensure that participants were reading for meaning: subjects were required to answer a yes/no question about the previous sentence. The questions were randomised throughout the test battery with a ratio of one question to every three sentences. The question appeared on the screen immediately after the sentence and the subject answered by pressing one of two buttons, Y for yes or N for no. This task was chosen as opposed to sentence repetition, which is also commonly used in self-paced reading tests, because we felt it to be more appropriate for (place less burden on) our deaf participants. The average number of correct responses was 90% for the deaf and 94% for the hearing.

3.2 Participants

Sixteen deaf and sixteen hearing adults from the Belfast, Northern Ireland region took part in the study. Only deaf individuals who were either born deaf or prelingually deafened and who had hearing parents were included. No subjects had cochlear implants. They were all educated in an oral system, learned English as their first language and could be termed high achievers within the deaf population. Both deaf and hearing participants had an education level of the UK General Certificate of Secondary Education or above; most of the deaf

participants had 'A' levels (university entrance examinations) or degrees. Reading assessment tests were not carried out, but we reasoned that in order to pass written examinations in various subjects at GCSE level the participants' reading ability must be adequate.

4. Experiment 1

4.1 Previous research

Aaronson & Scarborough (1976) and others have shown that in self-paced reading tests hearing readers produce a rise and fall pattern with shorter reaction times at functions words (determiners, auxiliaries, prepositions) than at content words (nouns, verbs). This may be attributed to the difference in time taken to access the different categories and/or to time taken to integrate material at the end of a phrase (in languages such as English, where function words typically are phrase initial).

4.2 Goals

The goal of Experiment 1 was to see if reading times for deaf readers differentiated between function words and content words, using simple sentence types. We considered it useful to include materials that could easily be read by most hearing children, since the literature indicates that the reading age of the average deaf eighteen year old is eight to nine years (Traxler 2000).

4.3 Materials

The experimental items were of four types:

(1) The clown entertained his audience (Det, N, V, Det, N)
(2) A teenager would choose this holiday (Det, N, Aux, V, Det, N).
(3) An old man walked up our street (Det, Adj, N, V, Prep, Det, N).
(4) The babies in each cot slept soundly (Det, N, Prep, Det, N, V, Adv).

(Det = Determiner, N = Noun, Aux = Auxiliary, Adj = Adjective, Prep = Preposition, Adv = Adverb). There were five tokens of each of the four sentence types.[2]

4.4 Results

Figures 1–4 below show the mean reading time for each group at each position in sentence types (1–4). In this and the two experiments reported below outliers were identified as scores higher than 2.5 times the standard deviation from the mean score and were replaced with the subject's mean score for that position in that condition.

[2] A complete set of the experimental materials used in all three experiments is available on request.

As can be seen from each figure, the deaf are reading much more slowly than the hearing participants and there is more variation in their reading times.

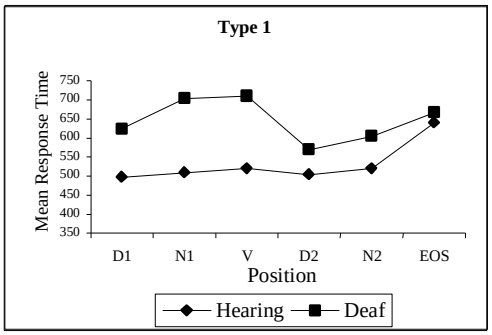


Fig1

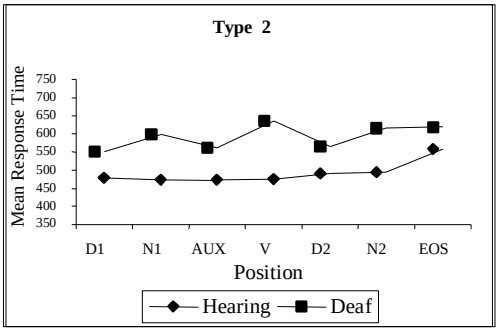


Fig.2

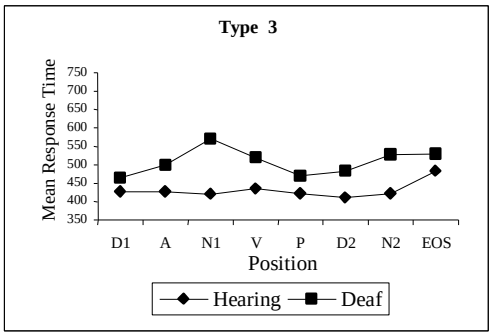


Fig.3

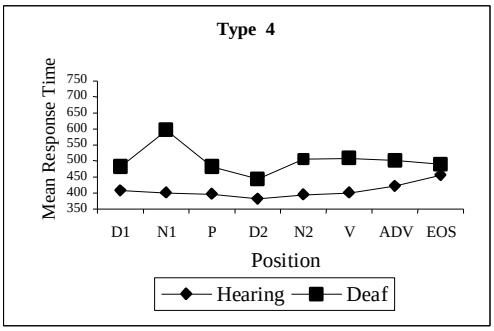


Fig.4

Figures 1–4: Mean reading times (msecs). Sentence Types 1–4.

An inspection of Figures 1–4 shows that the only position at which the reading times for the hearing change appreciably is at the end of sentence (EOS) position. Elevated reading times at EOS in studies with hearing participants are thought to reflect some aspect of sentence final processing and have been referred to as an end of sentence 'wrap-up effect' (Mitchell & Green 1978; Just & Carpenter 1980; Rayner, Sereno, Morris, Schmauder & Clifton 1989; Millis & Just 1994). An EOS effect for the deaf subjects is seen only for sentence type 1.

Figures 1–4 also show that reading times for the deaf are shorter at function words and longer at content words, as we expected based on the previous literature on hearing subjects. Paired sample t tests comparing the transitions from one position to the next were carried out to examine the effects of position. Reading times for the hearing group did not vary much from one position to another throughout the sentence. Of the 25 transitions from one word to the next for the four sentence types, only three were significant or near significant, and each of these three were for the transition between the last word and the EOS marker (Sentence type 1: N2 to EOS, $t(15) = -5.09$, $p < .001$, by subject, $t(4) = -2.61$, $p = .059$, by item. Sentence type 2, N2 to EOS, $t(15) = -3.66$, $p = .002$, by subject, $t(4) = -5.51$, $p = .010$, by item; Sentence type 3: N2 to EOS, $t(15) = -2.50$, $p = .025$, by subject, $t(4) = -6.00$, $p = .004$, by item).

For the deaf group, of the 25 transitions 12 were significant by subject, of which only one was the transition between the last word and the EOS marker (Sentence type 1: transition D1 to N1, t(15) = -3.72, p = .002, all probabilities here and below, two tail; V to D2, t (15) = 2.58, p = .021; N2 to EOS, t(15) = -2.31, p = .036. Sentence type 2: N1 to Aux, t(15) =2.48, p = .025. Sentence type 3: A to N1, t(15) = -2.56, p = .008; V1 to P, t(15) =2.75, p = .015; D2 to N2, t(15) = -2.56, p = .022). Sentence type 4: D1 to N1, t(15) = -3.17, p = .006; N1 to P1, t(15) = 3.36, p = .004; P1 to D2, t (15) = 2.51, p = .024; D2 to N2, t(15) = -2.72, p = .016). By item, nine of the 25 transitions were significant, of which again only one was for the last word to EOS marker, again for sentence type 1 (Sentence type 1: V1 to D2, t(4) = 3.40, p = .027; N2 to EOS, t(4) = -4.24, p = .013. Sentence type 2: Aux to V1, t(4) = -3.58, p = .023; V1 to D2, t(4) =3.93, p = .017; Sentence type 3, Adj to N1, t(4) = -3.11, p = .036; D2 to N2, t(4) = -2.78, p = .050. Sentence type 4: N1 to P1, t(4) = 3.17, p = .034; P1 to D2, t(4) =6.06, p = .004; D2 to N2, t(4) = -3.94, p = .017).

4.5 Discussion of Experiment 1

The data from deaf participants showed effects of position across the sentences types except at the end of sentence marker, whereas the hearing only produce an effect of position at EOS. Kelly (1995) did report an EOS wrap-up effect with deaf subjects. However, Kelly's analysis was carried out on the last word of the sentence and the word preceding it, whereas we measured EOS wrap-up between the last word in the sentence and the end of sentence marker. It may be that the difference Kelly found between the penultimate and final words in the sentence reflected the difference between function and content words, similar to that found for the deaf subjects in this experiment (Kelly's materials were not published and so we cannot directly verify this).

The overall reading patterns produced by the deaf are what we expected from previous research on hearing persons: reading times rose from determiners to nouns, and fell from verbs to determiners and other function words. For the hearing subjects, the absence of any effect of position except EOS wrap up can be attributed to the subordinate task we chose. Yes–no questions as a subordinate task have been found to lead to faster reaction times and smoother profiles across the sentence than sentence repetition as a subordinate task (Aaronson & Scarborough 1976).

5. Experiment 2

5.1 Previous research

A substantial literature in sentence processing by the hearing, using self-paced reading and other techniques, has shown that (1) overall, reading times are longer for wh-questions than for declaratives or yes–no questions, and (2) readers may incorrectly predict a position in the incoming sentence in which to place a *wh*-question word (Crain & Fodor 1985; Stowe 1986; Frazier & Clifton 1989; Stowe et al. 1991). With respect to (1), reaction times are expected to be longer in the subordinate clause in (5a) than in (5b):

(5) a. The chef wondered what he would cook the lamb with that night.
b. The chef wondered if he would cook the lamb with roast potatoes that night.

(5a) contains an embedded question, in which the processor must link the wh-word *what* to

the position after the preposition *with*. (5b) contains an *if* clause (an embedded yes–no question), in which no such linkage is required. (5a) also contains another potential position for the wh-word, as object of the verb *cook*. This location for the *wh*-word is shown to be incorrect when the phrase *the lamb* is encountered. Studies beginning with Crain & Fodor (1985) have shown that hearing subjects incorrectly anticipate a position for the *wh*-word after the verb. This is reflected in longer reading times to access the word after the determiner that follows the verb in *wh*-question sentences such as (5a) than in yes–no sentences such as (5b). It is reasoned that the processor initially places the *wh*-word as object of the verb, only to discover its mistake when it accesses the following determiner, leading to a delay in requesting the noun that follows the determiner. This is termed a 'filled-gap' effect – the place (gap) posited for the *wh*-word turns out to be filled by another noun phrase. This is one type of error known in the literature as a 'garden path' effect, the processor temporarily positing an incorrect analysis of the input.

5.2 Hypotheses

We expected the hearing subjects to show longer reading times in sentence such as (5a), and to show the 'filled gap' effect in encountering the object of the embedded verb. Following Quigley, Power & Steinkamp (op cit.), we expected holding a *wh*-word in memory would produce a greater effect for deaf readers than for hearing readers, and that either this greater load would enhance the effect of misplacement of the *wh*-word, and/or that it would cause a breakdown in parsing procedures normally used by hearing adults.

5.3 Materials

There were ten sentences frames of the type in (5a/b). The two versions of each sentence (*wh*/*if*) were counterbalanced across experimental lists, each participant seeing only one form of each sentence, for a total of five tokens of each condition per subject. The order of the sentences was arranged so that no two tokens of the same condition were presented adjacently.

5.4 Results

Figures 5 and 6 below show the mean reading time per word for each group at each position in the critical region (the verb and subsequent positions) of the sentence for each condition (*wh* (cond. 1) vs. If (cond. 2)) and syntactic position (V = verb, D = determiner, N = noun, P = preposition). As can be seen the pattern for the deaf is quite different from the pattern produced by the hearing subjects. Although for both groups reading times are generally longer for the *wh* condition than for the *if* condition, only the hearing group displays the expected pattern of an increase in reaction times at the determiner and noun. The deaf group, by contrast, show a pattern of higher reaction times at content words (the noun and verb) than at function words, reminiscent of their performance in experiment 1.

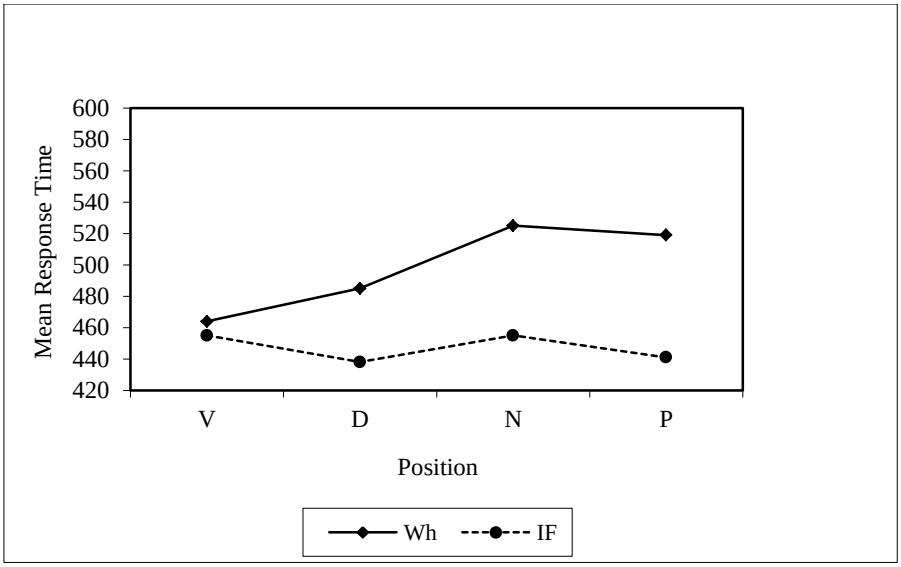


Figure 5: Mean reading times. Hearing group.

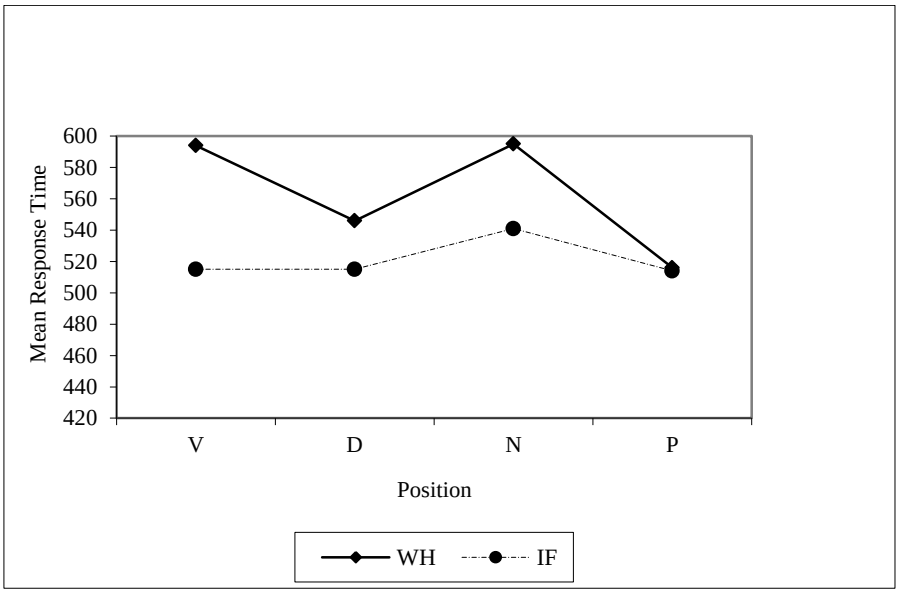


Figure 6: Mean reading times. Deaf group.

Four (position) x 2 (condition) repeated measures ANOVAs were carried out on the mean reading time for position in the critical region of the sentence. Analyses were carried out for both subjects (F_1) and items (F_2). Separate ANOVAs were conducted for the hearing and deaf groups.

The ANOVA for the hearing group revealed a significant effect of condition for subjects (F_1 (1,15) = 16.10, p = .001), but not for items. There was no effect of position for subjects, but there was an effect for items (F_2 (3,24) = 5.12, p = .019). There was a significant position x condition interaction for both subjects and items (F_1 (3,45) = 3.10, p = .047, F_2 (3,24) = 4.73,

p = .013). Paired sample t tests[3] for condition (*wh* vs. *if*) x position showed no significant difference in reading times between conditions at the verb. However, consistent with the literature on the filled-gap effect there was a significant difference immediately after the verb. Elevated reading times for the *wh*-condition were recorded at the three positions following the verb, i.e. at the determiner (t(15) =2.72, p = .016), noun (t (15) = 2.90, p = .011) and preposition (t(15)= 3.64, p = .002).

The results of the ANOVA for the deaf participants showed an effect of condition for subjects (F_1 (1,15) = 5.00, p = .041), but not for items. The deaf group also produced an effect of position for subjects (F_1 (3,45) = 4.02, p = .024), but not for items. There was no position x condition interaction for subjects or for items. Planned comparison t tests showed that there was a near significant difference in reading time between the *wh* items and the *if* items at the verb (t(15) = 2.05, p = .058) and a significant difference at the noun (t(15) = 2.24, p = .041), with elevated reaction times for the *wh* items. T-tests to examine the effect of position (transitions from one position to the next) obtained in the main ANOVA revealed a significant difference between the noun and preposition in the *wh* items, (t(15) = 4.25, p = .001) but not in the *if* items.

5.5 Discussion of Experiment 2

Our results for the hearing subjects are exactly as we predicted based on the previous literature, there are elevated reaction times to *wh* items, and there is a garden path effect in for these items. In contrast, no clear garden path effect was found for the deaf readers. Although there was a difference for the deaf readers at the noun between the *wh* and *if* conditions, this may be explained, not as a filled-gap effect, but as the pattern of elevated reaction times at content as opposed to function words, a pattern that is more exaggerated due to the overall slower reading times for the *wh* condition. In support of this interpretation, there was no significant difference at the verb for the hearing subjects (p = .609) but a near-significant difference for the deaf subjects (see above, p = .058).

6. Experiment 3

6.1 Previous research

Studies such as Frazier & Rayner (1982) have found evidence that hearing readers process according to the principle of Late Closure. Late Closure dictates that at a position that requires a choice between two alternative structures, the processor opts to attach incoming material to the phrase that has just been input, rather than to initiate a new phrase/sentence. This can lead to a garden path, when the correct analysis is to posit a new phrase/sentence, i.e. the correct analysis is Early Closure. The propensity for such error may be modulated by the length of time (number of words) that intervenes before it becomes apparent that the Late Closure analysis is incorrect, and a new phrase/sentence should have been posited. For example, in Frazier & Rainer's study, subjects were to better able to recover from a garden path of assuming that the noun phrase that follows the first verb is object of that verb in (6b) than in (6a), since the noun phrase is shorter and the misanalysis is likely to be less entrenched,

[3] In this experiment and Experiment 3 (below) t-tests were conducted only in the case of significant effects in the main ANOVA.

(6) a. Before the King rides his beautiful white horse is always groomed.
b. Before the King rides his horse is always groomed.

6.2 Hypotheses

For hearing individuals, we predicted processing difficulty in early closure sentences when the second verb is accessed (*is* in (6)), indicating that the preceding noun phrase is in fact not the object of the first verb. We predicted that for hearing individuals, as in the Frazier & Rayner study, the long versions of the early closure sentences would cause more processing difficulty than the short versions.

Deaf individuals have consistently been found to have shorter digit spans than hearing individuals (King & Quigley 1985). In view of this finding, we hypothesised that deaf subjects would experience greater processing difficulty than the hearing when trying to recover from an initial parsing decision which was incorrect.

6.3 Materials

The materials for this experiment were the same as those used by Frazier & Rayner (1982), with some lexical changes to accommodate Northern Irish English usage (for example, the word 'roommate' was replaced by 'flatmate'). In half the materials there was an initial subordinate clause followed by a main clause (as in 7); in the remaining half of the materials, there were three conjoined clauses (as in 8). Each sentence had four versions: Late closure long (LCL), early closure long (ECL), late closure short (LCS) and early closure short (ECS).

(7) LCL: Before the King rides his beautiful white horse it's always groomed.
ECL: Before the King rides his beautiful white horse is always groomed.
LCS: Before the King rides his horse it's always groomed.
ECS: Before the King rides his horse is always groomed.

(8) LCL: Anne was watching the people on the street she laughed and nobody knew why.
ECL: Anne was watching the people on the street were laughing and nobody knew why.
LCS: Anne was watching you she laughed and nobody knew why.
ECS: Anne was watching you were laughing and nobody knew why.

The materials consisted of sixteen sentences, half of which were of type (7) and half of type (8). Each subject saw only one version (Late Closure/Early Closure) of each sentence, for a total of four tokens of each condition (LCL, ECL, LCS, ECS). The sentences were ordered so that no two tokens of the same construction were ever presented consecutively.

6.4 Results

We averaged the reading time per word in each of three regions: prior to the syntactically ambiguous phrase (*his (beautiful white) horse* in 7), in the ambiguous phrase, and in the disambiguating phrase. Response times are reported separately for the adverbial materials (sentences of type 7) and the conjoined clause materials (sentences of type 8), since there was

a large difference between the sentence types.[4] Figures 7–10 below show mean reading times per region for each of the three regions, for each sentence type, for both the hearing and deaf subjects. One generalization that can be seen from an inspection of all four graphs is that for both the hearing and the deaf subjects, adverbial sentences are read slower than conjoined sentences.

Fig. 7 shows reading times for the late closure long items. On the adverbial items, reading times for the deaf subjects tend to decrease as they proceed through the sentence. Reading times for the hearing do not vary much in the three regions of these items. On the conjoined items, reading times for the deaf decrease as reading proceeds but less so than for the adverbial sentences. Reading times for the hearing subjects increase as they proceed through the sentence.

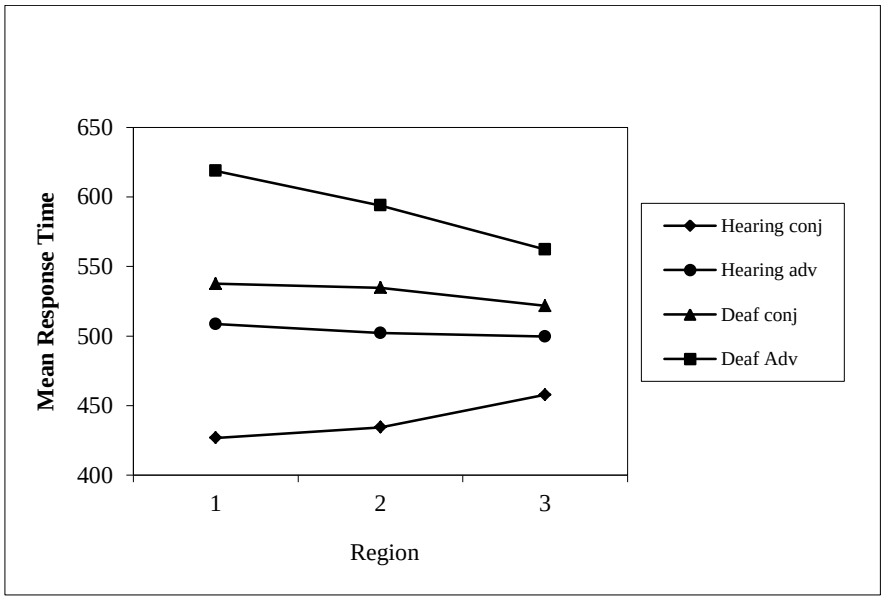


Figure 7: Mean reading times. *Late closure long by sentence type.*

[4] There were only two tokens each of the adverbial/conjoined structures per condition; the distinction between the structures was strong, nonetheless.

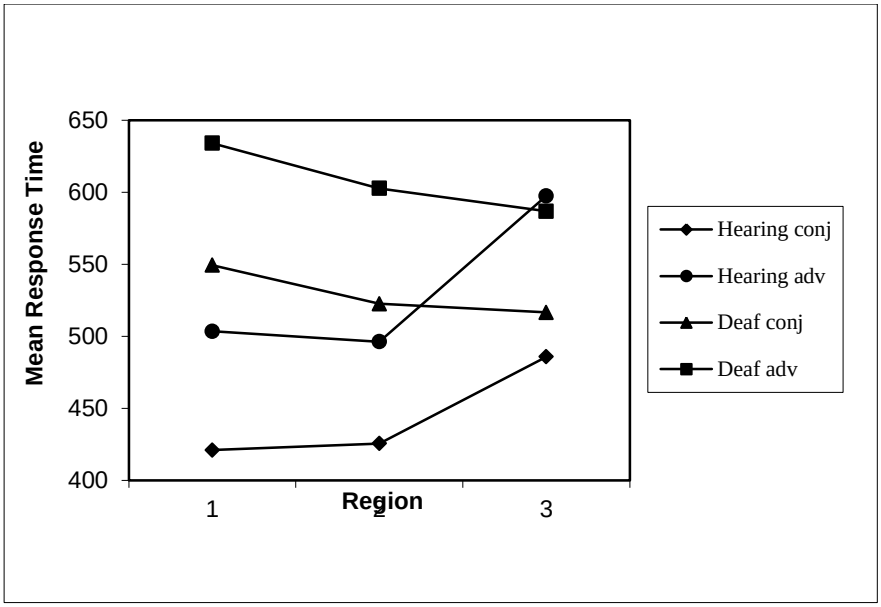


Figure 8: Mean reading times per word. Early closure long by sentence type.

Figure 8 shows the scores by region for the early closure long items x type. It is clear that the hearing subjects were garden-pathed in both the adverbial and conjoined items. However, the effect of the garden-path is greater for the adverbial items. The deaf were not garden-pathed in either the adverbial or conjoined items.

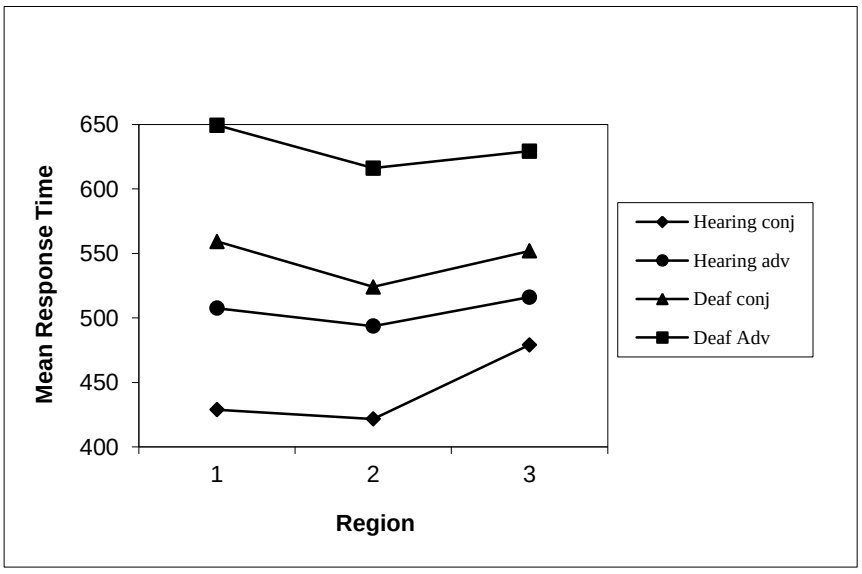


Figure 9: Mean reading times per word. Late closure short by sentence type.

Figure 9 shows the scores by region for the late closure short items x type. The pattern is similar across the three regions for both subject groups.

Figure 10 shows the scores by region for the early closure short items x type. It is clear that the hearing subjects were garden-pathed on both the adverbial and conjoined items.

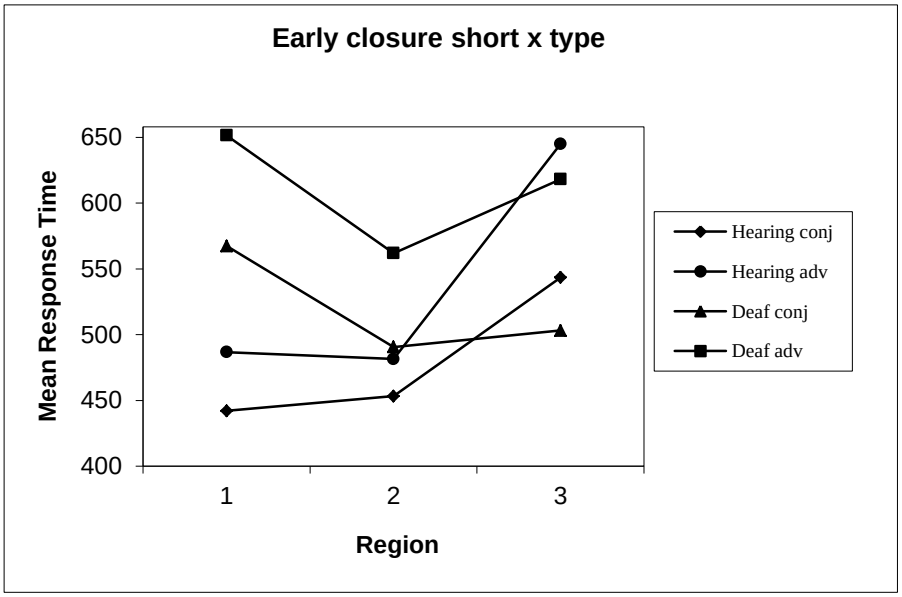


Figure 10: Mean reading times per word. Early closure short by sentence type.

ANOVAs were carried out for each group, with the factors 2 (sentence type) x 2 (closure) x 2 (length) x 3 (region). The ANOVA on the data from the hearing group revealed a significant effect of sentence type (F_1 (1,15) = 17.02, p = .001, F_2 (1,7) = 30.57, p = .001), an effect of closure for subjects, (F_1 (1,15) = 6.36, p = .023), but not for items, and a significant effect of region for both the subject and item analyses (F_1 (2,30) = 13.60, p = .002, F_2 (2,14) = 40.64, p = .001). The closure x region interaction was significant (F_1 (2,30) = 13.90, p = .001, F_2 (2,14) = 19.91, p = .001), as was the type x closure x region interaction (F_1 (2,30) = 4.36, p = .045, F_2 (2,14) = 5.09, p = .047). The length x region interaction was significant for subjects (F_1 (2,30) = 5.48, p = .012), but not for items.

Paired sample t tests revealed a significant difference between region 2 and region 3 of the late closure short conjoined sentences (t(15) = -2.76, p = .014) but not for the adverbial sentences. This indicates that the hearing subjects experienced greater processing difficulty in the disambiguating region of the late closure conjoined items than in the adverbial items.

The ANOVA for the deaf subjects revealed a significant effect of sentence type for subjects and items (F_1 (1,15) = 21.21, p = .001, F_2 (1,7) = 11.91, p = .011). There was a near significant effect of region for subjects and a significant effect for items (F_1 (2,30) = 3.61, p = .051, F_2 (2,14) = 8.69, p = .004). There was no effect of closure for either subjects or items. No other scores reached significance.

Paired sample t tests showed that the transition from region 1 to region 2 in the early closure short versions of both the adverbial (t(15) = 2.76, p = .015) and conjoined (t(15) = 2.69, p = .017) items was significant, but no significant difference was found between region 2 (the ambiguous region) and region 3 (the disambiguating region) in the early closure versions of the conjoined items (or between region 1 and region 3). There was a significant difference between region 2 and region 3 in the early closure short versions of the adverbial items (t(15) = -2.59, p = .020).

6.5 Discussion of Experiment 3

Our results for hearing subjects are different from those of Frazier and Rayner to the extent that in their study (a) no main effect of sentence type (adverbial vs. conjoined) and (b) greater garden path effects for the long early closure sentences were found. With respect to (b), Ferreira & Henderson (1991) argue that the resilience of a late closure misanalysis may depend not on the number of words intervening before the correct analysis is signalled, but on the structure of the intervening material. However, we used exactly the same materials as Frazier and Rayner, but still did not replicate their length effect. Possibly an explanation for both (a) and (b) lies in the difference in methodology between Frazier & Rainer's study, which used eye-tracking, and this one, which used self-paced reading.

Like the hearing subjects, deaf readers show longer readings times for adverbial clauses than for conjoined clauses, a finding that runs against the proposal by Miller (2000) that the deaf may treat subordinate clause markers as co-ordinating conjunctions. Despite this similarity, the basic result – a robust late closure effect – is found for hearing subjects, but not for deaf subjects.

7. Summary of the results of Experiments 1–3

Overall, the results of the three experiments we have reported show both similarities and differences in the reading patterns of deaf and hearing subjects. The similarities are:

- A pattern of longer reading times on content words than on function words in reading simple sentences (Experiments 1 for deaf readers and for hearing readers in other studies in the literature).[5]
- Longer reading times for *wh*-questions than for yes–no questions (Experiment 2)
- Longer reading times for preposed adverbial clauses than for conjoined clauses (Experiment 3).

The differences are:

- Little evidence of an end of sentence wrap up effect for deaf readers (Experiment 1)
- An absence of a filled-gap effect for deaf readers (Experiment 2)
- An absence of Late Closure garden-pathing for deaf readers (Experiment 3).

8. General discussion

8.1 Syntactic processing by deaf readers

We believe that the best explanation of the data from our experiments is that the deaf participants are not processing syntactically in the same manner as hearing participants. This is not to say that no syntactic analysis is taking place, but rather that, as proposed by Miller (2000), processing may be different in the deaf population. In the case of our experiments, the data indicates that the deaf are not processing in a manner that is guided by the processing principles that lead to the filled gap effect for hearing subjects in Experiment 2 and difficulty with early closure sentences in Experiment 3. The view that processing by the deaf is not syntactic in the same way as it is for hearing individuals is bolstered by the fact that an end of sentence wrap up effect is largely absent for the deaf hearers.

[5] As suggested in the discussion of Experiment 1, the lack of a distinction between content and function words for hearing readers in that experiment may reflect the choice of a subordinate task.

If our deaf subjects are not processing syntactically, then we need to explain the similarities between deaf and hearing subjects: the rise at content words and fall at function words pattern produced by deaf readers in this study and hearing readers in other studies, the longer reaction times for *wh* as compared to *if* questions in this study, and the longer reaction times for adverbial as compared to conjoined clauses in this study. Also, how can we account for the fact that comprehension by our deaf subjects on the secondary yes–no question task was at such a high level (90%)?

The first of the similarities in processing indicates that lexical storage and access is similarly organized for deaf and hearing readers. The second similarity requires that we acknowledge that processing a *wh*-question entails for the deaf, as for the hearing, a working memory burden due to storage of the *wh*-word. The third similarity may arise because the adverbial subordinating conjunction/preposition initiates a semantic analysis that requires a more complex mapping between the main and subordinate clause than is required in the case of the mapping between conjoined clauses. As stated above, what our deaf subjects appear to lack are syntactic strategies that in the case of *wh*-sentences act to minimize memory load, and in the case of early closure sentences to resolve a syntactic ambiguity.

So then why are the deaf so competent when it comes to answering the yes–no questions in the subordinate task? Recent literature on hearing individuals has argued that in certain circumstances the sentence processor may rely on a 'shallow' analysis – i.e. on an analysis that is not fully specified with respect to syntactic structure and the reference of, for example, pronouns (Townsend & Bever, 2000; Sanford & Sturt 2002; Clahsen & Felser 2006). On examination, the yes/no questions used in the secondary task could have been answered without an understanding of the details of the syntax of the sentence. For example, the question for sentence (9) below, 'Were the customers happy?' could be answered by reference to semantic information in the sentence i.e. 'the angry customers' and 'everyone seemed extremely upset'.

(9) The angry customers were cursing the manager all the waiters were running back and forth and everyone seemed extremely upset.

Even if the deaf subjects in this study were not processing syntactically in the same way as hearing readers, they would still have been able to answer the yes/no questions by relying on semantic information. Since the comprehension questions were used to ensure that subjects were reading for meaning and that they attended to the primary task without distraction, the questions used were adequate for the task. We cannot, however, take the results of the question task to reflect that the deaf subjects correctly analyzed details of the syntactic structure in all of the test sentences.

We wish to emphasise that we are not claiming that deaf readers do not process syntactically, but rather that they are not processing in the same way as hearing readers. It may be that the deaf are more heavily influenced by factors that have been shown to play a role in processing by hearing readers. Ferreira, Christianson & Hollingworth (2001) found that hearing subjects who read sentences such as (10) were often satisfied with an incorrect interpretation, thinking that Anna dressed the baby,

(10) While Anna dressed the baby spit up on the bed.

The longer the ambiguous phrase, the less likely readers were to achieve successful re-analysis. Ferreira et al. suggest that the incorrect interpretation based on the initial misparse

lingers and is more likely to do so in sentences with a long ambiguous phrase because the parser has been committed to the incorrect syntactic structure for longer. They also point out that misinterpretation is somewhat selective and does not hold of the interpretation is implausible. If, as Ferreira et al. claim, individuals who are aware of ambiguity and who perform reanalysis do not end up with a perfect interpretation for early closure sentences but accept 'good enough representations', then perhaps the deaf generally interpret sentences using a 'good enough' strategy, i.e. keep processing until you think you understand the intended meaning. This fits well with the results of Miller (2000), who found the deaf were more reliant on semantic cues in comprehending sentences than the hearing, and that the oral deaf were more so than signing deaf subjects.

8.2 Why do the deaf not develop the same parsing strategies as the hearing?

Although there is now a large body of work documenting research on phonological coding skills in the deaf population (Paul 1998; Perfetti & Sandak 2000; Izzo 2002; Miller 2002; Burkholder & Pisoni 2003, among others), no clear consensus has emerged on the degree to which ability in this area can account for lack of reading success in the deaf. Nonetheless, it has been suggested that a rich oral language is necessary for reading skills to develop successfully (Clay 1985).

Inefficient phonological coding has been argued to be at the core of reading failure in hearing readers (Ehri 1999). Since phonological skills are based on sound, the hearing child is able to internalize phonological knowledge as a matter of course, but this is not the case for the deaf child. It is well documented that the phonological ability of deaf individuals is poor (Conrad 1979; Hanson 1989; Lillo-Martin, Hanson & Smith 1991). Some evidence suggests that deaf readers with a high level of skill do use phonological encoding (Bebko 1998; Perfetti & Sandak 2000; Miller 2002). However, Miller (2002) found that although deaf readers who were raised orally did process phonologically, they did not have the same level of skill as their hearing counterparts. Other studies have shown that although some deaf readers use phonological coding, this is not the case with all deaf readers (Conrad 1973, 1972; Lillo-Martin et al. 1991).

Prosody in speech and punctuation in written text can prevent garden-pathing in some cases. Bader (1998: 1) claims that 'for certain kinds of syntactic ambiguities re-analysis is prosodically constrained'. He argues that this explains why some garden path sentences are easier to interpret than others. Consider the following pairs of sentences (11) and (12), in which the reader will be garden-pathed in the (b) versions. Bader suggests that (11b) produces a greater garden path than (12b) because in (11b) the prosodic structure has to be revised as well as the syntactic structure whereas in (12b) the prosodic structure can be left intact (see also Kjelgaard & Speer 1999, for a study of the role of prosody in late/early closure structures, and Frazier, Carlson & Clifton 2006, for a review of the importance of prosody for language comprehension).

(11) a. In order to help the little boy Jill put down the package she was carrying
b. In order to help the little boy put down the package he was carrying

(12) a. Peter knew the answer immediately
b. Peter knew the answer would be false.

Given that the deaf participants in this study did not exhibit garden-path effects, we consider it possible that lack of prosody during language acquisition may have resulted in atypical

development. Prosodic features such as stress, pitch and length are an integral part of speech and the hearing child benefits from such prosodic information as a matter of course. The relation between use of prosody and the development of particular reading patterns in the deaf can only be an indirect one. Prosodic information can resolve Early/Late Closure ambiguities, but not examples such as (12) or the garden path of wrongly anticipating a position for a question word (experiment 2). In the course of development the deaf reader may come to focus on lexical information as a safe option for how to figure out the meaning of a sentence, leading to strategies that can result in typical garden paths being side-stepped.[6]

9. Conclusion

The absence of increased reading times associated with syntactic processing (the end of sentence wrap up effect, filled gap effect and the late closure effect) in the data from the deaf readers in each of our three experiments indicates that they were not processing the materials in the same way as the hearing participants. We suggest that lack of access to prosodic information in the input during language acquisition may result in deaf children developing atypical processing strategies. We propose that the longer reading times produced by both the deaf and hearing participants in sentences with embedded *wh*-questions in experiment 2 and in the adverbial sentences in experiment 3 can be attributed to lexical processing and/or an increased burden on memory.

Acknowledgements

We are grateful to two reviewers for YPL for their detailed and constructive comments.

References

AARONSON, D. & SCARBOROUGH, H.S. 1976. Performance theories for sentence coding: Some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance* 2, 56-70.

BADER, M. 1998. Prosodic influences on reading syntactically ambiguous sentences". In Fodor J D & Ferreira F (Eds.) (1998) *Reanalysis in Sentence Processing,* 1-47. Dordrecht. Boston. London. Kluwer Academic Publishers.

BEBKO, J. 1998. Learning, language, memory and reading: The role of language automatization and its impact on complex cognitive activities. *Journal of Deaf Studies and Deaf Education*, 3, 4-13.

BERENT, G. 1988. An assessment of syntactic capabilities. In: Strong, M. (ed.) *Language Learning and Deafness*, 133-161.

BURKHOLDER R. & PISONI, D. 2003. Speech timing and working memory in profoundly deaf children after cochlear implantation. *Journal of Experimental Child Psychology,* 85, 63-88.

CLAHSEN, H. & FELSER, C. 2006. How native-like is non-native language processing? *TRENDS in Cognitive Sciences*, 10, 564-570.

CLAY, M. 1985. *The Early Detection of Reading Difficulties.* Portsmouth, N.H.: Heinemann.

[6] As a reviewer notes, sign language 'probably conveys some kind of prosody', and it would be valuable to compare persons educated in the oral system and who learnt sign language as a second language (as in this study), to subjects who were native signers.

CONRAD, R. 1972. Short term memory and the deaf: A test for speech coding. *British Journal of Psychology* 63, 173-180.

CONRAD, R. 1973. Some correlates of speech coding in short-term memory for the deaf. *Journal of Speech and Hearing Research* 16, 375-384.

CONRAD, R. 1977. The reading ability of deaf school leavers. *British Journal of Educational Psychology*, 47, 138-148.

CONRAD, R. 1979. *The Deaf School Child*. London, UK: Harper & Row.

CRAIN, S. & FODOR, J.D. 1985. How can grammars help parsers? In: Dowty D. R., Karttunen L. & Zwicky A. M. (eds.) *Natural Language Parsing*, Cambridge, England: Cambridge University Press, 94-127.

EHRI, L. 1999. Phases of development in learning to read words. In: Oakhill J. & Beard R. (eds.) *Reading Development and the Teaching of Reading: A Psychological Perspective.* Oxford, U.K.: Blackwell, 79-108.

FERREIRA, F., BAILEY, K. & FERRARO, V. 2002. Good-enough representations in language comprehension". *Current Directions in Psychological Science* 11, (1) 1-15.

FERREIRA, F., CHRISTIANSON, K. & HOLLINGWORTH, A. 2001. Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research* 30, (1) 3-20.

FERREIRA, F. & HENDERSON, J.M. 1991. Recovery from misanalysis of garden-path sentences. *Journal of Memory and Language* 30, 725-745.

FRAZIER, L., CARLSON, K. & CLIFTON, C. 2006. Prosodic phrasing is central to language comprehension. *TRENDS in Cognitive Sciences* 10 (6) 244-249.

FRAZIER, L. & CLIFTON, C. 1989. Successive cyclicity in the grammar and parser. *Language and Cognitive Processes 4* (2) 93 – 126.

FRAZIER, L. & RAYNER, K. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14, 178-210.

HANSON, V. 1989. Phonology and reading: Evidence from profoundly deaf readers. In: Shankweiler D. & Liberman I. (eds.) *Phonology and Reading Disability: Solving the Reading Puzzle.* Ann Arbor, MI: University of Michigan Press, 69-89.

HATCHER, C. & ROBBINS, N. 1985. The development of reading skills in hearing impaired children. In: King, C. & Quigley, S. (eds.) *Reading and Deafness*. London, U.K.: Taylor and Francis.

IZZO, A. 2002. Phonemic awareness and reading ability: An investigation with young readers who are deaf. *American Annals of the Deaf* 147, 18-28.

JUST, M.A. AND CARPENTER, P.A. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review* 87, 329-354.

KELLY, L. 1995. Processing of bottom-up and top-down information by skilled and average deaf readers and implications for whole language instruction. *Exceptional Children* 61 (4), 318-334.

KELLY, L.. 2003. Considerations for designing practice for deaf readers. *Journal of Deaf Studies and Deaf Education* 8, 230-249.

KING, C.M. & QUIGLEY, S.P. 1985. *Reading and Deafness* London. Taylor and Francis.

KJELGAARD, M.M. & SPEER, S.R. 1999. Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language*, 40, 153-194.

LILLO-MARTIN, D., HANSON, V. & SMITH, S. 1991. Deaf readers' comprehension of complex syntactic structure. *Advances in Cognition, Education and Deafness.*Washington, DC: Gallaudet University Press.

LILLO-MARTIN, D., HANSON, V. & SMITH, S. 1992. Deaf readers' comprehension of relative clause structures. *Applied Psycholinguistics* 13, 3-30.

MARSCHARK, M. 1997. *Raising and Educating a Deaf Child.* New York: OUP.

MARSCHARK, M. & SPENCER, E.P. (eds.) 2003. *Oxford Handbook of Language and Education.* OUP.

MILLER, P.F. 2000. Syntactic and semantic processing in Hebrew readers with prelingual deafness. *American Annals of the Deaf* 145, 436-448.

MILLER, P.F. 2002. Communication mode and the processing of printed words: Evidence from readers with prelingually acquired deafness. *Journal of Deaf Studies and Deaf Education* 7, 312-329.

MILLIS, K.K. & JUST, M.A. 1994. The influence of connectives on sentence comprehension. *Journal of Memory and Language* 33, 128-147.

MITCHELL, D.C. & GREEN, D.W. 1978. The effects of context and content on immediate processing in reading. *Quarterly Journal of Experimental Psychology*, 30, 609-636.

PAUL, P. 1998. *Literacy and Deafness: The Development of Reading, Writing and Literate Thought.* Needham Heights, MA: Allyn & Bacon.

PAUL, P.. 2003. Processes and components of reading. In: Marschark M. & Spencer E. (eds.) *Oxford Handbook of Deaf Studies, Language and Education.* Oxford, U.K.: Oxford University Press.

PERFETTI, C. & SANDAK, R. 2000. Reading optimally builds on spoken language: Implications for deaf readers. *Journal of Deaf Studies and Deaf Education* 5, 32-50.

QUIGLEY, S., POWER, D. & STEINKAMP, M. 1977. The language structure of deaf children. *Volta Review*, 79-84.

QUIGLEY, S. 1982. Reading achievement and special reading materials. *Volta Review* 84, 95-106.

RAYNER, K., SERENO, S.C., MORRIS, R.K., SCHMAUDER, A.R. & CLIFTON, C. JRN 1989. Eye movements and on-line language comprehension processes. *Language and Cognitive Processes* 4, 21-49.

SANFORD, A. & STURT, P. 2002. Depth of processing in syntactic comprehension: Not noticing the evidence. *TRENDS in Cognitive Sciences*, 6, 382-386.

SNYDER, L. & DOWNEY, D. 1991. The language-reading relationship in normal and reading disabled children. *Journal of Speech and Hearing Research* 34, 129-140.

STOWE, L.A. 1986. Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes* 1, 227-245.

STOWE, L.A., TANENHAUS, M.K. & CARLSON, G.N. 1991. Filling gaps on-line: Use of lexical and semantic information in sentence processing. *Language and Speech* 34, 319-340.

TOWNSEND, D.J. & BEVER, T.G. 2001. *Sentence comprehension.* Cambridge, Mass. London. MIT Press.

TRAXLER, C. 2000. The Stanford Achievement Test, 9th Edition: National norming and performance standards for deaf and hard of hearing students. *Journal of Deaf Studies and Deaf Education* 5, 337-348.

TRAXLER, M., CORINA, D. & MORFORD, J. 2010. Deaf signers reading written English sentences: Syntax and age-of-acquisition affect reading times. Poster presented at the CUNY sentence processing conference, New York University.

VAN HOOGMOED, A., VERHOEVEN, L. SCHREUDER, R. & KNOORS, H. 2011. Morphological sensitivity in deaf readers of Dutch. *Applied Psycholinguistics* 32, 619-634.

WATERS, G. & DOEHRING, D. 1990. Reading acquisition in congenitally deaf children who communicate orally: Insights from an analysis of component reading, language and memory skills. In: Carr T & Levy B (eds.) *Reading and its Development: Component Skills Approaches*. San Diego, CA: Academic Press, 323-368.

WAUTERS, L., TELLINDS, A. VAN BON, W. & MAK, M. 2007. Mode of acquisition as a factor in deaf children's reading comprehension. *Journal of Deaf Studies and Deaf Education* 13, 175-192.

Linda Coulter
29E Ballygelagh Rd
Ardkeen
BT22 1JG
Northern Ireland
email: linda.coulter@tiscali.co.uk

Helen Goodluck
Department of Language and Linguistic Science
University of York
Heslington
York
YO10 5DD
England
email: helen.goodluck@york.ac.uk

REFERENCE SAMPLE SIZE AND THE COMPUTATION OF NUMERICAL LIKELIHOOD RATIOS USING ARTICULATION RATE

VINCENT HUGHES[1], ASHLEY BRERETON[2] & ERICA GOLD[1]

[1]University of York

[2]University of Liverpool

Abstract

This paper explores the effects of variability in the amount of reference data used in quantifying the strength of speech evidence using numerical likelihood ratios (LRs). Monte Carlo simulations (MCS) are performed to generate synthetic data from a sample of existing raw local articulation rate (AR) data. LRs are computed as the number of reference speakers (up to 1000), and the number of tokens per reference speaker (up to 200) is systematically increased. The distributions of same-speaker and different-speaker LRs and system performance (log LR cost (C_{llr}) and equal error rate (EER)) are assessed as a function of the size of the reference data. Results reveal that LRs based on AR are relatively robust to small reference samples, but that system calibration plays an important role in determining the sensitivity of the LRs to sample size.

1. Introduction

Forensic voice comparison (FVC) commonly involves the analysis of the speech patterns in a recording of an unknown criminal's voice and a recording of a known suspect's voice with regard to the competing propositions of the prosecution (same-speaker) and defence (different-speakers). Across forensic disciplines the likelihood ratio (LR) is increasingly accepted as the "logically and legally correct" (Rose & Morrison 2009: 143) framework for the gradient assessment of such comparison evidence (Robertson & Vignaux 1995, Aitken & Taroni 2004). The outcome is gradient in that it not only indicates whether the evidence supports the prosecution or defence, but also provides an estimate of the relative strength of the support. The LR involves an assessment of both the similarity between the evidential samples and the typicality of within- and between-speaker variation in the relevant population (Aitken & Taroni 2004). The assessment of typicality is an essential element of the LR framework, since it allows the expert to estimate the probability of finding the speech evidence assuming the offender is another random member of the relevant population (for issues relating to the definition of the relevant population in speech see Hughes & Foulkes 2012, Morrison et al. 2012). Quantification of typicality is conducted using a set of representative reference data.

However, a practical concern in FVC is the amount of reference data needed to ensure a meaningful estimate of the strength of evidence. Consistent with general sampling principles, the answer is determined by "how accurate the experimenter wishes the estimate to be" (Wackerly *et al.* 2008: 421). That is, the greater the amount of representative data, the more precise the model of the population (Rose 2012). A small number of studies have investigated this issue for FVC. Ishihara and Kinoshita (2008) describe a *population size effect* when computing LRs based on distributional characteristics of fundamental frequency (f0) using small numbers of reference speakers. In their study, same-speaker (SS) and different-speaker (DS) LRs were overestimated by up to 1000 times using 10 reference speakers compared with a baseline using 120 speakers. Hughes and Foulkes (2012) investigated the effect of

different numbers of reference speakers (up to 120) and numbers of tokens per reference speaker (up to 13) on the distribution of LR scores and system performance (i.e. how well the system separates SS pairs from DS pairs) using polynomial representations of formant trajectories from spontaneous GOOSE (/uː/) vowels. Consistent with Ishihara and Kinoshita (2008), scores were found to be more stable and system performance improved with greater than 20 reference speakers. Further, DS LR scores were shown to be very sensitive to the number of tokens per reference speaker, displaying no stability between the 2- and 13-token conditions. Same-speaker pairs were more robust to variation in the amount of data per reference speaker.

The findings of these studies suggest that LRs are generally unstable and misrepresentative when small numbers of reference speakers and small numbers of tokens are used. The results reflect an imprecise estimation of the variation in the population when using small amounts of reference data. That is, the addition of speakers or tokens to a small sample affects the distributions of within- and between-speaker values more than the addition of speakers or tokens to a much larger set of existing data. Consistent with the law of diminishing returns, with a given amount of data the addition of more representative data will have little effect on the overall distribution. Very little work in FVC has considered such an upper limit at which the inclusion of more reference data has an asymptotic effect on LRs and system performance. Yet, the efficiency and cost-effectiveness of the numerical LR approach is, at least to some extent, dependent on knowing how much reference data is enough to produce robust estimates of the strength of evidence.

The relative lack of research in this area is in part due to a lack of sufficiently large amounts of raw data. Monte Carlo simulations (MCS) offer a potential solution to this problem. They involve generating synthetic values from known properties of the distributions of within- and between-speaker variation of a given variable in a given population. Synthetic data can be built from population statistics presented in previous research (e.g. mean and standard deviation (SD) when the distribution is assumed to be normal, although the assumption of normality is not a pre-requisite; as in Rose 2012) or using some existing set of raw data. Whilst Monte Carlo simulations avoid the need for extremely large amounts of raw data, there is a non-trivial *a priori* assumption that the true distribution of the variable in the population is known (or can be well estimated). This is because the distribution of the resulting synthetic data is defined by the properties of the input. In this respect, Monte Carlo simulations are not predictive. Despite having knowledge of the distribution of a variable in a population, MCS are still necessary in investigating how LR performance is affected by sample size since the Multivariate Kernel Density (MVKD) procedure proposed by Aitken & Lucy (2004) includes N speakers as part of the kernel density estimation and N tokens per speaker in determining suspect and offender variance and covariance matrices (Aitken & Lucy 2004: 12-13, Rose 2012, 2013: 94).

Initial exploration of Monte Carlo methods for FVC is offered by Rose (2012), who synthesised F1, F2 and F3 midpoint values for Australian English /aː/ for up to 10,000 speakers based on the distribution of values in Bernard (1967). Using Lindley's (1977) procedure to investigate individual (univariate) formant performance and the MVKD (Aitken & Lucy 2004) formula to investigate combined formant performance, SS LRs were computed based on real case data and assessed as a function of the number of reference speakers. Output was compared against the *true* LR, which is defined as the LR computed using the maximum amount of reference data. As such, the *true* LR is based on the most precise estimation of within- and between-speaker variation in the relevant population. The magnitude of the LR was found to be equivalent to the *true* LR (based on 10,000 speakers) with the inclusion of 30 or more reference speakers. However, Rose's study is limited by the

lack of DS comparisons, since there is no assessment of performance as a function of sample size or calibration based on weights from a development set. The procedures for modelling potential correlations between formants are also not made explicit.

The present study builds on Rose (2012) by using Monte Carlo simulations to investigate the effect of reference sample size on the outcome of numerical LRs based on an analysis of local articulation rate (AR). A set of existing raw data of 59 speakers is firstly analysed with regard to how precisely it estimates patterns in the relevant population, defined in terms of regional background, age, sex and class. MCS are then used to generate normal distributions for 941 synthetic speakers from which up to 200 tokens per speaker are generated. LRs are computed for development (20 speakers) and test (20 speakers) sets extracted from the raw data and the distributions of LRs and system validity are assessed as a function of sample size. Following Rose (2012), results are compared against the *true* LR performance based on the largest set of reference data (up to 1000 speakers, up to 200 tokens per speaker).

2. Methods

2.1. Data

The data consisted of local articulation rate (AR) measurements, quantified as the number of phonological syllables per second within multiple memory stretches (Jessen 2007). AR was chosen primarily because it is a simple, univariate variable, which can be synthesised relatively straightforwardly. However, the extent to which the findings based on AR can be generalised to other variables is limited by the lack of inherent speaker-discriminant value. This is confirmed by a small variance ratio (calculated as the between-speaker SD^2 divided by the within-speaker $mean^2$; Rose *et al.* 2006) of 0.326, which suggests that within-speaker variability in AR is generally higher than between-speaker variation. Whilst a multivariate variable with good speaker discriminatory power would be preferable for investigating sample size, unidimensionality makes AR a good candidate for the preliminary exploration of Monte Carlo methods for speech data. Local AR was chosen over global AR (across a recording) since multiple tokens are needed to estimate within-speaker variation when computing LRs. Further, local AR is a more meaningful forensic resource since it captures individual variability across utterances (Miller *et al.* 1984).

The data were collected as part of Gold (in progress). Local AR was extracted for 100 speakers from Task 2 recordings of the DyViS database (Nolan *et al.* 2009). The DyViS speakers are all young (aged 18-25), male speakers of Standard Southern British English (SSBE) from the University of Cambridge and as such are considered sociolinguistically homogeneous. Task 2 involves a telephone conversation with a mock accomplice, who is demographically matched to the subjects. Since raw AR values are extracted from a single session, the data necessarily provide an under optimistic estimation of the within-speaker variation found in real casework (Rose 2011; Morrison *et al.* 2012). It is expected that LR performance will be poorer in realistic, non-contemporaneous conditions.

Gold (in progress) identified the onset and offset of between 26 and 32 memory stretches per speaker, defined as a period of "fluent speech containing a number of syllables that can easily be retained in short-term memory" (Jessen 2007: 54). Memory stretches were chosen as a unit for measuring AR based on Gold (in progress), who found no significant differences in performance compared with inter-pause stretches. Further, Gold (in progress) suggests that memory stretches are better for FVC as they can be extracted efficiently without requiring

precise measurement of individual pauses. Following Künzel (1997), "fluent speech" was defined as the absence of pauses, hesitation phenomena and repair processes. Each token was then calculated as the total number of phonological syllables divided by the total duration (in seconds) of the memory stretch. Memory stretches generally contained between 7 and 11 syllables.

For each speaker, the first 26 tokens were used in the analysis (the largest number of tokens shared by all speakers). Mean and SD of AR values were calculated by-speaker and converted to z-scores to identify univariate outliers. On the basis of an outlying SD with z > $\pm$3.29 (p < 0.01) (Tabachnick & Fidell 2007: 73), one speaker was removed from the analysis. With this speaker removed z-scores were again calculated, but no values exceeded the $\pm$3.29 threshold. Of the remaining 99 speakers, 20 were selected at random as development data and a further 20 as test data. The remaining 59 speakers were used as reference data from which synthetic reference speakers and tokens were generated.

2.2. Modelling

In this study, Aitken and Lucy's (2004) MVKD formula was used to compute LRs, which models within-speaker variance with a normal distribution and the between-speaker distribution with a kernel density made up of Gaussians from each reference speaker (Morrison 2011a: 243). Using the modelling procedures in the MVKD formula as a starting point, a two-stage process for synthesising data was developed. Firstly, normal distributions for each synthetic speaker were generated by sampling synthetic means and SDs from the raw data. This is because MVKD models within-speaker variation with an assumption of normality. From the synthetic normal distributions $N(\mu, \sigma)$ a second round of simulations were conducted to generate synthetic tokens for each of the synthetic speakers. Prior to conducting MCS, two issues with the raw data were addressed. The first relates to the choice of distribution from which synthetic mean and SD values are sampled. Figures 1 displays the histograms of raw mean and SD values by-speaker fitted with normal distributions.

Skew and kurtosis were calculated to assess how well normality models the data. Following Tabachnick and Fidell (2007: 79), skew was analysed by dividing the skewness (S) by the standard error (S_s), defined as $S_s = \sqrt{6/N}$ where N is the number of observations (in this case 59), to give a z-score (where z > $\pm$1.96 = p < 0.05 and z > $\pm$3.29 = p < 0.01). Skew was non-significant for both means and SDs (p > 0.4). Kurtosis was analysed by dividing the kurtosis statistic by twice the standard error (Tabachnick & Fidell 2007) to generate a z-score. For both sets of data, kurtosis was also found to be non-significant (p > 0.24). Given the statistical assessment of normality and visual inspection of Figures 1 and 2, the normal distribution fits the data sufficiently well.

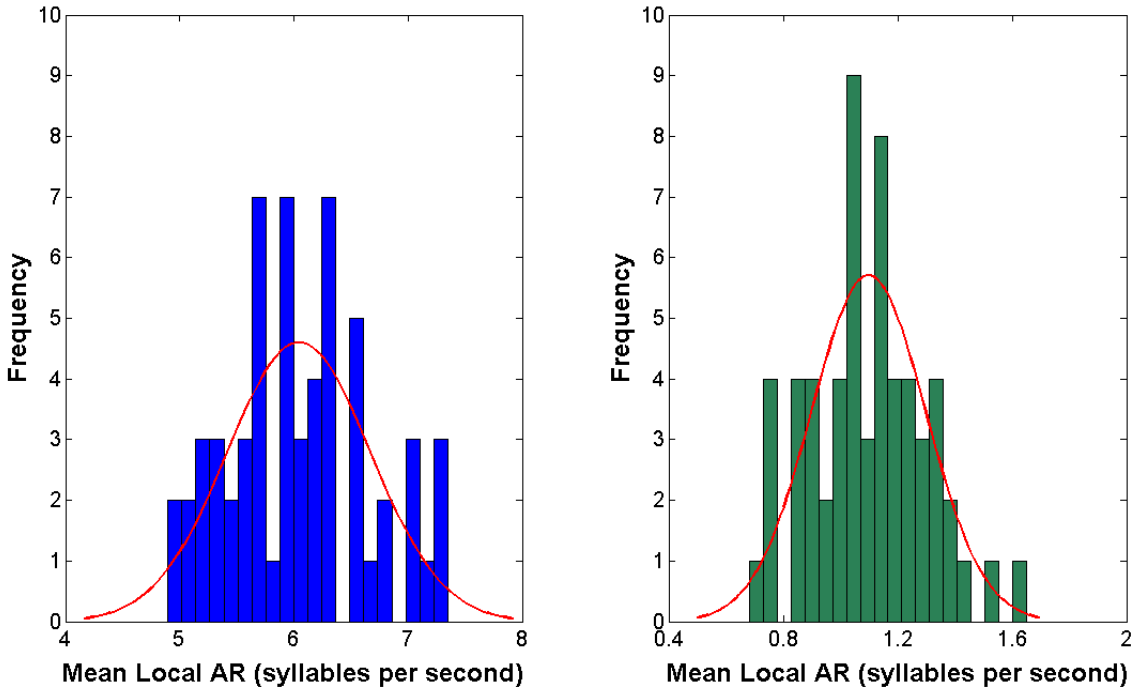


Figure 1: Histograms of mean (left) and SD (right) of AR by speaker fitted with normal distributions

The second issue relating to the raw data is whether the 59 raw speakers provide a sufficiently precise estimate of patterns in the relevant population. To assess how well the sample of raw data approximates the distribution of values in the relevant population, it is necessary to know *a priori* the shape of that distribution. In the absence of this knowledge, the alternative is to assess how the distributions of means and SDs vary as data is added. Independent samples t-tests were calculated for means and SDs using the values for all 59 speakers compared against values for 10 speakers. Values for each speaker were then added consecutively to the smaller data set and the t-test re-run. Welch's Correction (Welch 1947) was applied to account for unequal sample size and an assumption of unequal variance across the sets of data. The results are analysed with regard to the *p*-value where $p = 1$ is equivalent to the two samples having the same normal distribution.

Figure 2 shows that there is no significant difference for AR means in the distribution of values for as few as 10 speakers compared with the distribution using all 59 speakers. Despite an initial dip with small numbers of speakers, *p* increases towards 1 after 25 speakers. For SDs, *p* is relatively low (0.1) with small numbers of speakers, although at no point is the difference between the distributions significant (at either the 1% or 5% levels). Whilst predictably *p* approaches 1 as the number of speakers increases, there is considerable similarity in the distribution of SDs after 40 speakers. Further, the means and SD are consistent with expectations about the range of potential within- and between-speaker variation reported in Goldman-Eisler (1968). As such it is considered that the distributions based on 59 speakers provide a sufficiently precise estimate of the relevant population.

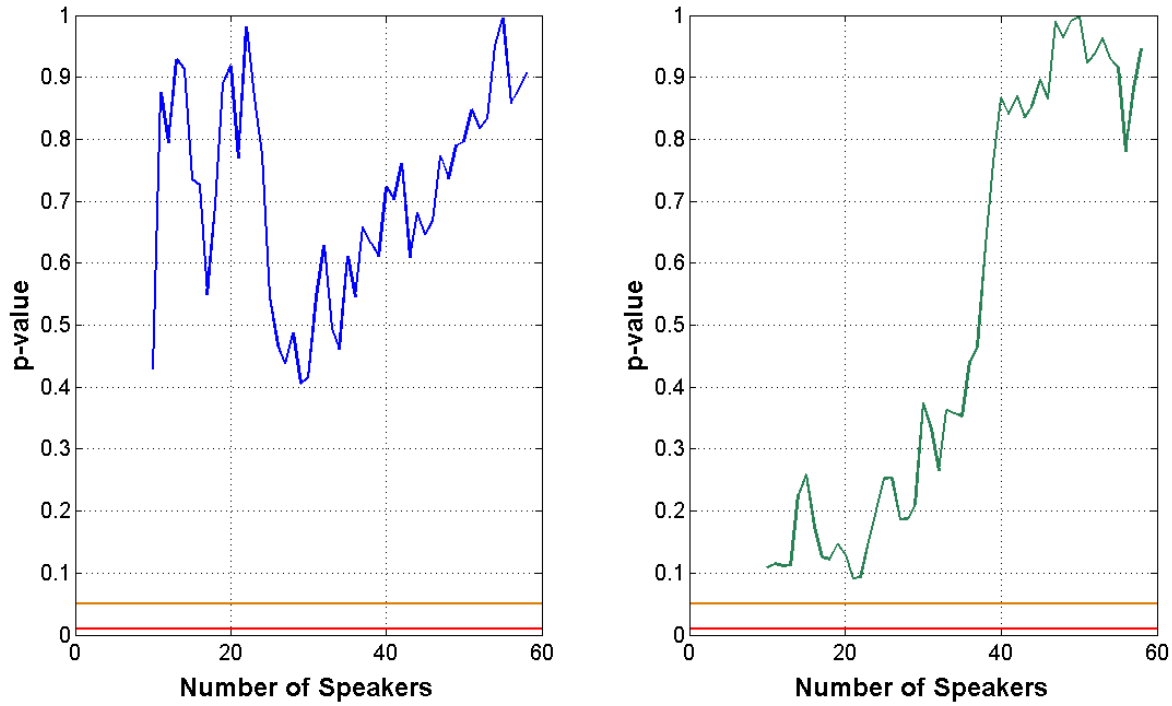


Figure 2: p-values based on t-tests comparing the distributions of means (left) and SDs (right) for the number of speakers on the x axis against that with all 59 raw speakers with 1% (red) and 5% (orange) significance marked

2.3. Monte Carlo simulations (MCS)

The following sections explain the procedures used for creating synthetic tokens of local AR for synthetic speakers based on the set of raw data.

2.3.1. Generating synthetic mean local AR

Mean local AR is denoted by *x*, where x_i is a value for a single speaker ($_i$ is speaker number). Based on the testing of normality in §2.2, the distribution of raw x_i values is converted to a normal probability density function (PDF) with mean of 0 and SD of $\frac{1}{\sqrt{2}}$, $N(0, \frac{1}{2})$, by applying the transformation:

$$z = \frac{(x - \mu_x)}{\sqrt{2}\sigma_x}, \tag{1}$$

where μ_x is the mean of the raw means and σ_x is the SD. This transforms values within the raw *x*-space to normalised values within the *z*-space where the aim of the Monte Carlo simulations is to generate synthetic z_i values from the preferentially scaled PDF. This is done using the inverse of the cumulative distribution function (CDF). The CDF uses integration to calculate the area under the PDF between $-\infty$ and z_i such that:

$$CDF(z) = \int_{-\infty}^{z} N(z, 0, \frac{1}{2})\, dz. \tag{2}$$

Given that the normal distribution is so widely used, a special function called the *error function* (erf) (Wang *et al.* 1989: 333) has been assigned to the integral ($\int$) meaning that it is possible to generate a CDF based on a normal PDF in the following way:

$$\int_{-\infty}^{z} N\left(z, 0, \frac{1}{2}\right) dz = CDF(z) = \frac{1 + \text{erf}(z)}{2} \tag{3}$$

where:

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-t^2} dt \tag{4}$$

With the CDF defined as above, normally distributed z_i values can be synthesised using a random variable $Z_i \in [0, 1]$ (i.e. a random number between 0 and 1). Using the inverse CDF ($CDF^{-1}(z)$), a single synthetically generated z_i value is defined as:

$$CDF\left(CDF^{-1}(z)\right) = z = \frac{1 + erf\left(CDF^{-1}(z)\right)}{2} \tag{5}$$

$$\therefore \quad 2z - 1 = erf(CDF^{-1}(z)) \tag{6}$$

$$CDF^{-1}(z) = erf^{-1}(2z - 1). \tag{7}$$

As demonstrated in Figure 3, using a random value for Z_i and with explicit knowledge of the inverse CDF, a synthetic z_i can be generated in the following way:

$$CDF^{-1}(Z_i) = z_i. \tag{8}$$

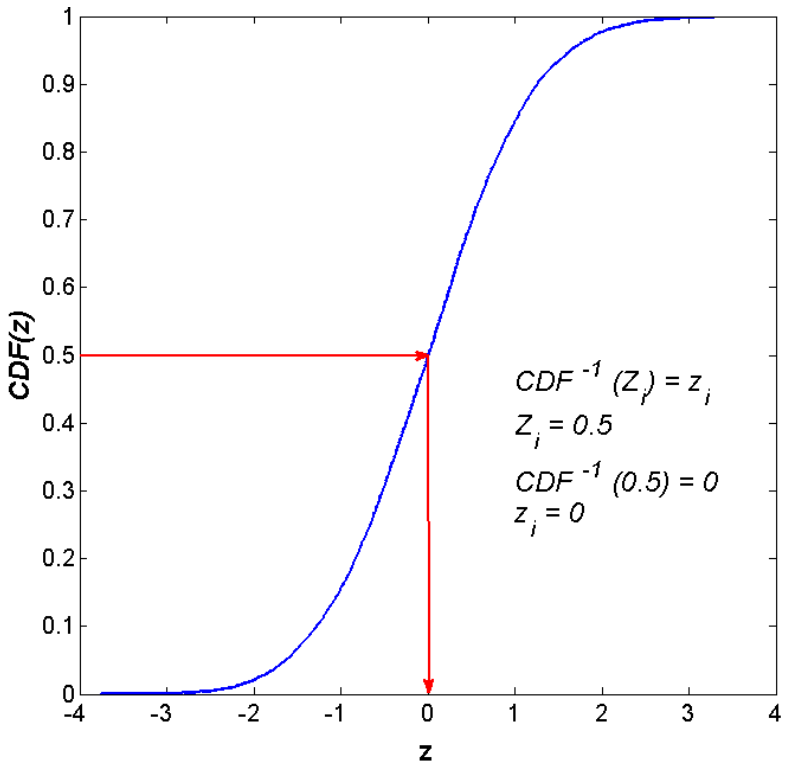


Figure 3: Example of the inverse CDF of mean local AR used to generate a synthetic z_i of 0 based on a random Z_i of 0.5 (z_i = 0 equates to x_i = 6.044 (i.e. the mean of the raw data))

Synthetic z_i values are then transformed back into the linguistically meaningful x-space by:

$$x_i = (\sqrt{2}\sigma_x * z_i) + \mu_x, \tag{9}$$

and used as the mean value for the normal distribution of a single synthetic speaker.

This process is repeated over a number of simulations (n). By the law of large numbers (Wackerly *et al.* 2008: 451), the distribution of $z = (z_1, z_2, \ldots, z_n)$ will converge on $N(0, \frac{1}{2})$ as $n \to \infty$. Therefore, with large n the synthetically generated values will have approximately the same normal distribution as the raw values.

2.3.2. Generating synthetic SD of local AR

The SD of local AR is denoted by y such that y_i is the SD for a single speaker. To generate synthetic y_i values, it is necessary to account for the correlation between the means and SDs in the raw data. Figure 4 reveals a significant (Pearson's r = 0.3964; p = 0.0019), positive correlation such that speakers with higher average AR generally display greater within-speaker variability. Potentially, this is because speakers with higher mean AR are able to exploit a wider range of variability, particularly in higher rates. Since the mean and SD are seemingly not independent a further (simple) projection was incorporated into the simulation of SDs. Rather than sampling from a normal PDF based on the mean and SD of the raw SDs (as with the raw means), $N(a\,x_i + b, \beta)$ was used where the linear trend line determines the mean ($ax_i + b$) and variance around the trend line (residuals) determines the SD (β) (see Figure 4). The mean of the normal distribution from which synthetic SD values are generated, therefore varies as a function of the associated synthetic mean value (x_i).

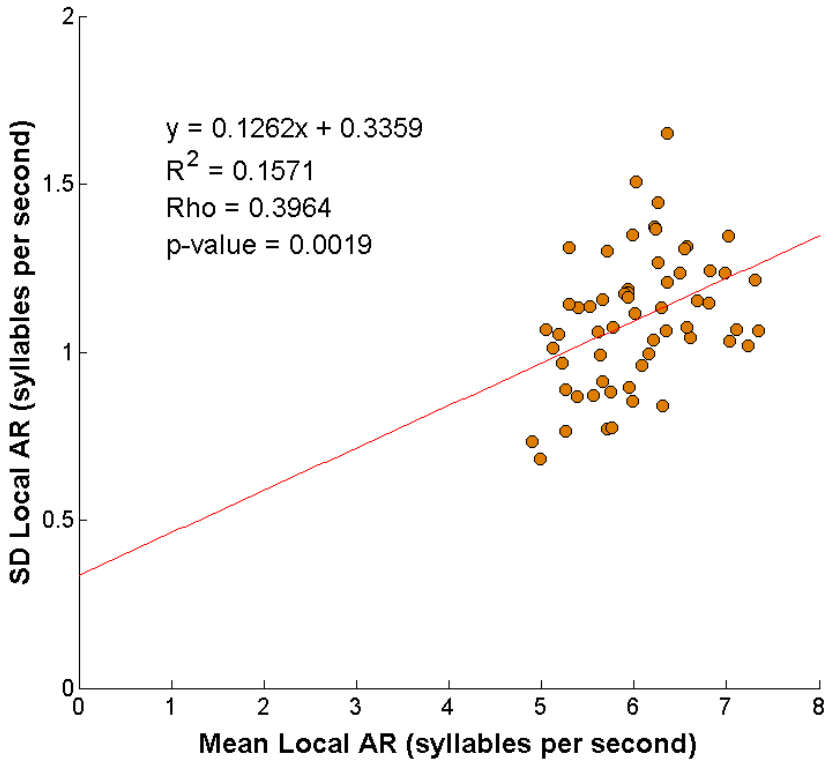


Figure 4: Mean local AR plotted against SD of local AR (syllables/ second) for each of the 59 raw speakers

Following the same procedure as before, synthetic y_i values were generated by converting $N(a\,x_i + b, \beta)$ to a normal PDF for each synthetic x_i. Again, the inverse CDF was used to

transform a random variable $Z^*_i \in [0, 1]$ into normalised z^*_i values, before transforming back to the *y*-space. The synthetic mean and SD values represent the normal distribution $N(x_i, y_i)$ for a new synthetic speaker. From this distribution, individual local AR tokens were synthesised using the same procedures as above. The process of generating synthetic means and SDs was performed 941 times. The synthetic speakers were pooled with the existing 59 raw speakers to create a reference sample of up to 1000 speakers. For the synthetic speakers, up to 200 tokens per speaker were generated. For each of the 59 raw speakers, MCS based on the mean and SD of the 26 raw tokens per speaker were used to generate an additional 174 tokens per speaker.

2.3.3. The synthetic data

The distributions of means and SDs in the raw data, synthetic data and all reference data combined (raw + synthetic) based on 26 tokens per speaker were compared to assess how well MCS approximates patterns in the raw data. Table 1 reveals minimal difference in the mean of the means (μ_x), with the raw data displaying a μ_x 0.023 higher than the synthetic data. The SD of the means in the synthetic data is higher than that in the raw data, although the difference is again negligible (0.015). *p*-values were generated from a comparison of the raw data and the synthetic data, as well as the raw data and all of the reference data, using independent t-tests. The differences between distributions were found to be non-significant, with *p* approaching 1 in both cases (Table 1).

	Mean	**SD**	**t-test (*p*-value)**
Raw data (59 speakers)	6.044	0.627	-
Synthetic data (941 speakers)	6.021	0.642	0.7954
Pooled data (1000 speakers)	6.023	0641	0.8065

Table 1: Mean and SD of mean local AR (syllables/ sec) for the raw data, synthetic data and all reference data

There are marginal differences in the distributions of SD (*y*) values, with μ_y 0.0029 higher for the raw data than for the synthetic data (Table 2). The differences between the sets in terms of σ_y are also marginal with SD in the raw data 0.008 greater than in the synthetic data. Again, paired independent t-tests were performed using the raw data and synthetic data, and the raw data and all reference data combined. In both cases, the differences were non-significant with *p*-values much closer to 1 than for the means.

	Mean	**SD**	**t-test (*p*-value)**
Raw data (59 speakers)	1.098	0.199	-
Synthetic data (941 speakers)	1.095	0.191	0.8989
Pooled data (1000 speakers)	1.095	0.191	0.9049

Table 2: Mean and SD of SD of local AR (syllables/ sec) for the raw data, synthetic data and all reference data

2.4. The present study

A MatLab implementation (Morrison 2007) of Aitken and Lucy's (2004) MVKD formula was used to compute raw LRs. For both the development and test sets independently, 20 SS and 380 DS LRs were computed as the number of speakers in the reference data was systematically increased by 1 starting with 10 and ending with 1000. To test the effect of the number of reference speakers, only the first 26 tokens per speaker were included. Using a random reference sample of 200 speakers, LRs were also computed as a single token per speaker was added to the reference data up to a maximum of 200 tokens. For both experiments, LR scores were transformed using a base-10 logarithm to account for the skew in the distribution of raw LRs. Using log LRs, zero represents the threshold, whereby positive values offer support for the prosecution (same-speaker) and negative values offer support for the defence (different-speaker). The magnitude of the log LR indicates the strength of the support for prosecution of defence.

For the test set, $\log_{10}$ LRs were calibrated based on weights generated from scores for the development set. Calibration was performed using a robust implementation (Morrison 2009) of Brümmer's (2007) logistic-regression-based procedure (Morrison 2013). Following Rose (2012), the effects on log LRs of the number of speakers and tokens per speaker are presented in the form of boxplots, which include the median, interquartile range (1^{st} to 3^{rd} quartile) and overall range (including outliers). The magnitude of LR output is assessed with reference to Table 3. The verbal scale provides a qualitative expression of the strength of numerical LR data which may be better understood by the court. For the purposes of the present study the verbal scale also allows broad differences in LR performance to be assessed.

Range of $\mathbf{Log_{10}}$ LR	**Verbal expression**
$\pm 4 \rightarrow \pm 5$	Very strong evidence
$\pm 3 \rightarrow \pm 4$	Strong evidence
$\pm 2 \rightarrow \pm 3$	Moderately strong evidence
$\pm 1 \rightarrow \pm 2$	Moderate evidence
$0 \rightarrow \pm 1$	Limited evidence

Table 3: Verbal expressions of $\log_{10}$ LRs according to Champod and Evett's verbal scale (2000:240)

Both equal error rate (EER) and log LR cost (C_{llr}) (Brümmer and du Preez 2006) are used as metrics of system validity. Validity refers to how well the system (i.e. the variable and the particular set of data) is able to separate same- (SS) and different-speaker (DS) pairs. EER provides a "hard", "error-based" (Brümmer and du Preez 2006: 230) measure of validity dealing with binary accept-reject decisions EER has an operating point at which the number of false hits (DS pairs offering support for the prosecution) and false misses (SS pairs offering support for the defence) are equal. As such, EER is not a forensically realistic metric of validity since in forensic casework the imperative is to avoid false hits. C_{llr} is a gradient, "soft" detector, which penalises the system for high contrary-to-fact LRs (Rose 2010). In both cases, optimum validity (i.e. complete separation of SS and DS pairs) is 0. C_{llr} becomes sub-optimum as it approaches 1, whilst values of greater than 1 are considered very poor performance (Morrison 2011b).

3. Results

3.1. Number of reference speakers

Figures 5 and 6 display the distributions of calibrated same-speaker (SS) and different-speaker (DS) $\log_{10}$ LRs as a function of the number of reference speakers. Across conditions, SS comparisons predominantly achieve LRs equivalent to 'limited' support for the prosecution, with the majority of $\log_{10}$ LRs approaching 0 (neutral evidence; no support for prosecution or defence). Of the 20 SS comparisons, only one achieves contrary-to-fact support for the defence, although this value never exceeds -0.1. DS comparisons generally perform worse, with over 50% of the 380 comparisons in each condition consistently achieving positive $\log_{10}$ LRs. However, in all conditions the magnitude of the contrary-to-fact DS LRs never exceeds 0.4. The poor overall strength of evidence achieved and the high proportion of false hits, reflects the poor value of AR as a speaker discriminant.

Figure 5 reveals striking consistency in the distribution of $\log_{10}$ LRs as the number of reference speakers increases. The largest difference in the distribution of LRs from the *true* LR (based on 1000 speakers) is found with 50 speakers. However, even in this case the differences are incredibly small (difference in medians = 0.011, difference in ranges = 0.1399). There is marginal underestimation of the median, inter-quartile range and overall range with the smallest amount of reference data (10 speakers). Further, the greatest instability in the distribution of log LRs is found with small amounts of reference data, although the fluctuation is very small.

DS LRs (Figure 6) were also remarkably robust to the effects of differences in reference sample size. Across all conditions the median fluctuates maximally within a range of 0.017. As with SS LRs, there is underestimation of the median (i.e. closer to zero), inter-quartile range and overall range with only 10 speakers compared with the *true* LR. Further, there is minor instability in the distribution of LRs found between the 10- and 50-speaker conditions, with some individual DS LRs increasing by up to 0.17. However, given that the overall range of LRs is always between 'limited' support for the prosecution and 'limited' support for the defence, it is considered that the LRs from the 10-speaker condition adequately capture the *true* distribution of DS LRs for this data set.

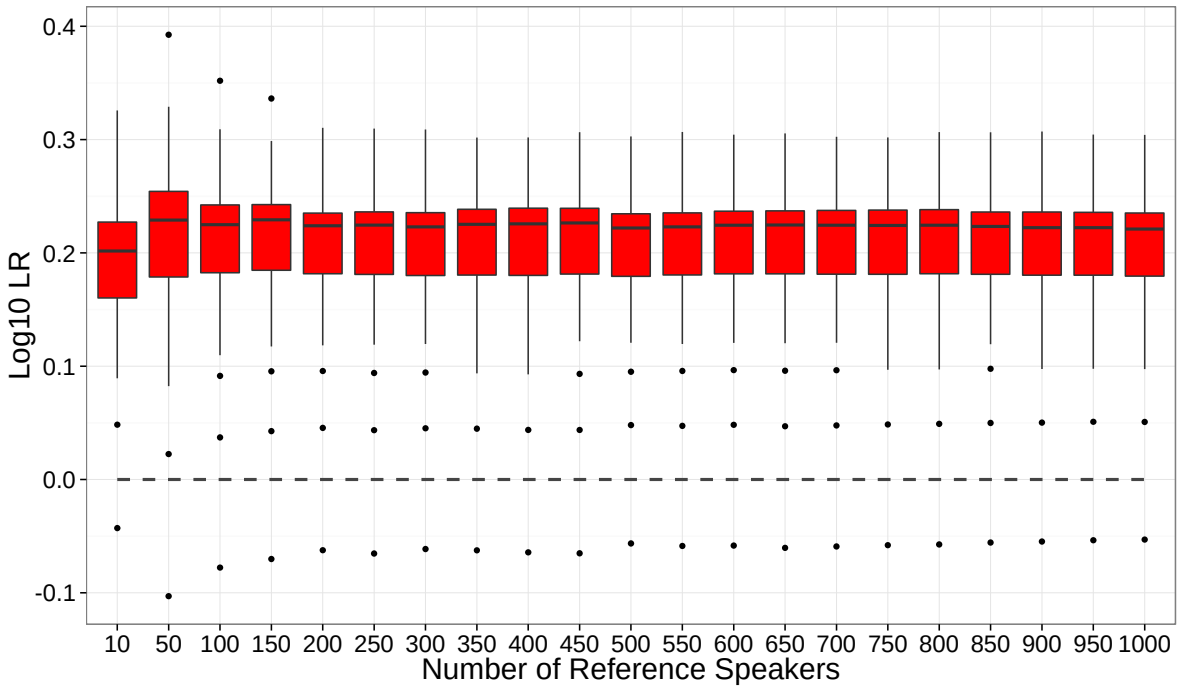


Figure 5: Calibrated SS $\log_{10}$ LRs as a function of the number of reference speakers (where mid line = median, filled box = interquartile range (containing middle 50% of the data), whiskers = scores outside the middle 50%, dots = outliers, dashed line = neutral evidence ('unity')) (Rose 2012)

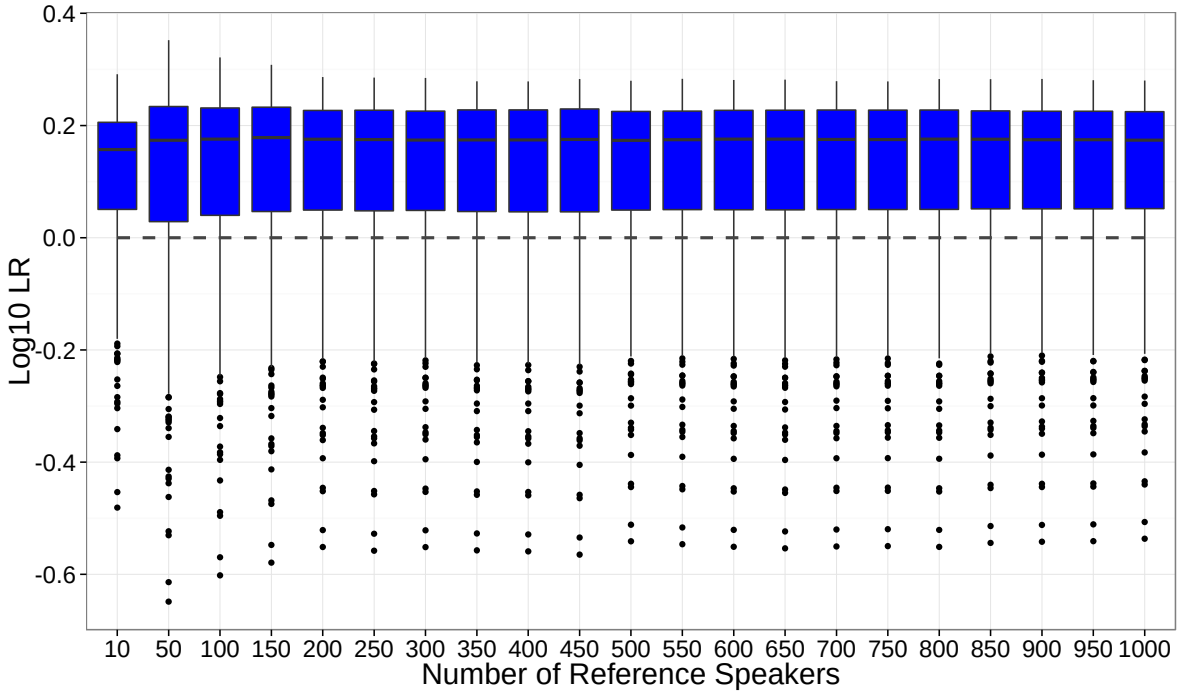


Figure 6: Calibrated DS $\log_{10}$ LRs as a function of the number of reference speakers

Figure 7 shows a line graph of EER as a function of the number of reference speakers, with the *true* EER (LRs based on 1000 reference speakers) plotted as a single line across all N speakers conditions as a means of comparison. The box-like fluctuations in EER are accounted for the fact that EER is a categorical metric with intervals dependent only on the

number of comparisons. Given that the variation in Figure 7 occurs within such a small percentage range, the fluctuations are attributed to a single, or small number of, comparison(s) being correctly/incorrectly categorised at given N speakers conditions. The EER of the *true* LRs is 35.1%. This means that when the proportion of false hits and false misses is equal, the system incorrectly classifies SS as DS and vice versa in 35.1% of pairs. Such poor performance reflects the very high proportion of DS pairs classified as offering support for the prosecution. There is some fluctuation in performance as the number of reference speakers increases. However, the variation appears to be random since the *true* EER is achieved with as few as 17 speakers. Indeed, the maximum extent of the fluctuation in EER performance is just 0.3% across all conditions, suggesting that categorical accept-reject performance of the system is relatively stable across different sample sizes.

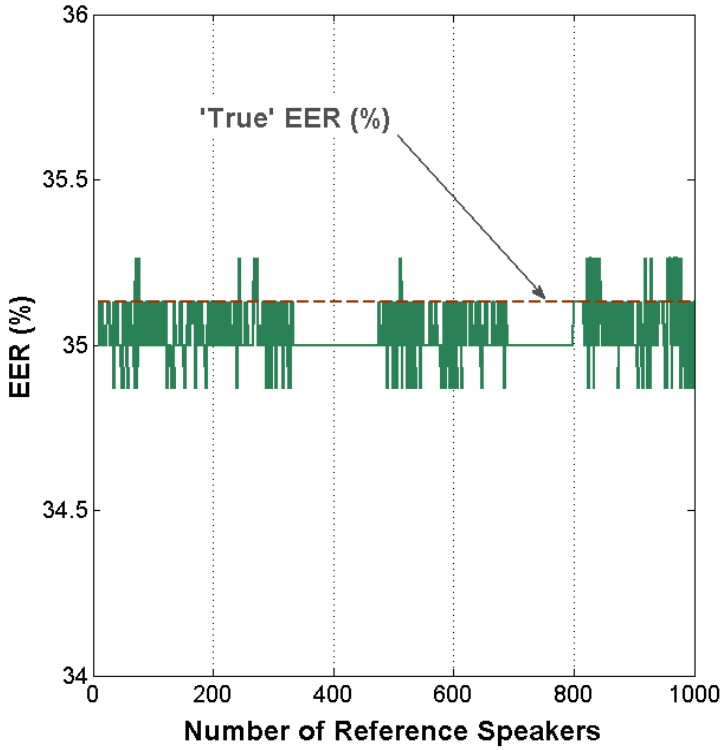


Figure 7: Equal error rate (EER, %) based on calibrated LRs as a function of the number of reference speakers

Figure 8 displays C_{llr} performance as a function of the number of reference speakers. The baseline C_{llr} achieved with the *true* LRs is approaching 1. As with EER, this reflects very poor system validity for AR. Performance based on C_{llr}, as a function of the number of reference speakers, is more systematic than for EER. There is overestimation of performance using C_{llr} with fewer than 200 speakers, such that the lowest value (best validity) is achieved with 57 speakers (0.963). With greater than 200 speakers performance appears asymptotic, although there is still marginal increase in C_{llr}. However, the overall range of C_{llr} remains very small since the C_{llr}s of the systems with very small numbers of speakers (10-20) are almost equivalent to the *true* C_{llr} of 0.971.

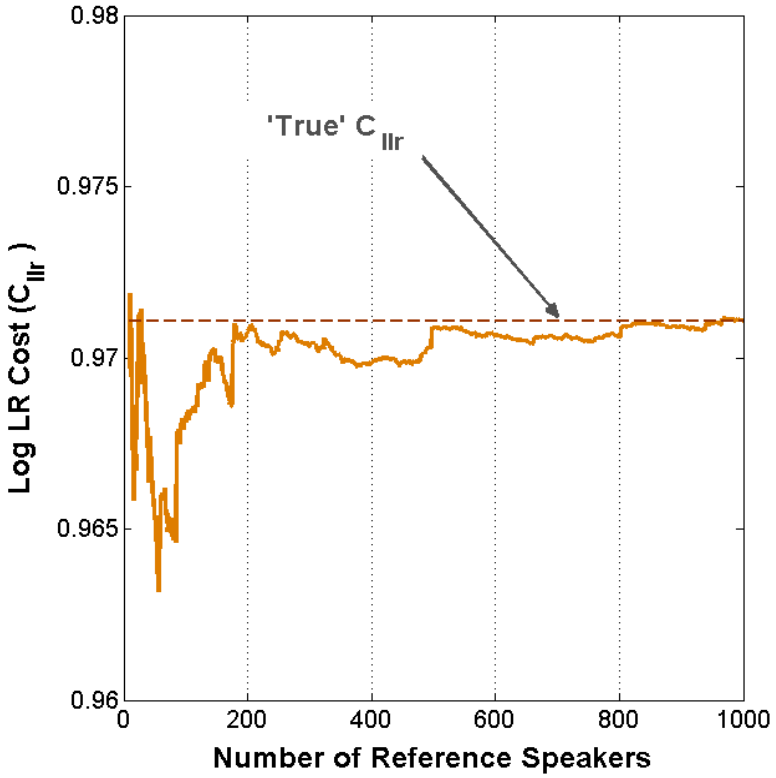


Figure 8: Log LR Cost (C_{llr}) based on calibrated LRs as a function of the number of reference speakers

Uncalibrated SS (Figure 9) and DS (Figure 10) LR scores were also plotted for the same set of test data. The uncalibrated scores display more complex sensitivity to the size of the reference data than calibrated LRs. For SS pairs the interquartile ranges of LR scores are always within one order of magnitude. There is underestimation of the median strength of evidence with small numbers of speakers (the lowest median SS score is achieved with 10 reference speakers). Further, there is marked instability in the interquartile range and overall range with smaller numbers of speakers compared with the *true* scores. The interquartile ranges and overall ranges with fewer than 200 speakers are also consistently underestimated.

Much more significant is the effect of different numbers of reference speaker on individual SS pairs. This is clearly seen in the lowest LR score (offering the most contrary-to-fact support for the defence), which is classed as an outlier. With between 10 and 50 reference speakers this score is around -0.5, equivalent to 'limited' support for the defence. By the inclusion of 150 reference speakers, the score has increased by the equivalent of one order of magnitude (to 'moderate' support for the defence). Assessment of individual scores suggests that certain SS pairs are more affected by the size of the reference sample than others.

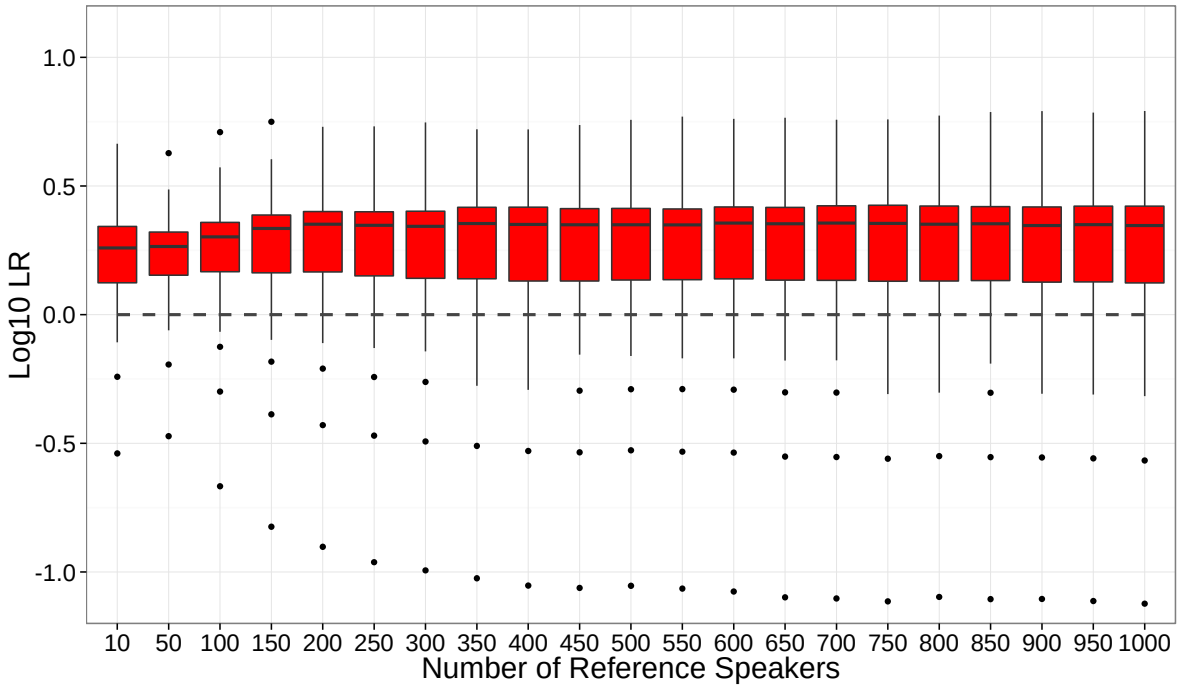


Figure 9: Uncalibrated SS $\log_{10}$ LRs as a function of the number of reference speakers

A similar pattern is found in Figure 10, although the effects are greater for DS pairs. Whilst the median DS score is relatively robust to sample size, the interquartile range and overall range is considerably underestimated with smaller samples. As with SS pairs, this underestimation is greatest in the 50 speaker condition. There is also greater variability in the distributions of scores when using comparatively small amounts of data. Again, the most significant effects are found in the highest, outlying values. For the two most extreme negative DS scores, the strength of evidence increases from less than -2 ('moderate' evidence) to over -3 ('moderately strong' evidence), equivalent to an increase of two orders of magnitude between the 10- and 1000-speaker conditions. Other smaller outliers increase by one order of magnitude as the number of reference speakers increases. This suggests that the magnitude of the score relative to the distribution plays some role in determining sensitivity to sample size.

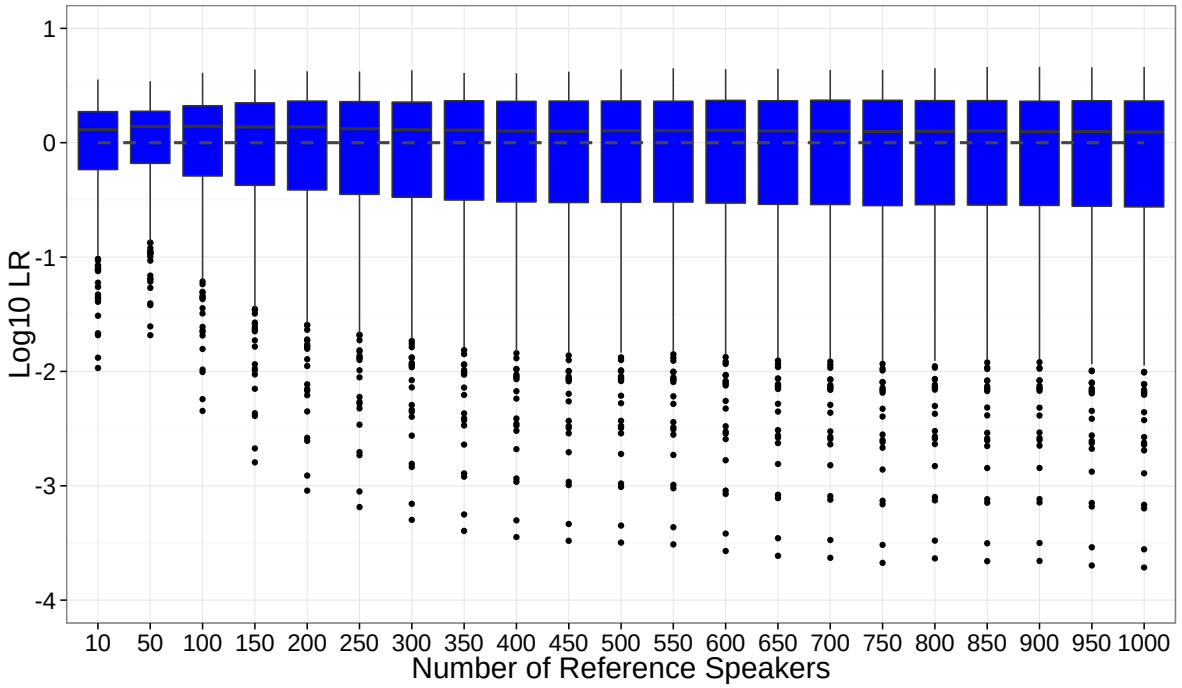


Figure 10: Uncalibrated DS $\log_{10}$ LRs as a function of the number of reference speakers

3.2. Number of tokens per reference speaker

In this section, LRs are computed using a random sample of 200 speakers and between 10 and 200 tokens per speaker. Output is assessed as a function of the number of tokens. Figure 11 shows the distributions of calibrated same-speaker (SS) $\log_{10}$ LRs as the number of tokens per reference speaker increases. There is very minimal overestimation of the median and range when using smaller samples, such that the highest median strength of evidence (0.239) is reached with 5 tokens per speaker and the highest range reached with 6 tokens. As in §3.1, the extent of variation as a function of the number of tokens per speaker is minimal with all bar one of the SS comparisons consistently achieving a LR that supports the prosecution. The magnitude of the calibrated SS LRs is consistently close to the zero turning point (unity).

Figure 12 reveals a similar pattern in terms of calibrated different-speaker (DS) $\log_{10}$ LRs. The median remains essentially the same across all conditions, even when very small numbers of tokens are included in the reference data. The interquartile and overall ranges are marginally overestimated with small numbers of tokens compared with the distribution of *true* LRs. This is reflected in the decrease in the strength of evidence for the two most extreme negative DS LRs, although in absolute terms these LRs increase by less than 0.1 between the 10- and 200-token conditions. In all conditions, DS LRs are maximally spread within a range of two orders of magnitude (between 'limited' support for prosecution and 'limited' support for defence). Further, the middle 50% of DS comparisons consistently achieve LRs offering contrary-to-fact support for the prosecution, although their absolute magnitude is relatively low (never greater than 0.3).

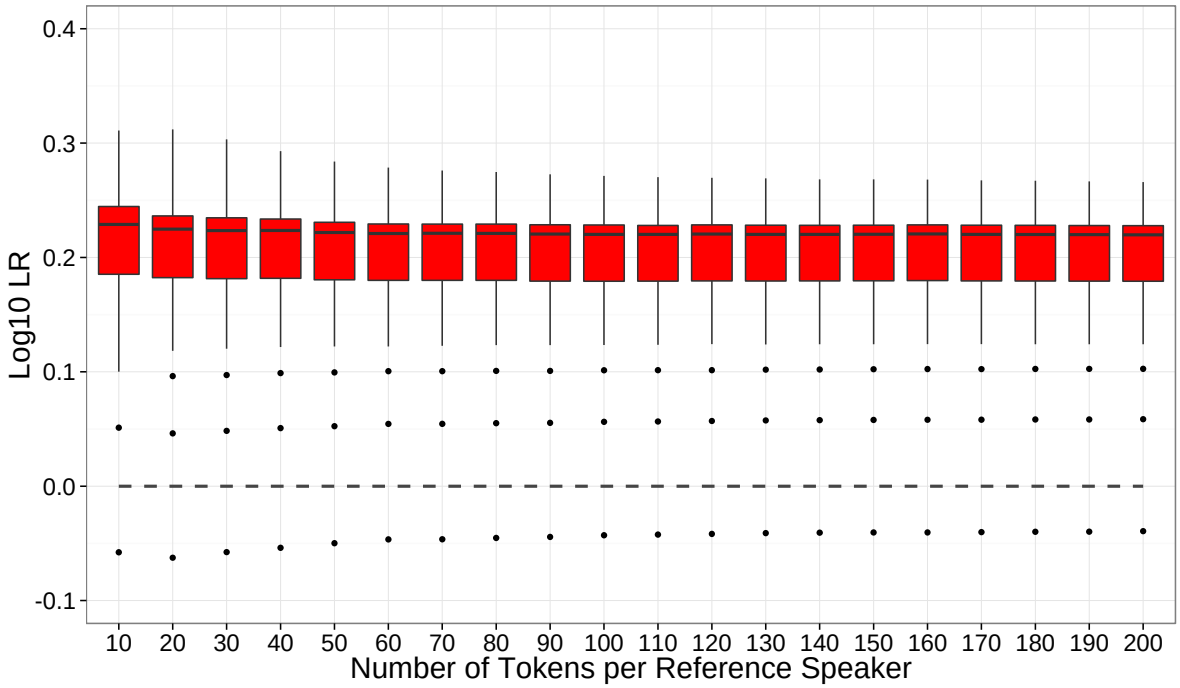


Figure 11: Distributions of calibrated SS $\log_{10}$ LRs as a function of the number of tokens per reference speaker

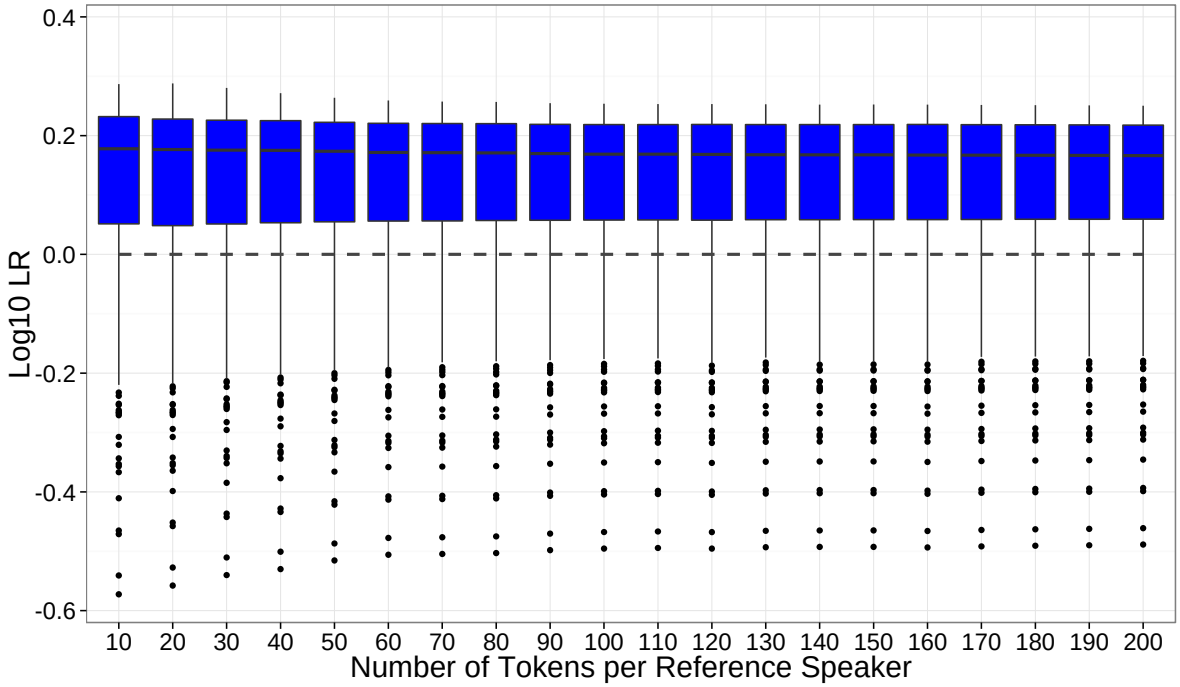


Figure 12: Distributions of calibrated DS $\log_{10}$ LRs as a function of the number of tokens per reference speaker

Consistent with Figures 11 and 12, EER is relatively robust to the number of tokens per reference speaker (Figure 13). EER based on the maximum amount of reference data is 35%. With the inclusion of greater than 96 tokens per speaker EER remains consistent (35%). However, the same EER is achieved with just 2 tokens per speaker suggesting that increasing

the number of tokens does not offer anything in terms of categorical system validity. As with Figure 8, there is variability in EER although this is not systematic. Further, the variability is within a very narrow range (maximally 0.26%) and as such can be assumed to be of little practical interest. Figure 14 displays C_{llr} based on calibrated LRs as a function of the number of tokens per reference speaker. Relative to the *true* LR baseline, there is a small amount of overestimation of how well the system performs when using small amounts of data. The system with the lowest C_{llr} is based on just 6 tokens per speaker. After this point C_{llr} increases until performance appears asymptotic with greater than 100 tokens per speaker, although the range of observed C_{llr} variability is rather small (maximally 0.005).

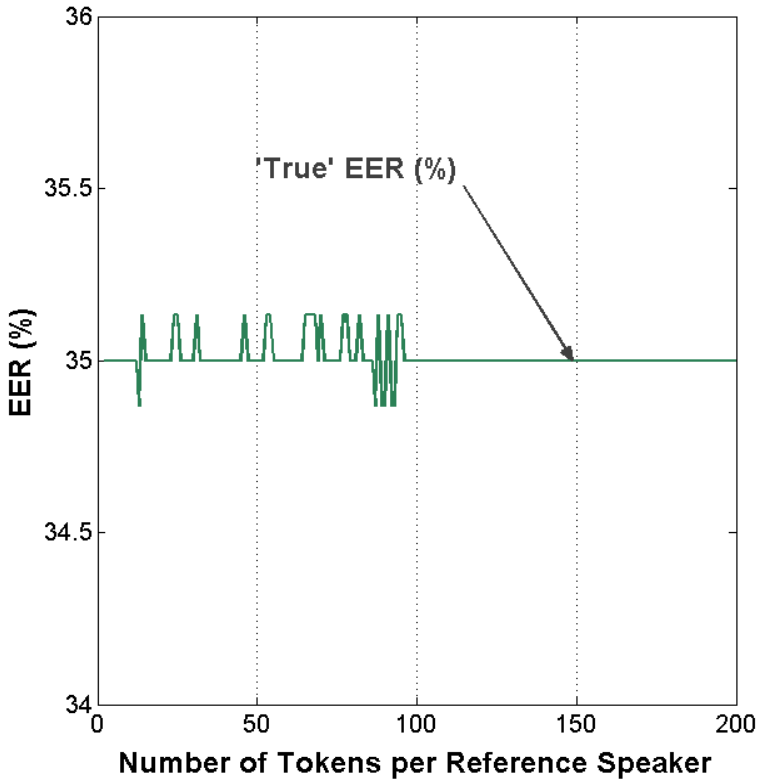


Figure 13: Equal error rate (EER, %) based on calibrated LRs as a function of the number of tokens per reference speaker

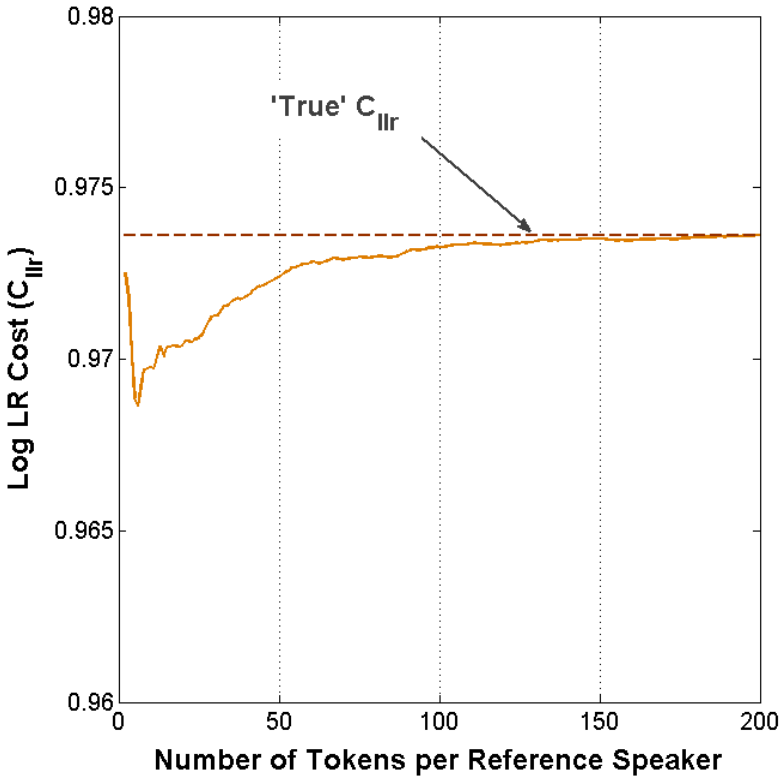


Figure 14: Log LR Cost (C_{llr}) based on calibrated LRs as a function of the number of tokens per reference speaker

As in §3.1.2, calibration plays a role in minimising the effects of small numbers of tokens. Figure 15 reveals considerable differences in the distributions of scores using small reference samples compared with the *true* LR. The median is underestimated (i.e. closer to zero) to a small extent with fewer than 50 tokens. Between the 10- and 50-token conditions there is also greater variability in the median. The interquartile range is narrower with smaller numbers of tokens (and narrowest with 20 tokens). Again the outlying, contrary-to-fact values are affected to the greatest extent. Considering the outlier with the largest negative value, strength of evidence increases by one order of magnitude (increase = 0.97) across all conditions from 'limited' to 'moderate' support for the defence.

The effects of small sample size are more dramatic for DS scores than for SS scores (Figure 16). The median strength of evidence decreases as the number of tokens per speaker increases. As such the median based on 10 tokens is positive whilst the median based on 200 tokens is negative. However, in absolute terms the differences in the medians is relatively small (0.17). As in Figure 15, the interquartile range of DS scores is underestimated when using smaller numbers of tokens per speaker, only stabilising after 100 tokens. Generally, strength of evidence increases (more support for the defence) with larger amounts of data per reference speaker. This is reflected in the magnitude of the most outlying negative scores. Considering one particular outlying DS pair, there is an increase in strength of evidence between the 20 and 200 token conditions equivalent to the difference between 'moderately strong' and 'very strong' support for the defence. In terms of the magnitude of the $\log_{10}$ LR scores this is an increase of two orders of magnitude.

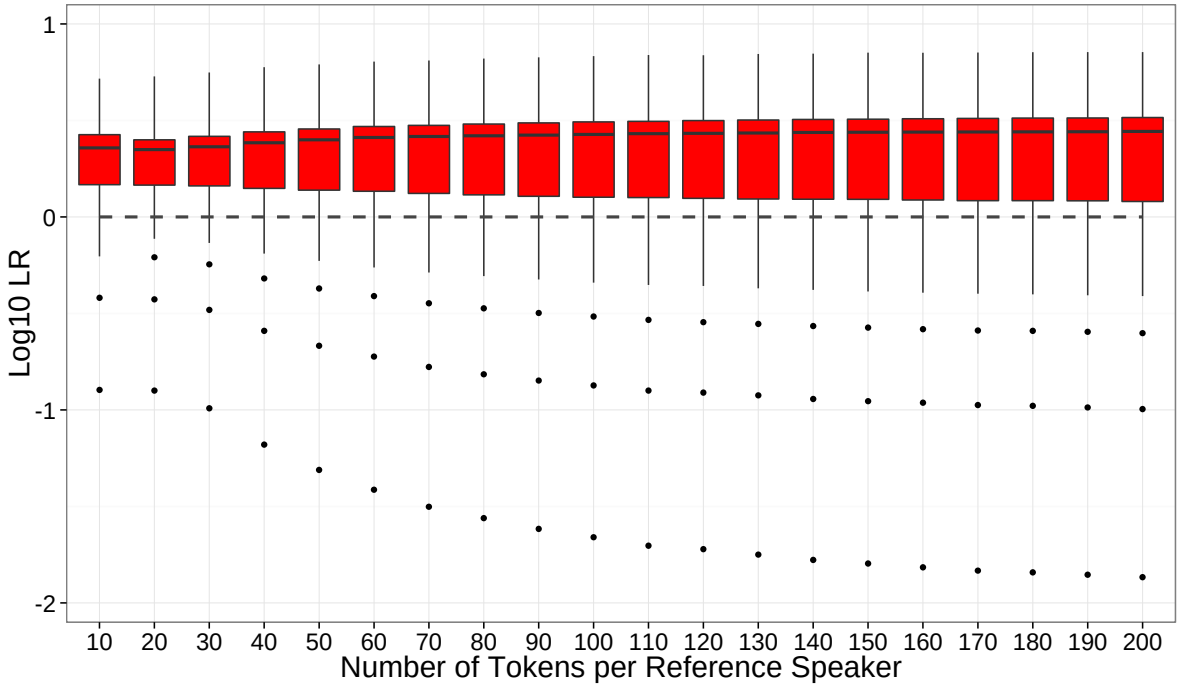


Figure 15: Uncalibrated SS $\log_{10}$ LR scores as a function of the number of tokens per reference speaker

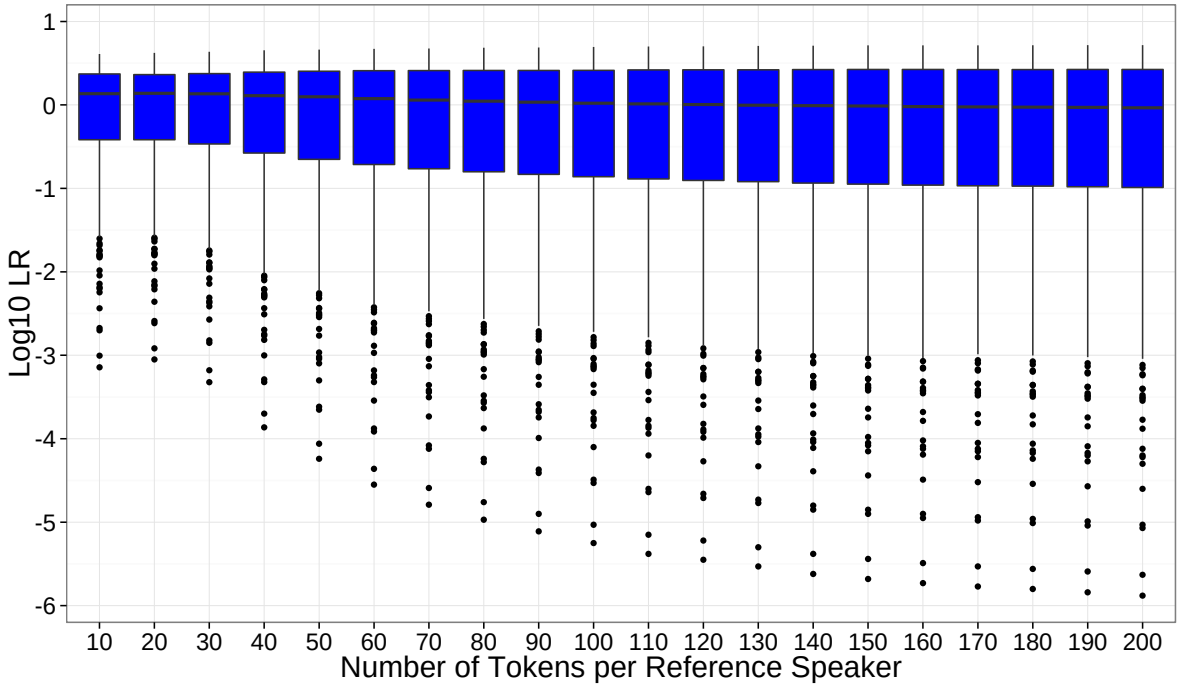


Figure 16: Uncalibrated DS $\log_{10}$ LR scores as a function of the number of tokens per reference speaker

4. Discussion

4.1. Effects of sample size on calibrated LRs

The results in §3.1.1 and §3.2.1reveal that calibrated SS and DS LRs for this set of data are relatively robust to the number of reference speakers and the number of tokens per reference speaker. Limited variability was found in the distributions of LRs with smaller amounts of reference data. However, the medians, interquartile ranges and overall ranges of LRs with the smallest amounts of data were consistently within the same order of magnitude as the distribution of *true* LRs. These findings suggest that a fairly precise estimate of the magnitude of calibrated SS and DS LRs can be achieved using 10 reference speakers and just 2 tokens per speaker. Whilst not considered directly in this study, there are, however, potential interactions between the number of speakers and the number of tokens to consider.

Consistent patterns were also found with regard to the validity of the systems based on calibrated LRs. EER remained stable as the number of speakers and the number of tokens per speaker increased, despite random variation within a very narrow range. Such variability can be explained by the inherently poor performance of AR as a speaker discriminant. Since the calibrated LRs are very close to zero, slight changes in the distributions of within- or between-speaker variation in the reference data can cause marginally positive values to become marginally negative and vice versa. Since EER deals only in absolute false hits and false misses such minor fluctuation has a direct effect on validity. This is an inherent limitation of using EER as a measure of performance, particularly for poor discriminants.

C_{llr} overcomes this limitation of EER by considering the gradient magnitude of the 'errors' made by the system. C_{llr} was found to be at its lowest when using smaller numbers of speakers and tokens per speaker. As such, the C_{llr} validity of calibrated LRs is systematically overestimated with small amounts of data relative to the *true* LRs. This improved performance is attributed to the slightly wider interquartile range for DS comparisons based on small amounts of reference data. Given that the interquartile ranges across all conditions in Figures 6 and 12 are positive (i.e. support for defence), the wider interquartile range means that the first quartile is closer to zero. Since the magnitude of a proportion of the 'errors' is lower when using smaller amounts of data, C_{llr} is also lower.

4.2. The role of calibration

As highlighted in §3.1.2 and §3.2.2, calibration has played an important role in reducing the sensitivity of LRs to small amounts of reference data. As such, the calibrated results in this study are not comparable with Ishihara and Kinoshita (2008), Hughes and Foulkes (2012) or Rose (2012) since these studies did not calibrate scores. The uncalibrated results are very much consistent with previous studies in that scores are misrepresentative and unstable when using small numbers of speakers and tokens. However, whilst previous studies found overestimation of scores within a wider range when using small samples, the scores in this study were underestimated and within a narrower range with small amounts of reference data. This highlights that different variables are affected by sample size in different ways.

The importance of calibration may also be specifically related to AR. The calibration procedure used in this study is configured to improve C_{llr}. For both experiments, the ranges of uncalibrated SS and DS scores increase as a function of the amount of reference data resulting in more contrary-to-fact scores of a higher magnitude when using large amounts of

reference data compared with smaller samples. As such, calibration weights generated for systems based on more reference data are greater than those based on less reference data. Despite calibration improving C_{llr} to different degrees for different conditions, Figures 8 and 14 suggest that AR performance cannot be improved beyond a ceiling close to unity, due to its inherently poor discriminatory value. For better speaker discriminants the role of calibration relative to the size of the reference sample may be different.

The uncalibrated LR results do highlight three important general issues. Firstly, there appears to be an interaction between calibration procedures and the overall sensitivity of LRs to sample size, at least in the present study. Whilst calibration counteracts the effects of small sample sizes, calibrated LRs are spread over a narrower range and are much closer to zero compared with the uncalibrated scores. Secondly, certain pairs of samples are more susceptible to the effects of sample size than others. This may be related to the magnitude of the score relative to the rest of the distribution. Thirdly, the uncalibrated results in Figures 9, 10, 15 and 16 suggest that within- and between-speaker variation are actually very poorly estimated when the background model consists of small amounts of reference data. Therefore, in the absence of calibration, considerable caution should be exercised when interpreting the absolute or relative value of scores generated using a small reference sample.

4.3. MCS procedures for FVC

MCS have provided a valuable resource for investigating the issue of sample size in this study. The procedures implemented have been able to adequately generate a large amount of univariate data which captures the correlation between mean and SD of AR fairly well. In practical terms, MCS are easy to implement and can be used to generate a lot of data quickly and efficiently. Crucially, however, MCS are dependent on the assumption that the underlying distribution of within- and between-speaker variation in the relevant population is known, either through previous research or raw data. As such, caution is advised when implementing MCS procedures using already small sets of raw data. Further, even with larger sets of raw data, procedures for assessing the precision of the representative data should be implemented as a preliminary, exploratory tool.

5. Conclusion

This paper has considered the effects of sample size based on the number of reference speakers and tokens per reference speaker for assessing typicality when computing numerical LRs based on local AR. Calibrated LRs were found to be robust to sample size effects, whilst uncalibrated scores displayed much more sensitivity to the amount of reference data used. Although calibration has been shown to have value in minimising sample size effects for this data set, the generalisability of these results to other FVC variables remains an empirical question. More generally, the results highlight the importance of considering the potential effects of the amount of reference data used when computing LRs both in research and in casework. Future work should focus on testing the sensitivity of much better speaker discriminatory variables to sample size. Attention should also be directed towards developing MCS procedures for synthesising more complex, multivariate FVC variables.

Acknowledgements

This research is funded by a UK Economic and Social Research Council DTC scholarship and the Marie Curie Actions EC Grant Agreement No. PITN-GA-2009-238803 (Bayesian Biometrics for Forensics, BBfor2). We are grateful to Paul Foulkes, Peter French and Dominic Watt for feedback on earlier versions of this paper. Thanks also to Anil Alexander, Esam Alzqhoul, Niko Brümmer, Philip Harrison, Geoffrey Morrison and Balu Nair for scripts which facilitated data analysis. We are thankful to two anonymous reviewers for their valuable comments.

References

AITKEN, COLIN G. G. & LUCY, DAVID. 2004. "Evaluation of trace evidence in the form of multivariate data". *Applied Statistics 54*: 109–122.

AITKEN, COLIN G. G. & TARONI, FRANCO. 2004. *Statistics and the Evaluation of Evidence for Forensic Scientists* (2nd edition). Chichester: John Wiley.

BERNARD, J. R. 1967. "Some measurements of some sounds of Australian English". PhD dissertation, University of Sydney.

BRÜMMER, NIKO. 2007. "FoCal multi-class: toolkit for evaluation, fusion and calibration of multi-class recognition scores".
Online resource: http://sites.google.com/site/nikobrummer/focal

BRÜMMER, NIKO & DU PREEZ, JOHAN. 2006. "Application-independent evaluation of speaker detection". *Computer Speech and Language* 20(2/3): 230–275.

CHAMPOD, CHRISTOPHE & EVETT, IAN W. 2000. "Commentary on A.P.A Broeders (1999) 'Some observations on the use of probability scales in forensic identification". *Forensic Linguistics* 7(2): 238–243.

GOLD, ERICA. In progress. "Calculation of likelihood ratios using phonetic and linguistic features". PhD dissertation, University of York.

GOLDMAN-EISLER, F. 1968. *Psycholinguistics: Experiments in Spontaneous Speech.* London: Academic Press.

HUGHES, VINCENT & FOULKES, PAUL. 2012. "Effects of variation on the computation of numerical likelihood ratios for forensic voice comparison". Paper presented at the *International Association of Forensic Phonetics and Acoustics (IAFPA) Conference*. Universidad Internacional Menedez Pelayo, Santander, Spain. 5–8 August 2012.

ISHIHARA, SHUNICHI & KINOSHITA, YUKO. 2008. "How many do we need? Exploration of the population size effect on the performance of forensic speaker classification". In: *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)*. Brisbane, Australia. pp. 1941-1944.

JESSEN, MICHAEL. 2007. "Forensic reference data on articulation rate in German". *Science and Justice* 47: 50–67.

KÜNZEL, HERMAN. J. 1997. "Some general phonetic and forensic aspects of speaking tempo". *International Journal of Speech, Language and the Law* 4(1): 48–83.

LINDLEY, DENNIS. V. 1977. "A problem in forensic science". *Biometrika* 64: 207–213.

MILLER, J., GROSJEAN, F. & LOMANTO, C. 1984. " Articulation rate and its variability in spontaneous speech: a reanalysis and some implications". *Phonetica* 41: 215–225.

MORRISON, GEOFFREY S. 2007. "MatLab implementation of Aitken and Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation".
Online resource: http://geoff-morrison.net/Software/multivar_kernel_LR.m

MORRISON, GEOFFREY S. 2009. "Robust version of train_llr_fusion.m from Niko Brümmer's FoCal toolbox".
Online resource: http://geoff-morrison.net/Software/train_llr_fusion_robust.m

MORRISON, GEOFFREY S. 2011a. "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM". *Speech Communication* 53(2): 242–256.

MORRISON, GEOFFREY S. 2011b. "Measuring the validity and reliability of forensic likelihood-ratio systems". *Science and Justice* 51(3): 91-98.

MORRISON, GEOFFREY S. 2013. "Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio". *Australian Journal of Forensic Sciences* 45(2): 173–197.

MORRISON, GEOFFREY S., ROSE, PHILIP & ZHANG, CUILING. 2012. "Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice". *Australian Journal of Forensic Sciences* 44: 155–167.

MORRISON, GEOFFREY S., OCHOA, FELIPE & THIRUVARAN, THARMARAJAH. 2012. "Database selection for forensic voice comparison". *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop*. Singapore. 62-77.

NOLAN, FRANCIS, MCDOUGALL, KIRSTY, DE JONG, GEA & HUDSON, TOBY. 2009. "The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research". *International Journal of Speech, Language and the Law* 16(1): 31–57.

ROBERTON, BERNARD & VIGNAUX, G. A. 1995. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom.* Chichester: John Wiley.

ROSE, PHILIP. 2010. "Bernard's 18 – vowel inventory size and strength of forensic voice comparison evidence". In: M. Tabain, J. Fletcher, D. Grayden, J. Hayek & A. Butcher (eds) *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*. Canberra: ASSTA. pp. 30-33.

ROSE, PHILIP. 2011. "Forensic voice comparison with Japanese vowel acoustics – a likelihood ratio-based approach using segmental cepstra. In: *Proceedings of the 17th International Congress of Phonetic Sciences XVII*. Hong Kong, China. pp. 1718-1721.

ROSE, PHILIP. 2012. "The likelihood ratio goes to Monte Carlo: the effect of reference sample size on likelihood ratio estimates". Paper presented at the *UNSW Forensic Speech Science Conference.* Sydney, Australia. 3 December 2012.

ROSE, PHILIP. 2013. "More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends". *International Journal of Speech, Language and the Law* 20(1): 77-116.

ROSE, PHILIP, KINOSHITA, YUKO & ALDERMAN, TONY. 2006. "Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/". In: P. Warren & C. I. Watson (eds) *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*. Canberra: ASSTA. pp. 1941-1944.

ROSE, PHILIP & MORRISON, GEOFFREY S. 2009. "A response to the UK Position Statement on forensic speaker comparison". *International Journal of Speech, Language and the Law* 16(1): 139–163.

TABACHNICK, BARBARA. G. & FIDDELL, LINDA. S. 2007. *Using Multivariate Statistics* (5th edition). Boston: Pearson.

WACKERLY, DENNIS D., MENDENHALL III, WILLIAM & SCHEAFFER, RICHARD L. 2008. *Mathematical Statistics with Applications* (7th edition). London: Thomson.

WANG, Z. X. & GUO, D. R. 1989. *Special functions.* London: World Scientific.

WELCH, B. L. 1947. "The generalization of student's problem when several different population variances are involved." *Biometrika* 34(1/2): 28–35.

Vincent Hughes
Department of Language and Linguistic Science
University of York
Heslington
York
YO10 5DD
United Kingdom
email: vh503@york.ac.uk

Erica Gold
Department of Language and Linguistic Science
University of York
Heslington
York
YO10 5DD
United Kingdom
email: erica.gold@york.ac.uk

Ashley Brereton
Department of Mathematical Sciences
University of Liverpool
Liverpool
L69 3BX
United Kingdom
email: a.brereton@liverpool.ac.uk

LATE TALKING TODDLERS: RELATING EARLY PHONOLOGICAL DEVELOPMENT TO LATER LANGUAGE ADVANCE

MARILYN M. VIHMAN[1], TAMAR KEREN-PORTNOY[1], CHRISTOPHER WHITAKER[2], AMY BIDGOOD[3] & MICHELLE MCGILLION[4]

[1]University of York, [2]Bangor University, [3]University of Liverpool, [4]University of Sheffield

Abstract

Background. Expressive Late talkers are identified as children with an unusually small productive vocabulary for their age, in the absence of any other known neurological, sensory or cognitive deficit. Their lexical delay has been found to be associated with phonetic delay.

Aims. The two primary goals of this study are (1) to provide intensive analyses of phonetic and phonological characteristics of late talkers (LTs) at the end of the single-word period as a basis for comparing their speech with that of typically developing children (TDs), not at the same age but at the same developmental point; (2) to compare the relative phonological, lexical, morphological and syntactic advance of the same two groups 14 months later, based on analysis of spontaneous language use, and to relate this advance to phonetic and phonological resources at the earlier measurement point.

Methods and procedures. Time 1 analyses included assessment of volubility, size of consonant inventory, percent consonants correct and extent of consonant variegation and of the use of selected prosodic patterns or 'templates'. Time 2 analyses assessed advance in phonology (percent consonants correct), lexicon (diversity of verb and function word types), morphology (provision of obligatory morphemes) and syntax (MLU, IPSyn).

Outcomes and results. Although three groups differing in age at achieving Time 1 lexical criterion were identified (TDs, LTs and 'transitional' LTs or TLTs), there is little evidence of group differences in other measures of linguistic advance at either sampling point, when the groups are compared at the same lexical level. Exploratory statistical analyses using Canonical Correlations revealed that a combination of high age at Time 1, small consonant inventory and low phonetic variegation are strong predictors of low accuracy in consonant use and relatively poor lexicon, morphology and syntax at Time 2, while dependence on a limited set of phonological patterns at Time 1 was significantly correlated more specifically with slower morphological advance at Time 2.

Conclusions and implications. The study found that, once the groups are equated for lexical level, the linguistic skills of LTs as a group are not distinguishable from those of TDs in either the early period of phonological development or the year following the end of the single-word period. Nevertheless, based on the relation of Time 1 to Time 2 measures within individuals, the study also demonstrates that phonetic and phonological knowledge and skills constitute a key foundation for later linguistic advance, as regards grammar as well as phonology.

1. Introduction

Expressive Late talkers (hereafter, LTs) are identified as children with an unusually small productive vocabulary for their age alongside normal comprehension skills, with no other known neurological, sensory or cognitive deficit (Desmarais, Sylvestre, Bairatti & Rouleau 2008). Expressive LTs are commonly defined, based on parental report, as children producing fewer than 50 identifiable words at about age two and few if any word

combinations (e.g., Carson, Klee, Carson & Hime 2003; Mirak & Rescorla 1998; Paul & Jennings 1992).

A search for ways of predicting the later language development of late talkers has been the focus of a good many studies over the past 25 years. In an attempt to identify group differences on measures other than vocabulary these studies are generally based on comparison with typically developing children (TDs) who are matched in age to the LTs (Carson et al. 2003; Mirak and Rescorla 1998; Paul & Jennings 1992; Thal, Reilly, Seibert, Jeffries & Fenson 2004 also include a second age-matched at-risk group, children with unilateral focal brain injury). A central question for this research is whether late talkers follow a typical developmental trajectory once they start on regular word use, or whether their development is instead qualitatively different from that of TDs (Baird Pharr, Ratner & Rescorla 2000). One as yet unexplored way to tackle this question is to compare LTs to TDs at a *single lexical level*, and then to follow all of them over the same length of time, to answer these questions: Given a common initial point of lexical comparison, (1) do LTs display similar characteristics to their TD peers, or different ones – specifically, in terms of phonetics or phonology? (Thal, Oroz & McCaw 1995, included language- as well as age-matched pairings with their LTs but no follow-up analysis.) (2) Do LTs and TDs continue to advance at a similar or a different pace in phonology, morphology and syntax?

The primary goal of this study is to fill this gap by providing intensive analyses of phonetic and phonological characteristics of the speech of TDs and LTs at the same developmental point – namely, the end of the single-word period, when children typically have a cumulative production lexicon of some 50-100 words – and then comparing the relative phonological, lexical, morphological and syntactic advance of these children 14 months later, based on analysis of spontaneous language use, in the home, at both time points. In addition, by keeping lexicon size constant for the two groups at the first assessment point, the study can explore the effect of any phonetic or phonological difficulty or delay that may have accompanied the initial lexical delay on linguistic advance over the period in which the foundations of language are generally established.

Although LTs are usually identified by a lexical criterion, they are reportedly delayed on several phonetic measures within the single word period as well as in morphology and syntax at a later age (Carson et al. 2003; Paul & Jennings 1992; Rescorla, Mirak & Singh 2000; Thal et al. 1995). LTs are also reported to be much less voluble than TDs (Baird Pharr et al. 2000). The phonetic characteristics of LTs' productions have been reported to resemble those of much younger typically developing children: In comparison with their age-matched peers their vocalizations (both words and babble) are based more exclusively on vowels, syllabic consonants, and glides and glottals than on 'true' consonants (in addition to glottal stop and /h/, glides and syllabic consonants are considered not to have fully consonant-like supraglottal closure) and they produce few closed syllables and few consonant clusters (Baird Pharr et al. 2000; Carson et al. 2003; note that all of these studies are based on children acquiring English). Similar findings have been reported for LTs' early word forms when babble is excluded, especially for the more severely delayed children (Thal et al. 1995). Baird Pharr et al. (2000) reported that the syllable structures of LTs at age three are similar to those of typically developing children at age two. Thus delay in lexical development seems to be accompanied by delay in phonetic and phonological development.

Limited vocal practice or atypical babbling is often considered a likely source of delay in word learning (Baird Pharr et al. 2000; Paul & Jennings 1992; Rescorla & Ratner 1996; Stoel-Gammon, 1989): Practice with production through babble can be taken to bootstrap the ability to match target sounds to articulatory patterns to gain control over sound production

(DePaolis, Vihman & Keren-Portnoy 2011; Majorano, Vihman & DePaolis 2013). Thus, a delay in the development of phonetic and phonological structure may, in some children, reflect their failure to avail themselves of vocal practice as a route to phonetic facility.

The reasons for LTs' reportedly reduced vocal practice are unknown but could include low vocal or social orientation, for example, or slow overall neurological development (Locke 1994; Rescorla & Ratner 1996). Whatever the cause, limited vocal practice reduces opportunities for experiencing a match between input speech forms and one's own production, lessening the 'self-reward' that partially motivates vocal practice (Oller 2000) and possibly leading to lexical delay. However, once word production is well established, if phonetic difficulties persist, for whatever reason, these will have further repercussions not only on the articulation and planning of vocalizations but also on the child's ability to attend to and represent speech sequences (Keren-Portnoy et al. 2010; McCune & Vihman 2001). Phonetic limitations may thus serve as a significant bottle-neck in early language development, potentially limiting linguistic advance in other domains and inhibiting word learning both before and after first word use is observed.

Phonological delay may also signal other, more general cognitive difficulties. Children with Specific Language Impairment (SLI) or a family history of dyslexia often prove to have been LTs (Bishop & Snowling 2004; Thal et al. 2004). Bishop & Snowling (2004) claim that both disorders involve phonological deficits, but that while dyslexic children may have no other deficit, children with (classic) SLI have other linguistic deficits as well. Beckman, Munson & Edwards (2007) claim more specifically that children with SLI have problems abstracting phonological categories from the finer-grained representations of words; somewhat similarly, Dodd & McIntosh (2010) emphasize 'rule abstraction skills' as a corollary of phonological accuracy in two-year-olds. Accordingly, if phonological delay is a sign of failure to abstract phonological rules or categories, LTs may later show delays in other linguistic domains, due to their failure to generalize from the items they have acquired. The abstraction referred to here can be achieved through distributional analysis not only on the speech stream itself but also on the learned lexicon (Pierrehumbert 2003). Thus, both phonetic (practice-related) and phonological (abstraction-related) difficulties may lead to language delay.

What could be construed as evidence for success or failure in phonological abstraction? One long-standing approach suggests that phonological development involves a non-linear course, in which children's word shapes do not, at first, show consistent progress toward greater accuracy in matching the adult targets but instead follow a U-shaped developmental path. That is, children's earliest word forms are typically relatively accurate (Ferguson & Farwell 1975; Menn & Vihman 2011). Once a few (context-limited, or situationally primed) words have been produced (Vihman & McCune 1994), distributional learning typically applies to the new database constituted by a child's first word forms, leading to the development of word production patterns specific to the child ('word templates': Vihman & Croft 2007; Vihman & Keren-Portnoy 2013). This first step in systematization is often manifested as an apparent regression: Although some of the later-learned words reflect the child's patterns by exhibiting relatively accurate forms in relation to adult words 'selected' for the match, others ('adapted' words: Vihman & Velleman 2000) involve deviations from the adult form, sometimes radical (i.e., involving metathesis, consonant harmony, truncation of syllables or inclusion of child-specific 'favourite' segments or sequences), as the child assimilates the target word to whatever production patterns have become the most familiar. This account accords with models such as that of Pierrehumbert (2003), who distinguishes token-based, fine-grained (relatively accurate) representations from type-based, coarser-grained representations (which reflect phonological generalization, expressed in pattern or template

formation; see also Beckman et al. 2007). Thus a deficit in procedural leaning might result in a failure to develop 'favourite phonological patterns' or templates.

This study was designed to test three hypotheses, using comparisons based on language at a given lexical development point, rather than on age:

LTs will have a distinct developmental profile from that of TDs even when measured at the same developmental lexical point, in terms of phonetics and phonology.

The effects of either weak phonetic or weak phonological resources, or both, on later learning can be tested even if this first hypothesis is not supported – i.e., if weakness in these two domains is not found to persist following lexical delay. Our two further hypotheses posit roles for both phonetics and phonology in supporting lexical advance; they are thus neither interdependent nor in competition:

Children who exhibit slow early lexical advance, and who may have fewer *phonetic and/or phonological resources*, will advance more slowly in grammar and lexicon as well as in phonology over the year of language use that follows the single-word period.

Those LTs whose speech in the single word period exhibits little initial evidence of *phonological systematicity* (based on distributional learning or systematization, resulting in phonological templates [see below]) will show delay at a later age and thus be at potential risk for SLI.

2. Method

2.1 Participants

LTs were recruited through newspaper advertisements and brochures left in nurseries and doctors' offices, in North Wales and in York, England. Interested parents were asked to fill out the Oxford Communicative Developmental Inventory (CDI: Hamilton, Plunkett, & Schafer 2000) when their child reached age two; children producing fewer than 50 words at that point and few if any word combinations were retained for follow-up. An initial home visit was followed by phone contact until the child was reported to be producing some 50 words. We then made frequent home recordings until the child produced at least 25 different words spontaneously in a 30-minute session, including onomatopoeia with a conventional adult target form (e.g., *meow* or *boom* but not 'pig snort'). This '25-word-point' (25wp) session, which reflects a cumulative vocabulary of 50-75 words (Vihman & Miller 1988), serves as our Time 1.

Our control sample of TDs was made up of 12 children seen as part of an earlier longitudinal study in North Wales (Keren-Portnoy et al. 2010), recorded at least once a month from 11 months on through their 25wp. At 30 months of age all of the children were screened using the Bayley Scales of Infant and Toddler Development III (Bayley, 2006) and the Reynell Developmental Language Scales III (Edwards et al., 1997). We sought to arrive at two groups free of evident cognitive difficulty and with age-appropriate comprehension. The inclusion criterion was Bayley screening test scores reflecting 'competence in age-appropriate tasks' or 'emerging age-appropriate skills' on all but the Expressive Communication subtest, on which three LTs scored 'at risk for developmental delay'. All participants retained in the study scored within 4 months of chronological age on the Reynell-III Receptive Scale; the LTs scored 4.5 months or more below chronological age on the Reynell-III Expressive Scale. No hearing difficulties were reported at the time of the recordings analysed here.

Of the 12 children in the original TD sample, one failed to reach criterion on the Reynell-III Receptive Scale and was excluded, leaving a sample of 11 TDs. Out of 47 potential LTs (seen either in North Wales or in York), 26 were excluded for the following reasons: too advanced when first seen (10), below criterion on Receptive Scales (9), family relocation, fostering away, serious behavioural problems, untestable on Reynell-III (1 each), loss of contact (2), and bilingual (heavy use of Welsh at 25wp: 1). Note that although our TDs (and some of the LTs and TLTs) were recruited in North Wales, none were being raised as bilinguals. Six of the TDs' families used Welsh occasionally and two of the LTs' families sometimes used Afrikaans, but we observed few non-English words in our recordings.

Of the 21 children followed as potential LTs, 10 passed expressive criteria as TDs when tested on the Reynell-III. We retained them for follow-up at Time 2 as 'transitional LTs' (TLT), or 'late bloomers' (Mirak & Rescorla 1998). The remaining 11 children made up the LT group (cf. the three similarly defined groups in Carson et al., 2003.) In the final sample the age ranges of the three groups at the 25wp were largely distinct (see Table 1). No TDs overlap in age at 25wp with any LTs. Thus the division into groups based on performance on the Reynell-III was strongly correlated with age at the 25wp. Additional details on the participants are given in Table 1, which shows that the LT group had parents with a somewhat lower education level and a higher proportion of birth by C-section, ear infections and learning difficulties in the extended family.

Group: N	N male	N 1st-born	Mean age (range) / mos @ 25wp	Mean word tokens	Mean word combos	Parental education: one or more parents with university degree	Difficult or premature birth	Ear infection or gluey ear	Learning difficulties in extended family	Age range / Reynell express. Test	Reynell recep. score	Reynell express. score
TD: 11	6	9	21.08 (15-26)	123.34	1.1	55%	27% C-section, none premature	9%	11%	3;0.11-3;1.18	3;00-5;5	2;9-3;11
TLT: 10	6	1	26.8 (25-28)	117.73	2.6	50%	no C-section, 30% premature	0	0	2;5.28-2;6.21	2;5-3;1	2;2-2;9
LT: 11	8	6	31.64 (27-36)	118	2.8	27%	36% C-section, none premature	27%	18%	2;5.8-2;11.13	2;3-2;11	<1;9-2;4

Table 1. Participant characteristics, by group

2.2 Procedures

2.2.1 Time 1 (T1)

We recorded the children in their homes for 30 minutes of free play interaction with a familiar caregiver. Once the recording had been phonetically transcribed, word candidates were identified and tested for word status, following Vihman & McCune (1994). Imitations were included in some analyses (see below), but were not counted in establishing the 25wp. (One TD later found to have produced only 24 identifiable words was nevertheless retained in the study.)

2.2.2 Time 2 (T2)

We again recorded each of the children at home in similarly naturalistic interaction with a familiar caregiver (and the same observer, in most cases) 14 (+/-2) months after the child's 25wp. These recordings were transcribed orthographically, with phonetic transcription wherever the child's utterance deviated from the expected adult form; one member of the team later re-checked all T2 transcripts against the original recordings.

3. Analyses

3.1 Time 1

3.1.1 The database

For comparison with previous LT studies we established relative *volubility* based on all word tokens used, including imitations and conventional onomatopoeia but excluding unidentifiable vocalizations (i.e., babbling or jargon). Word tokens were included only if the target was unambiguous, without overlapping sound or noise (e.g. from toys, other speech etc.). All usable word tokens were included in the analyses that follow unless otherwise indicated.

3.1.2 Phonetic measures

Two variables were used to measure articulatory resources and planning, *consonant inventory* and within-word *consonant variegation*. In addition, we included counts of (word-final) *codas*, since previous studies have reported a tendency for LTs to use fewer codas (Rescorla & Ratner, 1996).

3.1.2.1 Consonant inventory (CInv)

For this analysis sample size was found to have a significant influence; accordingly, we restricted our sample to 25 word types, the maximum available for all the children. When necessary we sampled words randomly for exclusion until we reached 25 words. We disregarded imitated words and used only one phonetic token per word type, retaining (i) the child form used most frequently, wherever possible, or (ii) the most adult-like form.

For each child we tallied all the consonants for which up to two examples could be found, in any word position, excluding consonants used in only one word type. Non-English consonants or consonants used only in substitution for other target sounds were included, to obtain as complete picture of the child's phonetic resources as possible. For example, TD Ali had only three stops: [p] as onset consonant only in substitution for /b/: *back* [pæʲk], *bed* [pat] but as coda in *pump* [bɔ:p] and both [t] and [b] in onset position only: *two* [tɯ:], *stay* [tæˑɪ:] and *bed* [bad], *byebye* [babei].

3.1.2.2 Consonant variegation (CVar)

We assessed consonant variegation within each word token, extending Stoel-Gammon's (1989) Mean Babbling Level analysis (cf. also Paul & Jennings 1992). The analysis distinguishes three levels (disregarding differences in voicing): (1) no true consonant; (2) no more than one true consonant type; (3) more than one true consonant type. The final CVar score is based on mean level per word.

3.1.2.3 Coda use

We tallied the proportion of all word tokens with a coda in any syllable (counting no more than one coda per token).

3.1.3 Phonological measures

We carried out two analyses designed to establish the extent to which the child was, on the one hand, *accommodating* to the adult phonology or, on the other hand, *assimilating* adult word forms to an emergent phonological system, based on child-specific patterns.

3.1.3.1 Percentage Consonants Correct-Revised (PCC-T1)

We calculated number of consonants correct in relation to the target consonants (in the typical local adult pronunciation, with due account taken of common adult variants, such as glottaling of /t/ or stopping of /ð/, as well as of phrases typically 'run together' by adults, such as /ɪnˈiə/ for *in here*, where only /n/ would be required) for all tokens produced in the T1 session, based on Shriberg et al. (1997). All but five children produced over 80 usable vocalizations (mean, 123; range, 57 to 266); the full sample was used for each child. Added or misordered consonants were not counted as errors in this analysis, designed to identify *correct articulation*; phonemic changes were scored as errors but deviant pronunciations resulting in subphonemic change (e.g., [ç] for /s/, /ʃ/) were not.

3.1.3.2 Quantifying pattern use (Pscore)

To test the idea that some LTs might make less than expected use of child-specific word-shape patterns or phonological templates (as described in Vihman & Croft 2007) as a manifestation of an underlying deficit in procedural learning we developed a pattern use score ('Pscore'). We first established the main *generic* prosodic structures used, based on number and order of C[onsonant] and V[owel] slots (CV, CVC, CVCV…). To these we added patterns *specified* by particular manner of consonant (e.g., CVC with fricative or nasal

specified for the coda slot) or particular vowel or diphthong or consonantal sequence. Table 2 shows the patterns included, ordered by total uses in our sample, and the scores that we assigned them. We categorized all of the words analyzed for each child in terms of (1) overall prosodic structure and (2) whether or not they could be considered 'accurate', within the child's production limits ('selected': e.g., *bubbles* [bʌbɔ:]), or were instead 'adapted' by the child to fit into the category (e.g., *Harvey* [bæbi]). For this measure just one token of each word type was included, whether imitated or spontaneous. To be included in a child's Pscore a category had to account for at least 10% of that child's words; 'specified' categories were identified wherever possible, so that, for example, a child who produced 10% or more of their words with a [Vu] pattern (e.g., Jeremy, Table 1) would be scored as having a 'specified' Vu pattern for each of those words, while any remaining [Vi] words (if fewer than 10%) would be scored as fitting a generic VV pattern.

The goal of our analysis is to define idiosyncratic pattern use (often construed as emergent systematicity) on a gradient scale, with words adapted to 'specified' patterns providing the strongest indicators of such use, whereas words based only on a generic pattern and closely adhering to the adult target form (selected) provide little such evidence. Accordingly, Pscore points were assigned as follows: 1 point to 'selected' words that fit generic patterns (e.g., *no* [nəʊ]); 2 points to 'selected' words that fit *specified* patterns, including C_1VC_1V, or consonant harmony, as seen in *bubbles*, above; 2 points to word forms *adapted* to fit a generic category (e.g., *boat* [bəʊ:], produced by a child who sometimes produces codas); 3 points to words *adapted* to fit a *specified* category (e.g., *Harvey*, above). To arrive at a single score for each child we average the child's individual word scores, yielding *mean score per word*. The Pscore assigns more points to forms that are (a) more constrained and (b) less accurate and is thus not only a measure of pattern formation, generalization or systematization, but also an error score, potentially indicating the extent to which a child's output is guided by individual limitations.

word shape	uses	child	group	select: gloss	select: form	adapt: gloss	adapt: form	select score	adapt score
CVV	250	Colin	TLT	*no*	nəu	*boat*	bəʊː	1	2
C1VC1V	169	Lewis	LT	*bubbles*	bʌbɔː	*Harvey*	bæbiː	2	3
C1VC2	161	Jennifer	TD	*pink*	pɪkʰ			1	2
CV	114	Jude	TD	*sky*	kʰa	*flower*	la	1	2
C1VC2V	81	Ian	TD	*monkey*	maŋgi			1	2
CV[fricative]	73	Elise	LT	*eyes*	aɪs	*bike*	baɪs	2	3
VCV	64	Julie	TLT	*(there) we go*	iːgəu	*jigsaw*	iʃɔː	1	2
CVCVC	53	Ian	TD	*break it*	bəɪkət			1	2
CVu	41	Jeremy	LT	*go*	gəuː	*two*	təuː	2	3
CV[nasal]	38	Jane	TLT	*green*	giːn	*camera*	kʰæm	2	3
CV[glide]V	28	Jack	TLT			*strawberries*	dauːwi	2	3
Reduplication	22	Liam	LT	*nana*	nɑːna	*spot*	babɑː	2	3
Three or more syllables	19	Holly	T	*boing boing*	paipaipʔai	*This*	diʒidiʒa	1	2
CVʔ	18	Carl	TLT	*foot*	ɛfʰʌʔ	*Thomas*	ətɑʔ	1	2
C1VC1	18	Clarissa	LT	*cake*	gɛkʰ	*Dog*	gag̃kʰ	2	3
C1VC1VC	15	Dean	TLT	*saucers*	tʌtaːs	*Carrots*	kłækłaːs	2	3
CVCV[fricative]	13	Elise	TD			*rabbit(i)*	haʔpiːsː	2	3
V[glide]V	12	Tony	LT	*away*	awɛɪː	*Flowers*	aːwe	2	3
C	7	Andy	TD	*sh*	sː	*Moo*	m̩ːː	1	2
CVi	7	Tomos	T	*bye*	baːɪ	*Plane*	pɛɪ	1	2

CVCi	4	Sylvia	TD	*potty*	pɔti	*Beer*	bi:di	2	3
CVCVi	4	Owen	TD	*marmite*	məˈmai	*Harri*	ɦəwaɪ	2	3
CV[fricative]V	3	Andy	TD	*what's that*	wʌsæʔ	*Tortoise*	dəza	2	3

Table 2. Phonological patterns and scoring

Note: Each pattern is illustrated by forms produced by the highest user of that pattern.

3.2 Time 2

3.2.1 Percentage Consonants Correct-Revised (PCC-T2)

For the phonological analysis we included as much of each T2 transcript as needed to arrive at 100 word tokens (Shriberg et al. 1997). The session-time needed for this ranged from 11 to 32 minutes; for three children fewer than 80 different words were available. For the analysis of T2 transcripts, which are largely made up of connected speech, segments were again counted as correct where deviations reflected features frequently found in British English (omission of initial /h/, replacement of final /ŋ/ by /n/ or final /t/ by glottal stop).

3.2.2 MLU and IPSyn

Mean Length of utterance in words (MLUw) was based on 100 fully intelligible utterances (BROWN, 1973). The Index of Productive Syntax (IPSyn; Scarborough, 1990) was used to evaluate grammatical complexity. IPSyn considers 56 syntactic and morphological structures across four subscales (noun phrases, verb phrases, questions/negations and sentence structures).

3.2.3 Morphology

To assess morphological advance we established the percentage of morphemes supplied in obligatory syntactic or discourse contexts. For example, auxiliary *be* is omitted in "sweeties goin' in the custard" (LT pretending to pour contents of teacup into a bowl). Following Brown (1973), we initially included present progressive *-ing*, past irregular and regular *–ed* and third-person singular present indicative *–(e)s*, plural *–(e)s*, the articles *a* and *the* and both contractible and uncontractible uses of copula and auxiliary *be*. However, one speech sample per child provided sufficient obligatory contexts to calculate child provision of only five morphemes – *present progressive, articles (a, the), plural, contractible copula* and *contractible auxiliary*, each of which had over 10 obligatory contexts of use in all cases. Of these, two were excluded because they were at ceiling (90% use or better) for 13 (present progressive) or 21 children (plural). Of the three morphemes frequent and variable enough to justify analysis – articles, contractible copula and auxiliary *be* – identifiable opportunities for use were far more frequent for articles than for *be*. Accordingly, we combined the average percent use in obligatory context of each of the three morphemes into a single *morphology* score.

3.2.4 Lexicon

Lexical level was based on a count of all the verb and function word types used in the 100-utterance samples for each child. Diversity of verb use is often included among evaluations of children with SLI; the function words serve as a useful type-based complement to the token-based morphology score.

3.3 Reliability

T1 reliability was checked by comparing the two main transcribers' phonetic transcriptions of 3 minutes' recording from each of five children (1 TD, 2 TLTs, 2 LTs; 93 utterances). Agreement on consonant identity averaged 81%. To assess Pscore reliability a colleague trained in the procedure independently scored 3 children in each group, giving a correlation of .85 with the first author.

For T2 we compared number and identity of morphemes as well as agreement on consonant identity for the two main transcribers, based on 3 minutes' transcription from each of four children (2 TLT, 2 LT). Agreement on number of morphemes reached 95%, on identification of morphemes, 87%. Agreement regarding consonant identity was again 81%.

4. Results

4.1 Group comparisons

In order to test our first hypothesis, we compared the three groups on the T1 measures, expecting to find the LT scores to be the lowest and the TD scores the highest (see Table 3). We ran Kruskal-Wallis tests to compare the three groups on the full range of T1 measures, including age at 25wp. (These tests were chosen because of the small sample sizes, to avoid having to assume normality. However, we also ran ANOVAs: the only additional comparison to show a significant group difference – before but not after applying the Bonferroni correction to control for running seven different comparisons – was CInv.) Only age at 25wp was significantly different across groups: $\chi^2(2) = 26.176$, $p < .001$, even after applying the Bonferroni correction.

	LT mean (s.d.)	**TLT mean (s.d.)**	**TD mean (s.d.)**	**Overall mean (s.d.)**
age at 25 wp (in mos.)	31.53 (2.9)	27.25 (1.1)	21.24 (2.9)	26.66 (5.0)
volubility	133.82 (63.3)	118.00 (31.8)	117.73 (48.1)	123.34 (48.9)
CInv	6.36 (1.6)	7.90 (1.3)	7.09 (1.0)	7.09 (1.4)
% codas	21.45 (24.1)	21.90 (17.5)	29.40 (19.4)	30.43 (28.5)
CVar	2.02 (.2)	2.13 (.3)	2.16 (.2)	2.10 (.2)
PCC-T1	.53 (.2)	.56 (.1)	.65 (.1)	.58 (.1)
Pscore	1.60 (.4)	1.53 (.3)	1.37 (.2)	1.50 (.3)

Table 3. Time 1 measures

Post-hoc Mann-Whitney tests revealed significant differences in age at 25wp between each pair of groups (LT vs. TLT: $U = 5$, $p < .001$, LT vs. TD: $U = 0$, $p < .001$, TLT vs. TD: $U = 1$, $p < .001$, significant after Bonferroni correction). The strength of the age measure as a group distinguisher was expected, given that the group categorization was based on the Reynell-III, which includes a lexicon size assessment, also the basis for 'age at 25wp'.

Table 4 provides a descriptive picture of the T2 measures. The Kruskal-Wallis tests revealed no significant group differences at T2 (ANOVAs were also run, as for the T1 variables; none showed significant differences).

	LT mean (s.d.)	TLT mean (s.d.)	TD mean (s.d.)	Overall mean (s.d.)
PCC-T2	.64 (.2)	.71 (.1)	.73 (.1)	.69 (.1)
MLUw	2.72 (.9)	2.67 (.7)	2.66 (.65)	2.69 (.7)
% article use	52.81 (34.4)	64.43 (25.4)	70.80 (18.6)	62.57 (27.3)
% contractible copula use	63.05 (32.9)	66.09 (23.67)	74.19 (20.2)	67.89 (25.9)
% contractible auxiliary use	53.87 (40.1)	57.32 (36.1)	48.81 (34.1)	53.08 (35.9)
IPSyn	58.73 (16.7)	65.44 (12.3)	65.45 (12.8)	63.06 (14.1)
verbs	14.64 (4.2)	15.44 (8.0)	16.73 (4.8)	15.61 (5.6)
function words	19.09 (7.4)	22.11 (7.9)	23.09 (6.1)	21.39 (7.1)

Table 4. Time 2 measures

Note. PCC = percent consonants correct; MLUw = mean length of utterance in words; IPSyn = Index of Productive Syntax.

Thus when early- and later-learning children are compared on the basis of a lexically-determined sampling point we find few group differences, despite the large differences in the children's age ranges between groups (cf. Thal et al. 1995). Given this outcome, we treat age as a continuous rather than a categorical variable in the remaining analyses. This does not override the children's group allocation, since the categorization by Reynell-III scores was tightly aligned with age at 25wp. Treating age at 25wp as a continuous variable makes finer-grained distinctions possible between children who reached that lexical milestone at different ages, beyond their group assignment. We thus add age at 25wp to the other T1 predictors in testing hypotheses 2 and 3, alongside the phonetic and phonological measures calculated on the T1 samples.

4.2 Selecting variables for analyses of the relation between T1 and T2 measures

We first had to reduce the number of strongly inter-correlated measures at each time point. Percent *coda use* is highly correlated with both *CInv* ($r = .36$, $p < .05$, two-tailed) and *CVar* ($r = .78$, $p < .01$, two-tailed) – not surprisingly, as coda consonants provide an opportunity for a child to increase inventory size and to vary within-word consonant use. We thus excluded coda use from the set of T1 variables. No other two T1 variables were significantly intercorrelated.

At T2 the lexico-syntactic scores (lexicon, MLUw and IPSyn) are all highly intercorrelated (above .8). We therefore combined them into a single composite *lexico-syntax* score, arrived at by Principal Component Analysis (PCA). The first PCA dimension accounted for 89% of the information. Even after creating the single lexico-syntax score the resultant three T2 measures (morphology, lexico-syntax and PCC-T2) remained correlated to some extent: PCC-T2 with morphology: $r = .66$, $p < .01$, two-tailed, and PCC-T2 with lexico-syntax: r =

.57, $p < .01$, two-tailed; lexico-syntax and morphology just miss being significantly correlated: $r = .35$, $p = .05$, two-tailed).

4.3 Relating T2 attainments to T1 skills

To test Hypotheses 2 and 3, which pertain to the ability of T1 phonetic and phonological measures to predict T2 advance, we ran a canonical correlation analysis (Manly 2005). This analysis is appropriate in an exploratory study in which the predicted variables are correlated to a degree, making independent regressions inappropriate, and in which the sample size is relatively small. The canonical correlation analysis is a way to calculate a linear combination of the predictor variables (here, the resultant variable will be called Canonical Variate for T1, or CVT1) with the predicted variables (here, Canonical Variate for T2, or CVT2). These summary variables are chosen with the goal of *maximizing the linear correlation between them.* In other words, the analysis looks for the strongest possible relation between the two sampling points for all participants. Thus, the relation between the first set of CVT1 and CVT2 is chosen to be maximally high. Potential additional sets represent different, and independent, dimensions of the relation between predictor and predicted variables. Each successive set of canonical variates is chosen so as to be maximally correlated with one another, but to be independent of previous sets. The first set, having the highest correlation, is the most important (Manly 2005).

The canonical correlation analysis was used to test both Hypotheses 2 and 3. Hypothesis 2 is corroborated if a relation is found between age at T1 and phonetic (or phonological) measures, on the one hand, and T2 grammatical attainment, on the other. For Hypothesis 3 to be corroborated a more specific relation needs to be apparent – namely, between a high Pscore at T1 (indicating successful procedural learning) and faster advance by T2. The analysis included, for T1, lexical advance (age at 25wp), volubility, and phonetic (CInv and CVar) and phonological measures (PCC-T1 and Pscore). For T2 we included three distinct types of linguistic knowledge: Phonology (PCC-T2), morphology and lexico-syntax.

The analysis produced two sets of statistically significant canonical correlations (Table 4; canonical correlation I [CCI]: $\chi^2 = 40.00$, $df = 18$, $p < .01$; CCII: $\chi^2 = 19.31$, $df = 10$, $p < .05$). Interpretation of the canonical variates (Manly, 2005) is best achieved by looking at the correlations between each variate and the variables that make it up. For CC-I the T1 summary variate (CVT1-I) is most strongly associated with small consonant inventory, high age at 25wp and low variegation (Figure 1). It is also associated, to a lesser degree, with high volubility and low accuracy (PCC-T1), but not with pattern use. CVT1-I explains 25% of the variability at Time 1 and 34% of the variability at Time 2. CVT2-I is most strongly associated with low accuracy (PCC-T2) and, to a lesser extent, low morphology and low lexico-syntax scores. CVT2-I explains 60% of the variability at Time 2. CVT1-I and CVT-2-I are strongly linearly related ($r = .75$). The summary variate (CVT1-II) for CC-II (Figure 2) is most strongly associated with high use of patterns (Pscore), moderately with low accuracy (PCC-T1) and weakly with volubility; it is unrelated to any other T1 measure. CVT2-II is strongly associated with low morphology but with no other T2 measure. CVT1-II explains 18% T1 variability while CVT2-II explains 20% T2 variability. CVT1-II and CVT1-II have a strong linear relation ($r = .66$).

Canonical correlation	**Canonical Variate for T1 (CVT1)**			**Canonical Variate for T2 (CVT2)**		
	T1 variable	*coefficient*	*r*	*T2 variable*	*coefficient*	*r*
I	Age at 25wp	0.30	.63	Lexico-syntax	-0.16	-.67
	Volubility	0.45	.44	Morphology	0.08	-.61
	CVar	-0.26	-.54	PCC-T2	-0.95	-.99
	PCC-T1	-0.39	-.44			
	CInv	-0.48	-.65			
	Pscore	-0.11	.11			
	T1 variable	*coefficient*	*r*	*T2 variable*	*coefficient*	*r*
II	Age at 25wp	-0.41	-.06	Lexico-syntax	-0.14	-.09
	Volubility	-0.12	-.30	Morphology	-1.33	-.78
	CVar	-0.10	-.09	PCC-T2	0.91	-.05
	PCC-T1	-0.09	-.37			
	CInv	-0.18	-.17			
	Pscore	0.96	.90			

Table 5: Standardized canonical coefficients and correlations for the two canonical variates and the original variables to which they correspond

Note. Each canonical variate is the sum of the relevant standardized variables, each multiplied by the relevant coefficient

Figure 1: Correlations between each variate and the variables that make it up (CC-I).

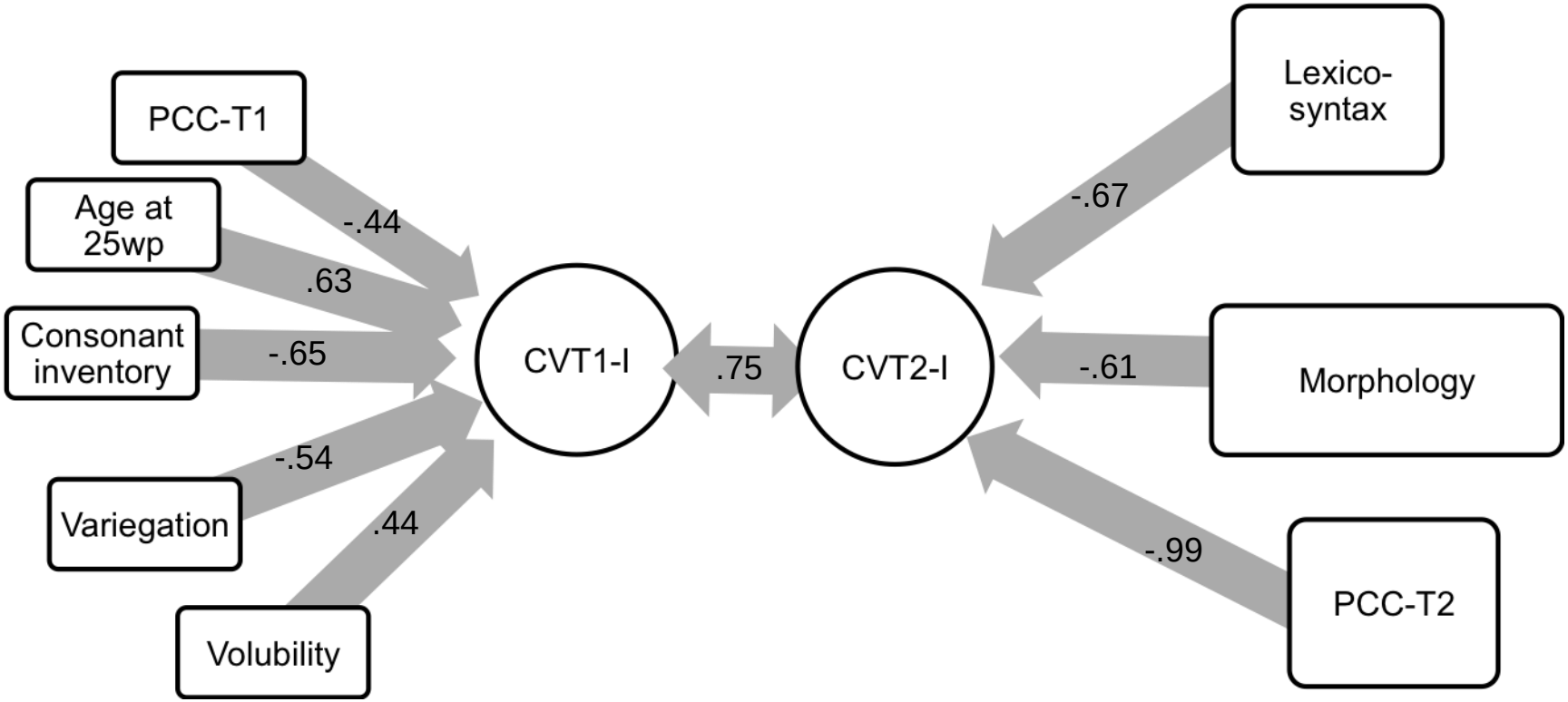
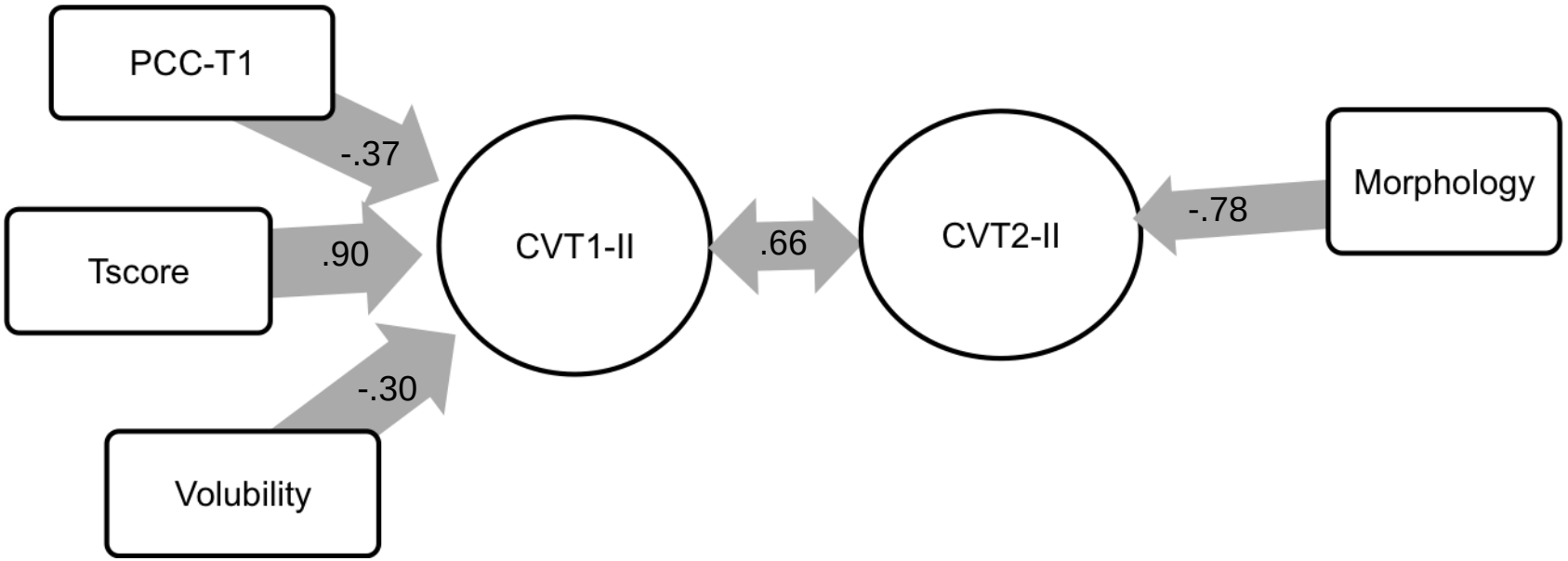


Figure 2: Correlations between each variate and the variables that make it up (CC-II).



5. Discussion

Unlike most previous LT studies, this study compared TDs and LTs – and, in the event, also TLTs – at a single developmental point, to investigate the predictive value of lexical, phonetic and phonological variables, measured at a time when all three groups had established a good starting vocabulary, for various domains of later linguistic advance. We refer to the T1 session as 'the end of the single word period', since the children are on verge of word combination (a mean of 1.1 [TDs], 2.6 [TLTs] and 2.8 [LTs] two-word or longer combinations occurred in this session – the increasing numbers of word combinations with age at 25wp perhaps reflecting the older children's longer experience processing language receptively at the point when they began actively making use of words in production). Note that the older TLTs and LTs also have a larger *receptive* vocabulary at T1 than the younger TDs, as all of our participants scored as age-appropriate on the Reynell-III receptive scales.

When measured at a comparable lexical level we found no evidence for weaker phonetics or phonology in any group relative to the others. This in itself is an important finding, which points to LTs being delayed without being 'deviant' or qualitatively different from TDs (or TLTs). Thus, reaching the end of the single-word period at an older age is not on its own enough to indicate a likelihood of continued delay in language development. However, as shown by the first canonical correlation, the combination of lexical delay with weak phonetic skills (small consonant inventory, low consonant variegation and, to a lesser degree, low accuracy) is related to low phonological accuracy, morphology and lexico-syntax a year later. The second canonical correlation revealed a more specific relation, regardless of age at 25wp, between non-adult like phonology (high Pscore and low accuracy) and low morphology a year later.

Although our samples are relatively small and the study accordingly exploratory, the findings suggest a direct relation between early delay accompanied by weak phonetic resources and slower advances not only in accurate word production but also in morphology and lexico-syntax, perhaps because phonetic resources are a necessary tool for achieving stable knowledge of both content and function morphemes, whether affixes or free morphemes. This recalls an earlier comment of Mirak & Rescorla (1998), that consonant diversity in production 'affects, or is affected by, the overall amount of lexical production, with the direction of this relationship being unclear' (p. 15). That is, phonetic and phonological advance are inextricably bound up with word learning, and word learning is likely a critical foundation for advances in grammar.

An unexpected finding is the relation between non-adult like phonology (persistent reliance on idiosyncratic patterns or templates) and morphology: This finding goes directly against our third hypothesis, according to which pattern use, despite its corollary, inaccuracy in reproduction of the adult target form, should be a positive indicator. We assumed that pattern use reflects emergent systematicity, which we take to be a normal concomitant of phonological advance, with its characteristic U-shaped curve, as well as evidence for successful procedural learning. Instead, high reliance on limited patterns, along with low accuracy (and, to a much lesser extent, high volubility, a paradoxical finding in this context), proved to be a strong predictor of poor T2 morphology.

However, pattern use can be taken to reflect output constraints as well as systematicity. As such, dependence on templates, or holistic matches to target, may sometimes be a useful 'holding strategy' while in other instances proving a negative indicator (see Velleman & Vihman, 2002). Note that in the current study high pattern use and late talking tended to be related: The LT group had the highest Pscores; a t-test comparing mean LT to TD Pscores

was marginally significant: p = .09, two-tailed). In fact, an extended period of holistic phonology could be paralleled by an extended period of holistic (i.e., not yet truly productive) morphology (Vihman, 1982). Note that successful acquisition of morphology requires, for some functional morphemes, prior analysis of words into stem and affix. Thus the Pscore appears to offer a usable predictor of later outcomes for morphology, although this is not what we had anticipated.

Our finding that pattern use is a *negative* indicator for LTs indicates that one or more of our original assumptions was mistaken. It is possible that our scoring failed to capture extent of pattern use, or that such use does not tap procedural memory after all, or that LTs who go on to present with SLI do not actually have deficient capacities for procedural memory. We believe that our measure does correspond relatively well to what has been described in the literature on templates, given that qualitative analyses of the data (cf., e.g., Vihman, DePaolis & Keren-Portnoy 2009) seem to correspond well to that which we arrived at here quantitatively. On the other hand, our attempt to meaningfully capture and compare pattern or template use at a single developmental point was likely misconceived: Previous studies suggest that such use may fluctuate over time and, most importantly, that the timing of its emergence and period of sway differs by child (Vihman & Keren Portnoy 2013).

How can we explain the fact that our T1 phonological measures predict morphology more strongly than they predict lexico-syntax, although both reflect grammatical knowledge? Our lexico-syntax score assesses aspects of grammatical advance such as, for example, provision of arguments, complex noun and verb phrases and function words and the development of a diverse vocabulary of verbs and function words. These aspects of language involve neither attention to sublexical speech units nor a requirement to match the phonetic details of target forms. In contrast, the morphology score reflects children's ability to match their production to highly specific target requirements; this is the ability that is captured, in opposite ways, by the two T1 phonological measures, accuracy (which requires attention to detail) and pattern use (which involves disregard for detail). This can explain the predictive relation between phonology at T1 and morphology at T2. A critical mediating factor is likely to be 'phonological memory', or the ability to find in one's articulatory repertoire the best segmental sequence to correctly reproduce a whole adult target, which requires flexible and well-practiced articulatory skills, good planning capacity *and* the ability to integrate perceptual experience with these skills (Keren-Portnoy et al. 2010; Jones , Macken & Nicholls 2004).

The canonical correlation analysis gives a clear picture of a scenario in which lack of accuracy with poor phonetic resources and slower initial lexical advance jointly predict less rapid later linguistic advance. In addition, although to a lesser extent, pattern use is specifically related to weaker morphological advance. While the fact that weak phonetic resources are related to less successful advance is not surprising, the finding regarding pattern use is: Templates are often found in the corpora of typically developing children; until lately they have been discussed only in the context of typical development (e.g., Vihman & Croft 2007), and have seldom been seen as a risk factor (but see Velleman & Vihman 2002). Our results suggest, however, that dependence on a routinized output, and a global rather than analytic approach to the matching of target forms, can be problematic when it comes to morphology. Clearly, more research is needed to understand the role of pattern or template use in development – what its benefits and its costs may be for later development.

In conclusion, timeliness in reaching the end of the single-word period in combination with adequate phonetic resources, on the one hand, and relative accuracy in word form production and freedom from routinized output, on the other hand, proved the strongest predictors of

satisfactory linguistic advance over a year later. More broadly, it seems clear that phonetic and phonological skills, practice, and knowledge provide a critical foundation for later morphological and syntactic knowledge and use.

References

BAIRD PHARR, A., RATNER, N. B. & RESCORLA, L., 2000, Syllable structure development of toddlers with expressive specific language impairment. *Applied Psycholinguistics*, 21, 429-449.

BAYLEY, N., 2006, Bayley Scales of Infant and Toddler Development (3rd ed.): Screening test manual (San Antonio, TX: Harcourt Assessment).

BECKMAN, M. E., MUNSON, B., & EDWARDS, J., 2007. "Vocabulary growth and the developmental expansion of types of phonological knowledge". In J. Cole & J. I. Hualde (eds.), *Laboratory Phonology, 9* (Berlin: Mouton de Gruyter).

BISHOP, D. V. M. & SNOWLING, M. J., 2004, Developmental dyslexia and Specific Language Impairment. *Psychological Bulletin*, 130, 858-886.

BROWN, R., 1973, *A First Language.* (Cambridge, MA: Harvard University Press).

CARSON, P. C, KLEE, T., CARSON, D. K. & HIME, L. K., 2003, Phonological profiles of 2-year-olds with delayed language development: Predicting clinical outcomes at age 3. *American Journal of Speech-Language Pathology*, 12, 28-39.

DEPAOLIS, R. A., VIHMAN, M. M., & KEREN-PORTNOY, T., 2011, Do production patterns influence the processing of speech in prelinguistic infants? *Infant Behavior and Development*, 34, 590-601.

DESMARAIS, C., SYLVESTRE, A., MEYER, F., BAIRATI, I. & ROULEAU, N., 2008, Systematic review of the literature on characteristics of late-talking toddlers. *International Journal of Language and Communication Disorders*, 43, 361-389.

DODD, B. & MCINTOSH, B., 2010, Two-year-old phonology: Impact of input, motor and cognitive abilities on development. *Journal of Child Language*, 37, 1027-1046.

EDWARDS, S., FLETCHER, P., GARMAN, M. HUGHES, A., LETTS, C., & SINKA, I., 1997. *Reynell Developmental Language Scales III: The University of Reading Edition* (London: NFER-Nelson).

FERGUSON, C. A. & FARWELL, C. B., 1975, Words and sounds in early language acquisition. *Language*, 51, 419-439.

HAMILTON, A., PLUNKETT, K. & SCHAFER, G., 2000, Infant vocabulary development assessed with a British CDI. *Journal of Child Language*, 27, 689-705.

JONES, D. M., MACKEN, W. J. & NICOLLS, A. P., 2004, The phonological store of working memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30, 656-674.

KEREN-PORTNOY, T., VIHMAN, M. M., DEPAOLIS, R., WHITAKER, C. & WILLIAMS, N. A., 2010, The role of vocal practice in constructing phonological working memory. *Journal of Speech, Language, and Hearing Research*, 53, 1280-1293.

LOCKE, J. L., 1994, Gradual emergence of developmental language disorders. *Journal of Speech, Language, and Hearing Research*, 37, 608-616.

MANLY, B. F. J., 2005, *Multivariate statistical methods: a primer* (3rd ed.). (Boca Raton: Chapman and Hall/CRC).

MCCUNE, L. & VIHMAN, M. M., 2001, Early phonetic and lexical development. *Journal of Speech, Language, and Hearing Research*, 44, 670-684.

MENN, L. & VIHMAN, M. M., 2011, "Features in child phonology: inherent, emergent, or artefacts of analysis?". In N. Clements & R. Ridouane (eds.), *Where Do Phonological Features Come From? The nature and sources of phonological primitives.* (Amsterdam: John Benjamins).

MIRAK, J. & RESCORLA, L., 1998, Phonetic skills and vocabulary size in late talkers: Concurrent and predictive relationships. *Applied Psycholinguistics,* 19, 1-17.

OLLER, D. K., 2000, *The Emergence of the Speech Capacity.* (Mahwah, NJ: Erlbaum).

PAUL, R. & JENNINGS, P., 1992, Phonological behavior in toddlers with slow expressive language development. *Journal of Speech and Hearing Research,* 35, 99-107.

PIERREHUMBERT, J. B., 2003, Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech,* 46, 115-154.

RESCORLA, L. & RATNER, N. B., 1996, Phonetic profiles of toddlers with Specific Expressive Language Impairment (SLI-E). *Journal of Speech, Language, and Hearing Research,* 39, 153-165.

RESCORLA, L., MIRAK, J., & SINGH, L., 2000, Vocabulary growth in late talkers. *Journal of Child Language,* 27, 293-311.

SCARBOROUGH, H., 1990, Index of Productive Syntax. *Applied Psycholinguistics*, 11, 1-12.

SHRIBERG, L. D., AUSTIN, D., LEWIS, B. A., MCSWEENY, J. L. & WILSON, D. L., 1997, The Percentage of Consonants Correct (PCC) Metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research,* 40, 708-722.

STOEL-GAMMON, C., 1989, *Prespeech and early speech development of two late talkers. First Language*

THAL, D. J., OROZ, M., & MCCAW, V., 1995, Phonological and lexical development in normal and late-talking toddlers. *Applied Psycholinguistics*, 16, 407-424.

THAL, D. J., REILLY, J., SEIBERT, L. JEFFRIES, R. & FENSON, J., 2004, Language development in children at risk for language impairment: Cross-population comparisons. *Brain and Language,* 88, 167-179.

VELLEMAN, S. L. & VIHMAN, M. M., 2002, Whole-word phonology and templates. *Language, Speech and Hearing Services in Schools,* 33, 9-23.

VIHMAN, M. M., 1982, The acquisition of morphology by a bilingual child: A whole-word approach. *Applied Psycholinguistics*, 3, 141-160.

VIHMAN, M. M. & CROFT, W., 2007, Phonological development: Toward a 'radical' templatic phonology. *Linguistics,* 45, 683-725.

VIHMAN, M. M., DEPAOLIS, R. A. & KEREN-PORTNOY, T., 2009. "A Dynamic Systems approach to babbling and words". In E. Bavin (ed.), *Handbook of Child Language.* (Cambridge: Cambridge University Press).

VIHMAN, M. M. & KEREN-PORTNOY, T. (EDS.), 2013, *The Emergence of Phonology.* (Cambridge: Cambridge University Press).

VIHMAN, M. M. & MCCUNE L., 1994, When is a word a word*? Journal of Child Language,* 21, 517-542.

VIHMAN, M. M. & MILLER, R., 1988. "Words and babble at the threshold of language acquisition". In M. D. Smith & J. L. Locke (eds.), *The Emergent Lexicon.* (New York: Academic Press)

Marilyn M. Vihman
Department of Language and Linguistic Science
University of York
Heslington
York
YO10 5DD
United Kingdom
tel. +44 1904 433612
email: mv509@york.ac.uk

OBJECT PREPOSING IN LATE ARCHAIC CHINESE

AIQING WANG

University of York

Abstract

This paper investigates the preverbal positioning of objects in Late Archaic Chinese (5th-3rdc BC; "LAC"). As an SVO language, LAC permits DP objects to front into preverbal positions in a medial domain below TP and above *v*P. Based on the relative ordering of preposed non-*wh*-constituents and negation, two positions can be found: a high position and a low position. The high position for the fronting of non-*wh*-objects displays topic properties, while the low position displays focus features. Nominal and pronominal objects in LAC may occur in either position; all preposed constituents occupy a specifier node of functional projections (Paul 2002, 2005), followed by an optional fronting marker as the head of the relevant functional categories. Within the medial domain, head-like elements are always in a fixed relative order: negatives precede modals of ability, and follow other modals. In terms of *wh*-DPs, D-linked *which*-phrases appear in the topic position, whereas non-D-linked *wh*-phrases are permitted exclusively in an extra (focus) position between the topic position and negation, triggered by the Intervention Effect of negation (Kim 2002). This paper also explores the underlying structure of *wh*-P.

1. Introduction

Archaic Chinese during the Warring States period (5th-3rdc BC) is referred to as Late Archaic Chinese (LAC).[1] Texts in LAC display predominant SVO word order, with objects appearing in a postverbal position. Examples[2] in (1) involve nominal and pronominal non-*wh*-objects.

(1) a. 齊 人 伐 燕。 (孟子•公孫醜 4thc BC)
Qi ren fa Yan.
Qi person attack Yan
"People of the State of Qi attacked the State of Yan."
b. 晉 人 用 之。 (國語•楚語上 5thc BC)
Jin ren yong zhi.

[1] I follow Chou (1963) and Wang (1958) in terming Classical Chinese during the Warring States period as Late Archaic Chinese (LAC), which exhibits distinctive characteristics. I also agree that around the Han Dynasty (2ndc BC-2ndc AD) after the Warring States period, there was a transitional period with multiple typological changes (Xu 2006).

[2] The primary sources of this paper are the Peking University corpus, the Academia Sinica electronic database, and the Sheffield Corpus of Chinese. The selected texts of these corpora are all received, representing a wide range of writing found in various time periods. In the LAC period, the corpora cover more than twenty key books written by different authors.

Jin person employ 3.Obj
"People of the State of Jin employed him."

However, there are contexts in which nominal and pronominal objects appear preverbally in the low TP-internal domain, as shown in (2a-b). Fronted objects in this paper are shown in boldface.

(2) a. 吾 百姓 之 不 圖 (國語•越語下 5thc BC)
wu **baixing** zhi bu [VP tu $t_{baixing}$]
I common.people ZHI not care.about
"I did not care about common people"
b. 若 子 不 我 信 (國語•楚語下 5thc BC)
ruo zi bu **wo** [VP xin t_{wo}]
if you not me trust
"if you do not trust me"

Moreover, when the object is a *wh*-phrase, it must front to a position below the subject and above *v*P, because LAC was a *wh*-fronting language,[3] and *wh*-in-situ did not emerge in Chinese until the Han Dynasty (2ndc BC-2ndc AD) (Aldridge 2010, Feng 1996). Examples (3a) and (3b) illustrate that both bare *wh*-words and complex *wh*-phrases move to a preverbal position when acting as direct objects. In (3b), the noun "city" is modified by a *wh*-operator *he* "what", and they form the only phrase preceding the *v*P. Example (3c), together with (3a-b), illustrates that both indirect and direct *wh*-objects raise to a preverbal position.

(3) a. 何 不 爲 乎?[4] (莊子•秋水 4thc BC)
He bu [VP wei t_{he}] hu?
what not do Q
"What do (I) not do?"[5]
b. 宋 何 役 之 不 會,
Song [**he** **yi**] zhi bu [VP hui $t_{he\ yi}$],
Song what battle ZHI not enter
而 何 盟 之 不 同? (左傳•昭公二十五年 5thc BC)
er [**he** **meng**] zhi bu [VP tong $t_{he\ meng}$]?
Conj what alliance ZHI not join
"What battle the State of Song does not enter, and what alliance (it) does not join?"

[3] Although LAC is a wh-fronting language, wh-in-situ is obligatory for the second complement in a double object construction, as shown in (i). Moreover, this is the only exception to the requirement of wh-preposing in LAC.
(i) 國 謂 君 何? (左傳•僖公十五年 5thc BC)
Guo wei jun he?
state call lord what
"What does the state call the lord?"
[4] As pointed out by an anonymous reviewer, wh-questions like (3a) tend to be exclamatives due to the clause-final interrogative particle 乎 hu.
[5] In this paper, the omitted constituents are recovered based on contextual information

c. 寡人 將 誰 屬 國?[6]
Guaren jiang **shui** [VP shu guo t_{shui}]?
I will who entrust state
"To whom will I entrust the state?"

It is worth mentioning that in LAC, objects may undergo both long- and short-distance movement. The examples in (2-3) demonstrate the short-distance raising of objects, while in examples (4a-b), non-*wh*- and *wh*-DPs undergo long-distance movement. As the object of an embedded verb, the pronoun 女 *ru* "you" in (4a) and *wh*-word 誰 *shui* "who" in (4b) move across a nonfinite complement clause boundary to a higher node.

(4) a. 餘 不 女 忍 殺 (左傳•昭公元年 5th c BC)
yu bu **ru** ren [VP sha t_{ru}]
I not you bear kill
"I cannot bear to kill you"

b. 公 誰 欲 與? (莊子•徐無鬼 4th c BC)
Gong **shui** yu [VP yu t_{shui}]?
Your.Majesty who want entrust
"To whom does Your Majesty want to entrust (the country)?"

Notwithstanding examples (2-4) which exhibit preverbal objects, the observations here support the view that LAC has always been an SVO language (Aldridge 2011, 2012a, Djamouri 2005, Djamouri and Paul 2009, Djamouri et al 2012, Meisterernst 2010, Peyraube 1996), so object preposing is derived, and should not be assumed as the vestige of basic OV word order, as proposed by Li and Thompson (1974), Wang (1958), La Polla (1993), Feng (1996), Xu (2006), among others.

I analyse the preverbal positioning of non-*wh*-DP objects in LAC, and propose two landing sites based on the relative order between negation and preposed elements. In terms of *wh*-DPs, they move to an extra position between the high position and negation. This paper is organised into three main sections. Section 2 introduces the preposing of non-*wh*-objects in LAC, including two landing sites, fronting markers, as well as the medial domain. Section 3 investigates the nature of two positions of non-*wh*-fronting and proposes that the high position displays topic properties, while the low position displays focus properties. Section 4 discusses *wh*-fronting, including the Intervention Effect of negation (Kim 2002) that triggers further

[6] This example involves a ditransitive verb 屬 shu "entrust". In LAC, there are three ways of packaging arguments in ditransitive constructions. The first approach is a disposal construction yi-DP-V-wh "yi state entrust who". Since this approach will generate a reverse DP-V order with that in (3c), it is ruled out. The second method is to place both arguments in postverbal positions, with the latter (the wh-word) being packaged as a PP: V-DP-P-wh "entrust state to who". This structure satisfies the V-DP word order of (3c), but it would entail an ellipsis of preposition. Besides, unlike NPs, wh-phrases cannot act as a prepositional complement in the form of P-wh due to the obligatory wh-fronting during the period of LAC. Therefore, the only option for packaging arguments of this ditransitive is a double object construction: V-wh-DP "entrust who state". In (3c), the wh-word shui moves from its unmarked base position to a preverbal position, generating the surface structure.

wh-movement to an extra position, and the underlying structure of *wh*-P.

2. Preposing of Non-Wh-Objects

In this section I focus on the preverbal positioning of non-*wh*-DP objects and propose two landing sites for object preposing based on the relative ordering of preposed non-*wh*-elements and negatives. Both positions allow nominal and pronominal objects, all of which occupy a specifier node of some functional projection (Paul 2002, 2005), accompanied by an optional fronting marker as the head of relevant functional categories. In addition, I explore the medial domain by illustrating intervening negative/modal elements and their relative order.

2.1. Two Positions for Non-Wh-Fronting

Previous research treats preverbal positioning of DP objects in LAC as focalisation, and states that the syntactic focus movement of *wh*-phrases only targets a node above negation while below modals (Aldridge 2006, 2010). Nevertheless, there must be two landing sites for non-*wh*-fronting between TP and *v*P, and evidence can be derived from the relative ordering of preposed constituents and negation to invalidate any approach involving one single projection. The relative ordering between preposed non-*wh*-DP objects and negation serves as the evidence that the landing sites of object fronting cannot be accounted for by an approach involving one single projection. Fronted nominal objects may precede or follow a negator, as illustrated by (2a), repeated here as (5a), and (5b) respectively. These examples are extracted from texts of distinct authors in the same period, so this fact proves that such positional discrepancy is not a diachronic feature.[7] 不 *bu* "not" in these examples is a neutral clausal negator simply denying the situation without affecting the aspect or mode. It is worth mentioning that NPs in these instances are accompanied by fronting markers ZHI (5a) and SHI (5b) respectively; their distribution and individual nature will be discussed in the next subsection.

(5) a. 吾 百姓 之 不 圖 (國語•越語下 5thc BC)
wu **baixing** zhi bu [VP tu $t_{baixing}$]
I common.people ZHI not care.about
"I did not care about common people"
b. 敢[8] 不 唯[9] 命 是 聽?[10] (左傳•昭公十二年 5thc BC)

[7] Examples (2a) and (2b) that are extracted from the same book exhibit positional asymmetry: the NP in (2a) precedes the negator, whereas the pronoun in (2b) follows negation. So this fact indicates that such positional discrepancy in LAC is not author-specific either.

[8] The debatable nature of 敢 gan "dare" is beyond the research scope of this paper, so I simply treat it as a verb.

[9] The semantic values and nature of this morpheme will be presented in the next subsection.

[10] Edith Aldridge (p.c.) has pointed out that example (5b) has a very different structure from (5a), because the negator and the NP do not occupy the same minimal clause. She also presented an instance where negation appears in the lower clause following 之 ZHI, as in (ii).

(ii) 不 唯 下 土 之 不 康靖。 (國語•吳語 5thc BC)
Bu wei xia tu zhi bu kangjing.

Gan bu wei **ming** shi [VP ting t_{ming}]?
dare not WEI order SHI listen
"How dare (I) not listen to orders only? (It is only orders (I) must follow.)"

Similarly, this observation of bi-positional non-*wh*-fronting also applies to pronouns: preposed pronominal objects may appear before or after negation. In (6a), a demonstrative pronoun 斯 *si* "this" precedes the aspectual negator 未 *wei* "not yet; not at all, never", but its minimal pair counterpart 之 *zhi* moves to a position lower than the same negator, as in (6b). As can be seen from (6b), apart from being a fronting marker (see (5a) and (6a)), 之 ZHI can also act as a fronted personal pronoun. It is notable that negatives usually "trigger" raising of pronouns, so pronoun fronting in the context of negation is prevalent in LAC. Pronouns raised into the preverbal position are predominantly, but not exclusively, negated by 未 *wei*.

(6) a. 吾 斯[11] 之 未 能 信。 (論語•公冶長 5th c BC)
Wu **si** zhi wei neng [VP xin t_{si}].
I this ZHI not.yet can be.confident
"I have not been able to be confident in this."
b. 未 之 能 行 (論語•公冶長 5th c BC)
wei **zhi** neng [VP xing t_{zhi}]
not.yet 3.Obj can execute
"before (he) can execute it"

Therefore, these facts suggest that LAC entails two landing sites for non-*wh*-fronting in the medial domain between TP and *v*P, with negation intervening in between. The high position is above negation, whereas the low position is below negation; evidence comes from the relative ordering between fronted non-*wh*-DPs and the negator.

2.2. Fronting Markers

In this subsection, I explore fronting markers that follow fronted non-*wh*-DPs in both positions. I show fronting marker 之 ZHI and 是 SHI exhibit discrepant properties and discriminating positional distribution.

As mentioned earlier, NPs in examples (5a-b) are accompanied by fronting makers ZHI and SHI respectively. The fronting maker ZHI can follow preposed non-*wh*-DPs either in the high position, or in the low position, accompanied by a matrix predicate 唯 WEI "be (the one

---

not WEI under land ZHI not peaceful
"It is not only the world that is not peaceful."

Nevertheless, I presume non-wh-DPs being in a lower clause does not deny the fact that they are below negation; in other words, the presence of clause boundaries does not affect the relative order between negation and preposed elements. With respect to (ii), I argue that this example concerns a subject focus-type cleft reading, and ZHI in this sentence is not a fronting marker. Since (ii) is irrelevant to object preposing, it is not a counterexample to my proposal.

[11] According to contextual information, the demonstrative 斯 si in (6a) is non-human, indicating the action of becoming a government official.

who/that)" that indicates assertive modality (Djamouri 2001, Meisterernst 2010) to form a cleft structure WEI … ZHI. Alternatively, ZHI may combine with a negative copula 非 FEI "not be" to form another cleft FEI … ZHI in the low position. ZHI never occurs together with the matrix predicate WEI or the negative copula FEI in the higher position (see (5a) and (6a)), but when ZHI appears in the lower position, cleft structures WEI/FEI … ZHI are obligatory (7a/b). As can be seen from (7b), in addition to being a fronting marker (see (5b)), 是 SHI can also act as a demonstrative pronoun and form a minimal pair with 斯 *si* "this".

(7) a. 将 不 唯 卫国 之 败 (左傳•成公十四年 5thc BC)
jiang bu wei **weiguo** zhi [VP bai t_{weiguo}]
will not WEI State.of.Wei ZHI ruin
"it is not only the State of Wei (he) will ruin"
b. 是 詩 也, 非 是 之 謂 也 (孟子•萬章上 4thc BC)
shi shi ye, fei **shi** zhi [VP wei t_{shi}] ye
this poem NMLZ FEI this ZHI interpret Decl
"this poem, is not interpreted as this"

Dissimilar to ZHI that is permitted in ether position, the fronting marker SHI, however, is restricted in the low position,[12] as in (5b). Furthermore, when SHI is employed as a fronting marker, it may combine with the matrix predicate 唯 WEI to constitute a cleft structure WEI … SHI (5b), or combine with the negative copula 非 FEI to form FEI … SHI (8). I hypothesise WEI … ZHI and WEI … SHI are underlying the same cleft construction, only with disparate fronting markers; this presumption also applies to FEI … ZHI/SHI. Nevertheless, although WEI … ZHI/SHI can be negated, and its negative form occupies the identical position with FEI … ZHI/SHI, the semantic value of its negative form is different from that of FEI … ZHI/SHI: the former conveys the meaning "it is not only who/that", whereas the latter means "it is not who/that".

(8) 今 王 非 越 是 圖 (國語•吳語 5thc BC)
jin wang fei **yue** shi [VP tu t_{yue}]
now emperor FEI Yue SHI contrive
"now it is not the State of Yue the emperor contrives"

It is noteworthy that fronting markers in both positions are optional, as illustrated by instances (9a) and (2b) (repeated as (9b)) respectively.

(9) a. 祀 不 可以 已 乎? (國語•楚語下 5thc BC)
Si bu keyi [VP yi t_{si}] hu?
propitiation not can cease Q
"Cannot (I) cease the propitiation?"
b. 若 子 不 我 信 (國語•楚語下 5thc BC)

[12] The claim that the fronting marker SHI occurs exclusively in the lower position is supported by lack of attested counterexamples.

ruo zi bu **wo** [VP xin t_{wo}]
if you not me trust
"if you do not trust me"

Both WEI and FEI are considered as predicates, because they can be marked by modal verbs (10a) and modified by adverbs (10b). In addition, WEI can be negated by the clausal negator 不 *bu* "not" (see (5b) and (7a)).

(10) a. 將 唯 命 是 從 (左傳•昭公十二年 5th c BC)
jiang wei **ming** shi [VP cong t_{ming}]
will WEI order SHI follow
"it is only the orders (they) will follow"
b. 君 今 非 王室 不 平安 是 憂[13]
jun jin fei [**wangshi bu pingan**] shi [VP you $t_{wangshi\ bu\ pingan}$]
Your.Majesty now FEI monarchy not peaceful SHI worry
"now it is not the monarchy being not peaceful that Your Majesty worries about"
(國語•吳語 5th c BC)

The distribution of fronting markers ZHI and SHI in the high and low positions for object fronting is hence shown in Table 1:

	High position		**Low position**	
之 **ZHI**	ZHI (5a) & (6a)	/ (9a)	WEI … ZHI (7a) FEI … ZHI (7b)	/ (9b)
是 **SHI**	*		WEI … SHI (5b) FEI … SHI (8) & (10b)	/ (9b)

Table 1: Distribution of ZHI and SHI

2.3. Landing Site of Non-Wh-Fronting

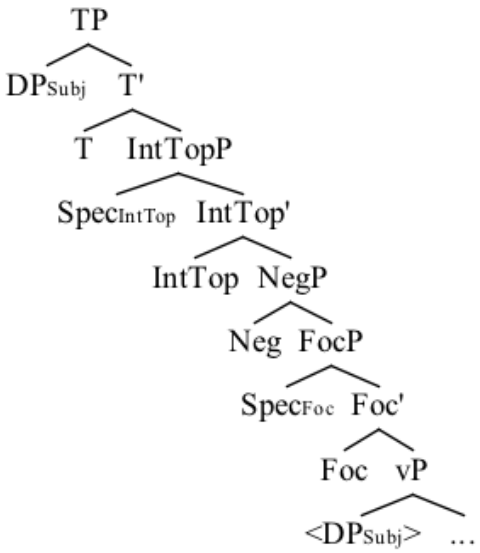
In terms of the landing site of object fronting, I adopt (and extend) the theory that preposed non-*wh*-DPs occupy a specifier position of some functional category below TP and above *v*P, instead of being adjoined to *v*P directly (Paul 2002, 2005). For one thing, supposing the analysis of object preposing targeting the edge of *v*P (Aldridge 2010, 2012a) is adopted, it will imply a single position for fronted non-*wh*-DPs, contrary to the above-mentioned instances entailing two preverbal positions. For another, fronting markers ZHI and SHI also lend further support for the proposal involving functional categories. As can be observed from the attested data (5a/b), ZHI/SHI is always immediately preceded by a preposed non-*wh*-DP, which means a single position on the edge of *v*P cannot accommodate these two constituents. Providing the

[13] Apart from the negative copula 非 FEI, this example contains an additional negative element 不 bu. However, this negator bu is in the lower nominalised clause, so it cannot be treated in the same way as negatives that determine the two positions for preposed non-wh-DPs.

assumption concerning functional projections is adopted, then fronted elements can occupy the specifier node, while fronting markers may act as the head of corresponding functional projections.

Therefore, I follow the ideas in Paul (2002, 2005) about different positions between TP and *v*P in modern Mandarin Chinese, and further extend them by proposing two distinct positions of object preposing, both of which are specifiers of functional projections. Consequently, the basic structure of object preposing in LAC is as follows:

(11)



As can be seen in (11), I posit both the high and low landing sites for the preposing of non-*wh*-objects are in the lower TP domain; the high position is above NegP, whereas the low position is below negation. For the time being, the corresponding projections for the high and low positions are referred to as IntTopP and FocP respectively, with IntTopP dominating FocP. Evidence of the nature of two landing sites will be discussed in Section 3.

2.4. Medial Domain

LAC allows medial elements between the subject and the verb in a canonical clause. These elements include auxiliaries,[14] adverbs, negatives and modals. Among auxiliaries, only passive markers can intervene between the subject and the verb, as shown in (12).

(12) a. 今 兄弟 被 侵 (韓非子•五蠹 3rdc BC)
jin xiongdi bei qin
now brother PASS attack

[14] In Chinese, tense, aspect and voice are not reflected in the morphology of the verb, so they have to be expressed by auxiliaries, which only take VPs a as complement and probably derive from full verbs through grammaticalisation (Peyraube 1999, Meisterernst 2008a).

"now brothers are attacked"

b. 吾 長 見 笑 于 大方 之 家。 (莊子•秋水 4thc BC)
Wu chang jian xiao yu dafang zhi jia.
I often PASS laugh by enlightened GEN sage
"I am often teased by enlightened sages."

Apart from acting as medial elements following immediately after the subject (12b), adverbials may precede the subject in canonical clauses, as illustrated in (12a). I argue in LAC, temporal or conjunctive adverbs are analogous to conjunctions, thus preceding the subject.

In the context of object preposing, however, only adverbials, negatives and modals can act as medial elements between the subject and *v*P. Since passives and object preposing are in a complementary distribution, passive markers never act as medial elements for object fronting. In terms of adverbials, they may precede the subject in the context of object preposing, as shown in (8), repeated as (13a). In addition to this, as illustrated by (13b), adverbs may also precede the modal verb (see 今 *jin*), or intervene between the preposed DP object and the verb (see 小 *xiao*). Given adverbials appear either in a relatively high or low position, they are not discussed as medial elements in this paper. Consequently, I only investigate negative/modal elements in the medial domain of object preposing.

(13) a. 今 王 非 越 是 圖 (國語•吳語 5thc BC)
jin wang fei **yue** shi [VP tu t_{yue}]
now emperor FEI Yue SHI contrive
"now it is not the State of Yue the emperor contrives"
b. 今 將 惠 以 小 賜 (國語•魯語上 5thc BC)
jin jiang **hui** yi xiao [VP ci t_{hui}]
now will benefaction YI fractionally grant
"now (you) will fractionally grant benefactions"

Both negatives and modal verbs are the head-like elements intervening in the medial domain between the subject and *v*P. These medial elements do not necessarily appear together, but if they do, they are always in a fixed relative order: negatives must precede modals of ability, yet follow other modals. As for fronted non-*wh*-DPs, they always appear immediately next to the negator, in an anterior or posterior position. Consequently, preposed non-*wh*-objects in the high position intervene between modals other than those of ability and negation, while fronted constituents in the low position follow negation and precede modals of ability. A linear format of the clausal positions and the medial elements is in (14):

(14) Interim version:

Subject > Other modals > <u>High position</u> > Negation > <u>Low position</u> > Modals of ability > *v*P

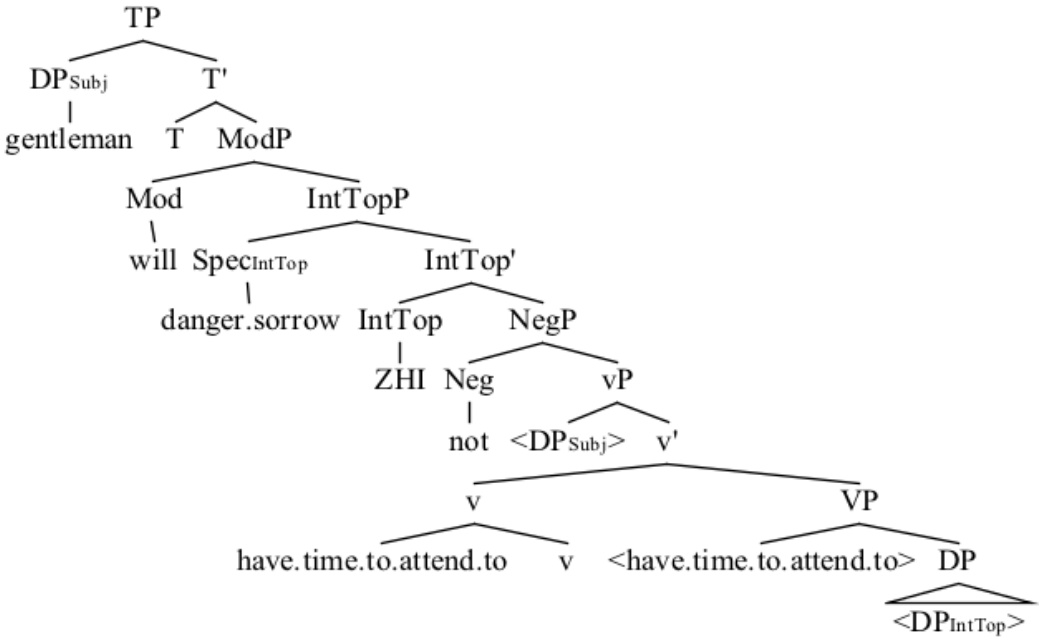
In example (15a), the fronted NP *xianai* "danger and sorrow" raises to a preverbal position above the negator *bu*, so this sentence involves the high position. The modal 將 *jiang* "will"

precedes the preposed NP in the high position, thus preceding the negative as well. Example (6b), repeated as (15b), involves the low position, in that the fronted pronoun is under the aspectual negator 未 *wei*; the modal of ability 能 *neng* "can" follows the preposed pronoun. Since I posit preposed objects are specifiers of functional categories, DPs in (15a) and (15b) occupy $Spec_{IntTop}$ and $Spec_{Foc}$ nodes respectively (see 15c/d).

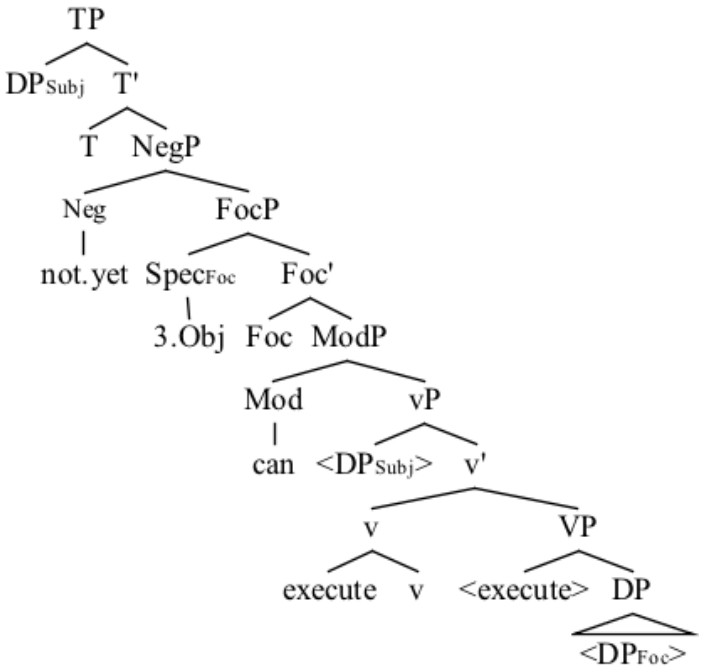
(15) a. 君子 將 險哀 之 不 暇 (國語•週語下 5th c BC)
junzi jiang **xianai** zhi bu [VP xia t_{xianai}]
gentleman will danger.sorrow ZHI not have.time.to.attend.to
"gentlemen will not have time to attend to the danger and sorrow"

b. 未 之 能 行 (論語•公冶長 5th c BC)
wei **zhi** neng [VP xing t_{zhi}]
not.yet 3.Obj can execute
"before (he) can execute it"

c.



d.



Example (16) demonstrates that in LAC, modals in front of negatives are not restricted to 將 *jiang* "will": this category contains 必 *bi* "must" as well.

(16) 彼 知 吾 將 用 之, 必 不 吾 予 也。[15]
Bi zhi wu jiang yong zhi, bi bu **wu** [VP yu *twu*] ye.
3.Subj know I will employ 3.Obj must not I give Nmlz
"He knows I will employ him, (so he) must not give (him to) me."
(管子 5thc BC)

In LAC, modals of ability entail 可 *ke*, 可以 *keyi* and 能 *neng*, all of which can be translated by "can" in English. I follow Meisterernst (2008a) in treating 可以 *keyi* as a disyllabic verb, rather than analysing 以 *yi* as a stranded preposition, a conjunction or a transitive verb. Moreover, I agree with Meisterernst's analyses (2008a) of 可 *ke* and its variant 可以 *keyi* in Han period Chinese; in LAC, they also predominantly express root possibility values, parallel to 能 *neng*. In LAC, when modal verbs 可 *ke*, 可以 *keyi* and 能 *neng* occur in a negative environment, they may express root possibility values (as shown in (6a) and (15b)), or deontic values (9a).

It is important to point out that sentences involving modals of ability should not be treated as passive constructions, as suggested by Pulleyblank (1995), Meisterernst (2008a), Aldridge (2007, 2010), and others, but as the object preposing construction. According to the approach of passivisation, the theme of the verb is predicted to raise out of the internal argument position to a subject position. Nevertheless, as can be observed from (17a) which involves 可 *ke*, a NP "monarch" occupies the subject position, accordingly, *zuoyou* "attendant" can only be regarded as moving to the object position. Similarly, the subject position of the sentence

[15] Example (16) describes the scenario that the minister of Lu knew the minister of Qi would employ the assassin Guan Zhong who was a potential threat to the State of Lu, so the minister of Lu refused to release Guan Zhong and "give" Guan Zhong to the minister of Qi. This sentence is an utterance of the minister of Qi where the subject is the minister of Lu, while the direct object (null in the second clause) is the assassin Guan Zhong.

involving 能 *neng* (see (6a), repeated as (17b)), is taken by *wu* "I", so the pronominal DP *si* "this" has to front to an object position. To reinforce this point, example (15b), repeated as (17c), is presented, in which the pronoun *zhi* "this" is clearly an internal argument fronted to the preverbal object position, because it is lower than the negator *wei*. So (17c) also helps to show that modals of ability in LAC are not passive markers.

(17) a. 故 人主 左右 不 可 不 慎 也。
Gu renzhu **zuoyou** bu ke bu [VP shen t_{zuoyou}] ye.
so monarch attendant not can not be.wary.of Decl
"So the monarch cannot not be wary of attendants."
(韓非子•說疑 3rd c BC)

b. 吾 斯 之 未 能 信。 (論語•公冶長 5th c BC)
Wu **si** zhi wei neng [VP xin t_{si}].
I this ZHI not.yet can be.confident
"I have not been able to be confident in this."

c. 未 之 能 行 (論語•公冶長 5th c BC)
wei **zhi** neng [VP xing t_{zhi}]
not.yet 3.Obj can execute
"before (he) can execute it"

Along with this, canonical sentences involving modals of ability can be transitive, which lends indirect support to the proposal that this type of sentences are not passive constructions. Examples (18a-c) illustrate clauses with 可 *ke*, 可以 *keyi* and 能 *neng* respectively.

(18) a. 士 何 如 斯 可 謂 之 達 矣?
Shi he ru si ke wei zhi da yi?
scholar what like this can call 3.Obj sagacious Perf
"How can scholars have been called being sagacious?"
(論語•顏淵 5th c BC)

b. 吾 不 可以 僭 之。 (左傳•哀公 5th c BC)
Wu bu keyi jian zhi.
I not can arrogate 3.Obj
"I must not arrogate it."

c. 吾 能 止 之。 (國語•晉語九 5th c BC)
Wu neng zhi zhi.
I can stop 3.Obj
"I can stop it."

Therefore, I take the view that clauses involving modals of ability entail object preposing, analogous to other examples in this paper.

3. Nature of Two Positions for Non-Wh-Fronting

In the previous section, I have demonstrated that there are two landing sites for the preverbal positioning of non-*wh*-DPs between the subject and *v*P. With respect to the nature of these two

positions, the higher position displays topic-like properties, while the lower position is focal; this statement is proved by a comparison between constituents in these two positions.

3.1. Internal Topic Position vs Focus Position

Before discussing the disparate features of the higher and lower positions, their similarity is addressed here: sentences involving both positions are compatible with constructions of a contrastive interpretation. Example (19a) involves the higher position, because preposed DPs in both clauses are above negatives. According to contextual information, this instance shows a scenario that an official Zichang did not even try to relieve the domestic crisis, whereas being busy with accumulating fortune insatiably. Obviously, what the official was supposed to do and his actual behaviour are contrasted with each other. Instance (19b) involves the low position: in the former clause, the fronted NP is lower than negation; while the latter clause contains the fronting marker SHI that occurs exclusively in the low position. (19b) describes a contrastive scenario: when choosing the residence, it was not the mansion itself that was taken as a criterion; instead, people regarded the quality of neighbours as the only criterion. Example (19c) also involves a low position, but it contains pronominal DPs.

(19) a. 是 之 不 卹, 而 蓄聚 不 厭
shi zhi bu [VP xu t_{shi}], er **xuju** bu [VP yan t_{xuju}]
this ZHI not relieve Conj accumulation not be.insatiable.for
"(he) does not relieve this, while is insatiable for accumulation (of fortune)"
(國語•楚語 5thc BC)

b. 非 宅 是 蒠, 唯 鄰 是 蒠。
Fei **zhai** shi [VP bu t_{zhai}], wei **lin** shi [VP bu t_{lin}].
not.be mansion SHI choose WEI neighbour SHI choose
"It is not the mansion (people) choose; it is only the neighbours (people) choose."
(左傳•昭公三年 5thc BC)

c. 我 無 爾 詐, 爾 無 我 虞。
Wo wu **er** [VP zha t_{er}], er wu **wo** [VP yu t_{wo}].
I not you deceive you not I deceive
"I do not deceive you, while you do not deceive me".
(左傳•宣公十五年 5thc BC)

Notwithstanding this common feature, the higher and lower positions exhibit two discriminating properties, which coincide with those of topics and foci respectively.

For one thing, fronted non-*wh*-DP objects in the lower position involve exclusive and exhaustive interpretations, whereas such implication is absent from constituents in the higher position. In (19b), apart from contrastiveness, exclusiveness is also expressed, in that the property of being chosen denoted by the presupposition is not held by the NP "mansion". Additionally, the matrix predicate 唯 WEI "be (the one who/that)" that indicates assertive modality also excludes the NP "mansion", rendering the clefted NP "neighbour" the only

option available. Moreover, elements in the lower position require an exhaustive interpretation. In (20), the matrix predicate WEI determines that the clefted pronoun "you" preceding the fronting marker SHI is the only option available. Parallel to that in (5b), 敢 *gan* accompanied by the negative 不 *bu* conveys the meaning of "dare not but" and implies a modality of obligation.

(20) 敢 不 唯 子 是 從? (左傳•哀公六年 5th c BC)
Gan bu wei **zi** shi [VP cong t_{zi}]?
dare not WEI you SHI follow
"How dare (people) not follow you only? (It is only you (people) must follow.)"

Other examples, not shown here, also demonstrate that only fronted non-*wh*-objects in the low position involve exclusiveness and exhaustivity, whereas preposed non-*wh*-DPs in the high position are incompatible with such interpretations, which is proved by lack of attested data. It is notable that there is another mismatch between these two positions: fronted constituents in the low position are compatible with WEI … ZHI/SHI cleft structures (see (5b), (7a) and (20)), whereas those in the high position are in complementary distribution with clefts, which is also proved by lack of data. Since WEI … ZHI/SHI cleft constructions correlate with exclusive and exhaustive interpretations in LAC, this mismatch also leads to a presumption that fronted DPs in the low position involve exclusive and exhaustive interpretations, while those in the high position do not. Furthermore, assuming cleft constructions correlate with focus (Kiss 1998), then a preliminary conclusion can be drawn that the low position for object preposing is focal, but its higher counterpart is not.

For another, there is no bipartition with fronted non-*wh*-DPs in the high position into the DP and a presupposition, but there is such a bipartition with non-*wh*-DPs in the low position. This asymmetry is illustrated by the fact that the whole VP in sentences involving the higher position can be negated and questioned; besides, the lack of bipartition into the preposed element and the presupposition is further demonstrated by the possibility of raised DPs in the higher position to occur in a list context. First, the former clause of (21) involving the high position shows that the entire VP, including the fronted object and the presupposition, is negated. Based on the following rhetorical question and contextual information, the former clause in (21) can be assumed to imply an "even" interpretation that the person does not even adore his own parents, not to mention others. So that means in example (21), it is not only the fronted DP that is negated; the verb is negated as well.

(21) 其 父母 之 不 親 也, 又 能 親 君 乎?
Qi **fumu** zhi bu [VP qin $t_{qi\ fumu}$] ye, you neng qin jun hu?
3.Gen parents ZHI not adore Decl then can adore lord Q
"(He) does not adore his parents, then how can (he) adore the lord?"
(韓非子•十過第十 3rd c BC)

Second, the lack of bipartition for sentences involving the high position is supported by the fact that the VP as a whole can be questioned. In example (9a), repeated as (22), both the

preposed NP *si* and the presupposed part are questioned.

(22) 祀 不 可以 已 乎? (國語•楚語下 5th c BC)
Si bu keyi [VP yi t_{si}] hu?
propitiation not can cease Q
"Cannot (I) cease the propitiation?"

Third, the fact that preposed objects in the higher position can appear in list contexts indicates the absence of bipartition with fronted non-*wh*-DPs in the high position into the DP and the presupposition. This fact also implies that the high position is not focal, because listing is the opposite of focalisation. In (23), DPs fronted into the high position appear in a list context.

(23) 宴語 之 不 懷, 寵光 之 不 宣,
yanyu zhi bu [VP huai t_{yanyu}], **chongguang** zhi bu [VP xuan $t_{chongguang}$],
chat ZHI not cherish favour.glory ZHI not appreciate
令德 之 不 知, 同福 之 不 受
lingde zhi bu [VP zhi t_{lingde}], **tongfu** zhi bu [VP shou t_{tongfu}]
virtue ZHI not understand common.blessing ZHI not accept
"(they) did not cherish the chat; (they) did not appreciate the glory of favour; (they) did not understand the virtue; (they) did not accept the common blessing"
(左傳•昭公十二年 5th c BC)

By contrast, sentences involving the low position display a bipartition into the fronted object and the presupposition. First, such bipartition excludes the presupposition from the scope of negation, which is a property of an association of focus pattern. In (24a), although the negator precedes both the matrix predicate WEI and the embedded verb "listen", only the matrix verb is negated. Second, example (5b), repeated as (24b), demonstrates another consequence caused by the bipartition: the question only applies to the preposed NP "order", while the presupposition remains unaffected. Third, absence of preposed objects in the lower position appearing in list contexts[16] supports the argument that there is a bipartition into DP objects fronted in the low position and their corresponding presuppositions. It is noteworthy that such restriction is not universal, because both DPs raised into the high position and those remaining in situ can occur in list contexts, as shown in (23) and (24c) respectively. So the fact preposed objects in the low position never appear in a list must be attributed to a reason other than locality restriction. Since focalisation is the opposite of listing, the lack of listed constituents in the low position implies that the low position could be focal.

(24) a. 鄭國 而 不 唯 晉 命 是 聽 (左傳•襄公九年 5th c BC)
zhengguo er bu wei **jin** **ming** shi [VP ting $t_{jin\ ming}$]
Zheng.state Conj not WEI Jin order SHI listen
"regarding the state of Zheng, it is not (only) the order of the state of Jin it listens to"

b. 敢 不 唯 命 是 聽? (左傳•昭公十二年 5th c BC)

[16] This claim is supported by lack of attested data.

Gan bu wei **ming** shi [$_{VP}$ ting t_{ming}]?
dare not WEI order SHI listen
"How dare (I) not listen to orders only? (It is only orders (I) must follow.)"

c. 故 制 之 以 義, 旌 之 以 服,
Gu zhi zhi yi yi, jing zhi yi fu,
so formulate 3.Obj with justice indicate 3.Obj with uniform
行 之 以 禮, 辯 之 以 名,
xing zhi yi li, bian zhi yi ming,
conduct 3.Obj with etiquette distinguish 3.Obj with terminology
書 之 以 文, 道 之 以 言。
shu zhi yi wen, dao zhi yi yan.
write 3.Obj with script narrate 3.Obj with utterance
"So (the emperor) formulated it with justice, indicated it with uniforms, conducted it with etiquette, distinguished it with terminology, wrote it with scripts, (and) narrated it with utterances."

(國語•楚語上 5thc BC)

The aforementioned discrepancies between the high and low positions for the preposing of non-*wh*-objects coincide with those between the internal topics and foci. To be more specific, non-*wh*-objects in the higher position are consistent with a topical interpretation: incompatibility with exclusive or exhaustive interpretation and the lack of bipartition into fronted objects and presuppositions. In terms of constituents in the lower position, they are consistent with a focal interpretation, because preposed non-*wh*-DPs in the low position involve exclusive and exhaustive interpretations and entail a bipartition into fronted elements and presuppositions.

To summarise, by illustrating the relative ordering between fronted non-*wh*-DPs and negatives as well as disparate positions of fronting markers, in Section 2.1 I have validated two landing sites for the preposing of non-*wh*-objects between TP and *v*P. Through comparing constructions involving two preverbal positions, I have further demonstrated in this subsection that the higher position for object preposing displays topic features, while the lower position is focal.

3.2. Nature of Fronting Markers

As shown previously in Section 2.2, the fronting maker ZHI can follow preposed non-*wh*-DPs either in the high or low position, while its counterpart SHI is exclusively permitted in the low position. Since in the previous section I have demonstrated that non-*wh*-DPs in the higher position are consistent with a topical interpretation, whereas those in the lower position are consistent with a focal interpretation, the asymmetry between fronting markers ZHI and SHI can be explained by their respective nature: ZHI can act as either a topic or focus marker, while SHI acts exclusively as a focus marker.

4. Preposing of Wh-Objects

In this section, I explore the preposing of *wh*-DPs in LAC, and propose that D-linked *which*-phrases raise to the internal topic position, while other *wh*-phrases front to an extra focal position between the topic position and negation. In terms of the landing site of preposed *wh*-DPs, it also occupies the specifier of functional projections, parallel to that of non-*wh*-DPs. The underlying structure of *wh*-P is discussed in this section as well.

4.1. Two Positions for Wh-Fronting

Similar to non-*wh*-DPs, *wh*-phrases also front to a position below the subject and above *v*P. However, unlike their non-*wh*-counterparts, fronting of *wh*-objects is obligatory in LAC, unless *wh*-constituents are the second complement in a double object construction. Examples (3a/b), repeated as (25a/b), illustrate a preverbal bare *wh*-word 何 *he* "what" and internally complex *wh*-phrases 何役 *he yi* "what battle" and 何盟 *he meng* "what alliance" respectively. Examples (25c) concerns another simplex *wh*-word 孰 *shu* "what/who". As for (25d), it contains a *wh*-indefinite 誰 *shui* "who".[17]

(25) a. 何 不 爲 乎? (莊子•秋水 4thc BC)
He bu [VP wei t_{he}] hu?
what not do Q
"What do (I) not do?"

b. 宋 何 役 之 不 會,
Song [**he** **yi**] zhi bu [VP hui $t_{he\ yi}$],
Song what battle ZHI not enter
而 何 盟 之 不 同? (左傳•昭公二十五年 5thc BC)
er [**he** **meng**] zhi bu [VP tong $t_{he\ meng}$]?

[17] According to Aldridge (2007), Archaic Chinese wh-elements are typically quantificational, but they can function as indefinites as well. Both instances of wh-elements being quantificational operators and indefinites are attested, although the latter is much rare. For instance, all the wh-elements in aforementioned instances are quantificational, while (24d) is an example involving a wh-indefinite. However, I do not concur with the claim that wh-movement in LAC is the result of prohibition on quantificational materials in VP (Aldridge 2006). First, fronted wh-elements in LAC are not restricted to quantificational operators; consequently, being non-quantificational, wh-indefinites are expected to remain in situ. However, this prediction is counterfactual, as illustrated by (24d) where the wh-indefinite is preposed to a preverbal position. Second, in (i), repeated as (iiia), as the second argument in a double object construction, the quantificational operator he "what" does remain in its base position, ie within VP. Third, the statement that quantified NPs are banned in VP is not borne out, because example (iiib) reveals the possibility of quantified NPs appearing within VP.

(iii) a. 國 謂 君 何? (左傳•僖公十五年 5thc BC)
Guo wei jun he?
state call lord what
"What does the state call the lord?"
b.下佐食 取 牢 一切 肺 于 俎 (儀禮 5thc BC)
xiazuoshi qu lao yiqie fei yu zu
xiazuoshi take sacrifice all lung from vessel
"xiazuoshi (the worshipper) took all the lungs of the sacrifices from the vessel"

Conj what alliance ZHI not join
"What battle the State of Song does not enter, and what alliance (it) does not join?"

c. 孰 不 可 忍 也! (論語•八佾 5th c BC)
Shu bu ke [VP ren t_{shu}] ye!
what not can endure Decl
"What (he) cannot endure!"

d. 誰 之 不 如, 可以 求 之。 (國語•晋語六 5th c BC)
Shui zhi bu [VP ru t_{shui}], keyi qiu zhi.
who ZHI not compare can follow 3.Obj
"If you don't measure up to *someone*, you can follow him."

(Aldridge 2010:45)

As can be seen from these examples, the landing position of *wh*-fronting is always above negation, which supports Aldridge's generalisation that *wh*-words never appear under negation (2010). In addition, *wh*-phrases appear between modals other than those of ability and negation, as shown in (26). Therefore, all *wh*-constituents in LAC seem to appear in the topic position.

(26) 將 何 不 忘 哉! (韓非子•喻老 3rd c BC)
Jiang **he** bu [VP wang t_{he}] zai!
will what not forget Decl
"What will (you) not forget!"

Nevertheless, D-linking in LAC determines that there are two types of *wh*-phrases: 1) *which*-phrases that are D-linked in the sense of Pesetsky (1987), and 2) other *wh*-phrases that are non-D-linked.[18] The former is unselectively bound by Q and is associated with a restricted set of possible answers known in the context, whereas the latter is novel in the discourse, and is not linked to any already existing entry. I hypothesise that *which*-phrases in LAC is topical, landing in the topic position above negation. In (27), 何-NP returns an old familiar entry in the filing system of discourse, so it expresses the meaning "which NP", and it appears in the topic position.

(27) 以 此 攻 城, 何 城 不 克?
Yi ci gong cheng, [**he cheng**] bu [VP ke $t_{he\ cheng}$]?
with this attack city which city not conquer
"(If I) attack cities with this, which city cannot (I) conquer?"

(左傳•僖公四年 5th c BC)

In terms of non-D-linked simplex *wh*-words and *wh*-elements other than *which*-phrases, they display features of foci, therefore, their usual landing site is supposed to be the low (focal) position. However, in LAC, negators trigger further *wh*-movement when c-commanding *wh*-elements, induced by the Intervention Effect (Kim 2002). Kim discusses the blocking effect

[18] The possibility of D-linking in LAC was pointed out by an anonymous reviewer.

on LF movement of *wh*-in-situ (termed by Hagstrom 1998 and Pesetsky 1999 as the Intervention Effect), and proposes that what induces an Intervention Effect in modern Mandarin is focus phrases, instead of negation or quantifiers in general (2002). It is obvious that in LAC, negation does display some kind of Intervention Effect when c-commanding *wh*-phrases that are in the focus position. Unlike its counterpart in modern Mandarin, the focus construction in LAC does not act as the barrier for *wh*-fronting, but as the constituent undergoing further movement.[19] However, *wh*-elements other than *which*-phrases would not be expected to move to a topic-like position, so I presume there must be an extra position exclusively for non-D-linked *wh*-DPs. Furthermore, this landing site for non-D-linked *wh*-DPs is focal, intervening between the topic position and negation. As (28a) shows, this sentence involves a "high" adverbial 何以 *heyi* "why" ("what for"), and it indicates that *wh*-elements other than *which*-phrases occupy a position following the internal topic (聖人 *shengren* "sage") while preceding negation. In (28b), the same 何-NP as that in (27) means "what NP", in that it is not bound by any known entry mentioned in the previous context; so it is focalised, appearing in the extra position exclusively for non-D-linked *wh*-DPs (as is also true for (25b)). Similarly, bare *wh*-words such as 何 "what", 誰 "who" and 孰 "what/who" also move the extra focalised position (see eg (25a), (25c) and (26)).[20]

(28) a. 聖人 何 以 不 可 欺? (荀子•非相 3rdc BC)
Shengren **he** [PP yi t_{he}] bu ke [VP qi $t_{shengren}$]?
sage what for not can deceive
"For what (people) cannot deceive sages?"
b. 先 君 若 问 与夷, 其 將 何 辭 以 對?
Xian jun ruo wen Yuyi, qi jiang [**he** **ci**] [VP[PP yi $t_{he\ ci}$] dui]?

[19] I find this kind of LAC-type Intervention Effect of negation applies to modern Mandarin as well. In wh-fronting sentences, simplex and complex wh-constituents intervene between the subject and negaters; providing wh-phrases land in a position following negatives, ungrammatical sentences would be generated:

(iv) a. Ni shenme bu [VP chi tshenme]?
You what not eat
"What do you not eat?"
b. * Ni bu shenme [VP chi tshenme]?
You not what eat
c. Ni shenme shuiguo bu [VP mai tshenme shuiguo]?
You what fruit not buy
"What fruit do you not buy?"
d. * Ni bu shenme shuiguo [VP mai tshenme shuiguo]?
You not what fruit buy

The other type of Intervention Effect in the sense of Kim (2002) can be illustrated as follows:

(v) a. ?* Zhiyou Lili kan-le na-ben shu?
Only Lili read-Asp which-Cl book
b. Na-ben shu zhiyou Lili kan-le?
Which-Cl book only Lili read-Asp
"Which book did only Lili read?"
(Kim 2002: 12)

[20] Due to the limitation of historical linguistics hence lack of abundant data, I cannot find any example to prove that 何-NP with an interpretation of "what-NP" and simplex *wh*-words indeed occur in a position below the topic position and above negation; I only assume they are lower than *which*-phrases.

former lord if ask Yuyi Mod will what utterance with reply
"If the former lord asks about Yuyi, what utterances will (I) reply with?"
(左傳•隱公三年 5thc BC)

Hence, the final version of the linear format of the clausal positions and the medial elements is as follows:

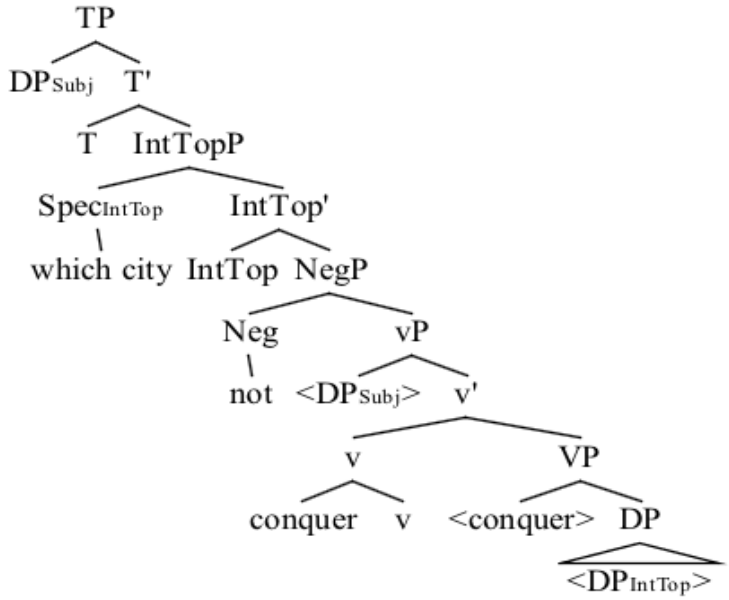
(29) Final version:

Subject > Other modals > Topic position > Extra focus position for *wh*-DPs > Negation > Focus position > Modals of ability > *v*P

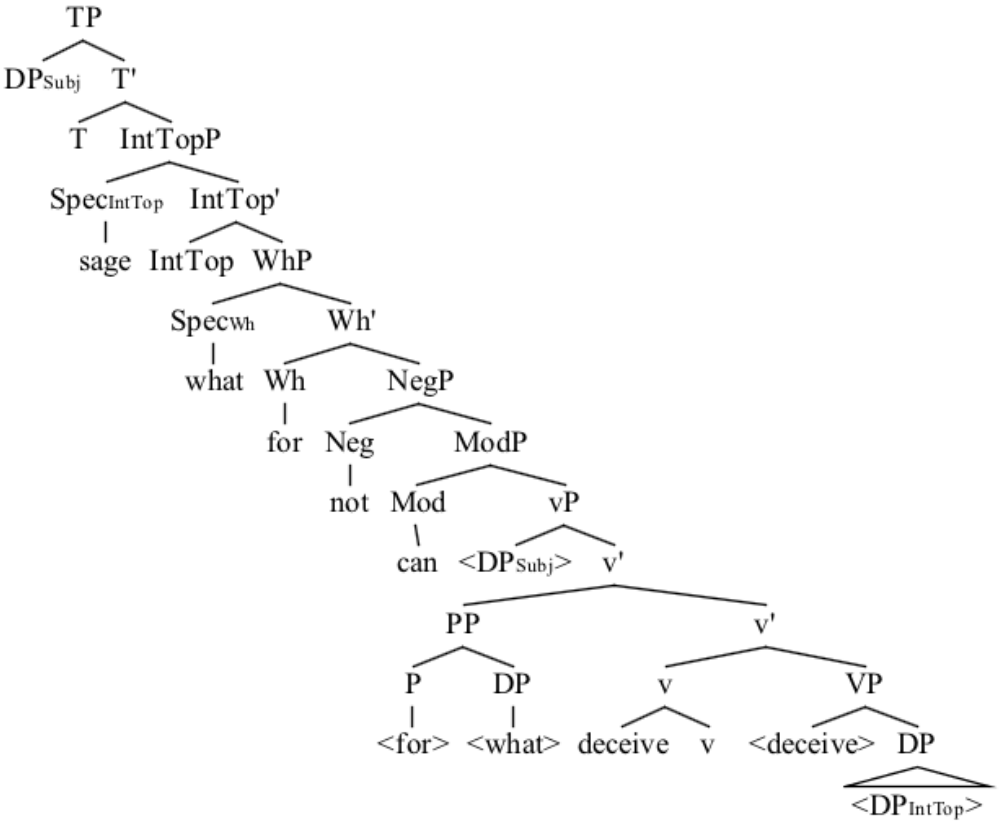
4.2. Landing site of Wh-Fronting

With respect to the landing site of fronted *wh*-DPs, I posit it also occupies the specifier position of relevant functional category between TP and *v*P, similar to that of non-*wh*-DPs. The landing site of *which*-phrases is the $Spec_{IntTop}$, while non-D-linked *wh*-objects occupy the specifier position of the extra focalised position termed "*Wh*P". The tree diagrams of (25) and (28a) are in (30a/b) respectively. It is worth mentioning that both the focus marker ZHI (but not SHI, because SHI never appears above negation) and raised prepositions occupy the head position of *Wh*P, so they are always in a complementary distribution: focus marker ZHI only follows DPs, but never coexists with PPs. However, if ZHI acts as a topic marker, it may occupy the head of IntTopP, thus coexisting with (and preceding) a preposition that is Wh^0. Such possibility is feasible in principle, although it is not borne out due to limited data.

(30) a.



b.



4.3. Wh-P

In (28), the DP movement strands a preposition occupying the head node of the functional projection, which is a robust aspect of LAC syntax. In LAC, the canonical order of prepositional phrases is P-DP (31a),[21] but the reverse order is also prevalent (31b). Moreover, since *wh*-in-situ is forbidden in LAC, if *wh*-elements act as the complement of a preposition, they always appear in the form of *wh*-P, even if internally complex (31c-d).

(31) a. 將 以 戈 擊 之 (左傳•昭公二十五年 5thc BC)
jiang yi ge ji zhi
will with spear attack 3.Obj
"(they) will attack him with a spear"

b. 弓 以 招 士 (左傳•昭公二十年 5thc BC)
gong [VP [PP yi t_{gong}] zhao shi]
bow with summon gentleman
"(he) summoned a gentleman with a bow"

[21] The claim of P-DP being the unmarked order is supported by the overwhelming proportion of P-DP structures relative to the DP-P pattern.

c. 將 何 以 守 國? (國語•周語上 5thc BC)
Jiang **he** [VP[PP yi t_{he}] shou guo]?
will what with guard state
"Will what with to guard the state?"

d. 先 君 若 问 与夷, 其 將 何 辭 以 對?
Xian jun ruo wen Yuyi, qi jiang [**he** **ci**] [VP[PP yi $t_{he\ ci}$] dui]?
former lord if ask Yuyi Mod will what utterance with reply
"If the former lord asks about Yuyi, what utterances will (I) reply with?"
(左傳•隱公三年 5thc BC)

Before discussing the *wh*-P structure, it is necessary to point out that the morpheme 以 *yi* "with/for" in LAC should be treated as a preposition, following the traditional analysis (see among others, Djamouri et al. 2012). The theory of 以 *yi* in LAC being a high applicative above VP within *v*P is not well supported (Aldridge 2012b). Aldridge presumes that the functional morpheme 以 *yi* heads a high applicative phrase, and she takes the *yi*-DP-VP order to be basic, based on its frequent occurrence in a 5thc BC text. The Appl head *yi* moves to *v*, generating *yi*-DP-VP, whereas the structure of VP-*yi*-DP is derived from VP fronting. She points out that extraction is impossible when *yi*-DP is postverbal: the *DP-VP-*yi* pattern with *yi* stranded in postverbal position is unattested. She posits the nature of *yi*-DP being ApplP rather than PP, and builds the theory on a hypothesis that both *wh*- and VP fronting target the outer specifier of *v*P, so the focus movement is blocked (2012b). There is no denying the fact that the *DP-VP-*yi* ordering is impermissible in LAC. However, this theory predicts the wrong order between postverbal *wh*-word and *yi*: the presumed VP-*yi*-*wh* pattern is not attested for independent reasons, namely, *wh*-phrases other than those acting as the second complement of double object constructions must front. Moreover, I adopt the theory that preposed *wh*-elements occupy a specifier position of some functional projection above *v*P, not the edge of *v*P, so the assumed blocking effect on partial *wh*-movement caused by VP fronting should not occur. The blocking effect predicts that there should not be instances of *yi* being stranded in a postverbal positionon. Nevertheless, the availability of VP-*wh*-*yi* instances in LAC (as shown in (32)) demonstrates that the blocking effect on *wh*-movement triggered by VP fronting does not exist. This example also illustrates that VP fronting targets a node above *v*P, instead of the specifier of *v*P, because VP is higher than the functional projection for preposed *wh*-elements that is above *v*P.

(32) 救 饑 何 以?[22] (國語•晋語四 5thc BC)
[VP Jiu ji **he** [PP yi t_{he}]]?
solve famine what with
"What to solve the famine with?"

[22] 救饑 *jiu ji* "solve famine" in (30) is treated as a VP, instead of a sentential subject, otherwise there would be no predicate in this sentence.

Another defect of the applicative approach lies in the analysis concerning negation in the *yi* construction. Aldridge generalises that both *yi*-DP and the following VP are in the scope of negation if *yi*-DP precedes the VP (Aldridge 2012b). However, this argument fails to take into consideration the asymmetry between sentential negation and constituent negation, as shown in (33a) and (33b) respectively. In (33a), it is obvious that both the *yi*-DP construction as well as the VP "harm people" are negated, because gentlemen do not harm people in any way. By contrast, negation in the former clause of (33b) only applies to *yi*-DP whereas the VP "obtain them" is not negated, because the verb "retain" in the posterior context implies that the action of obtaining wealth and status must have been presumed to be done, otherwise the action of retaining them would not have been mentioned.

(33) a. 君子 不 以 其 所 以 養 人 者 害 人。
Junzi bu yi [qi suo yi yang ren zhe] hai ren.
gentleman not with 3.Gen SUO with nurture person Det harm person
"Gentlemen do not harm people using that with which they nurture them."
(孟子•梁惠王上 4th c BC)

b. 不 以 其 道 得 之, 不 處 也。
Bu yi [qi dao] de zhi, bu chu ye.
not with 3.Gen means obtain 3.Obj not retain NMLZ
"(If people) do not obtain them by their means, (people) do not retain (them)."
(論語•裏仁 5th c BC)

In addition, the applicative theory predicts that if *yi*-DP were analysed as a PP within VP, negation in (34) would precede the verb and take scope over *yi*-DP; but the *Neg-V … *yi*-DP pattern is unattested (Aldridge 2012b).

(34) 域 民 不 以 封 疆 之 界,
[VP Yu min] bu yi feng jiang zhi jie,
enclose people not with close strong Gen border
固 國 不 以 山 溪 之 險,
[VP gu guo] bu yi shan xi zhi xian,
secure country not with mountain stream Gen steep
威 天下 不 以 兵革 之 利。 (孟子•公孫醜下 4th c BC)
[VP wei tianxia] bu yi bingge zhi li.
impress world not with military Gen advantage
"(A ruler) keeps his population not with tight borders, secures his country not with steep mountains and gorges, and impresses the rest of the world not with military might."
(Aldridge 2012b: 10)

Nevertheless, I argue that according to contextual information, (34) actually involves constituent negation, in that the negation only applies to the specific methods to realise those goals, excluding the goals themselves. For sentences involving constituent negation, they either take the form of VP-Neg-*yi*-DP, as in (34), or Neg-*yi*-DP-VP, as in (33b). That is to say, *Neg-V … *yi*-DP would not be a feasible pattern of constituent negation anyway: it represents

either sentential negation, or negating only the VP but not *yi*-DP, which is unreasonable. Therefore, the lack of *Neg-V … *yi*-DP pattern fails to invalidate *yi*-DP as a PP.

Therefore, owing to the imperfections of the high applicative analysis of 以 *yi*, I stick to the traditional view and treat *yi* as a preposition. Moreover, there is fronting from PPs concerning other prepositions in LAC that lends indirect support to the traditional approach, and relevant prepositional complements can be *wh*-DPs (35a-b) and non-*wh*-DPs (35c).

(35) a. 吾 誰 與 歸? (國語•晋語八 5th c BC)
Wu **shui** [PP yu t_{shui}] gui?
I who with classify
"With whom I am classified?"

b. 寡人 惡 乎 屬 國 而 可? (莊子•徐無鬼 4th c BC)
[CP[TP Guaren **wu**$_i$ hu t_i shu guo]] er ke?
I whom to entrust nation Conj good
"If I entrusted the nation to whom would it be good?"
(Aldridge 2010: 35)

c. 八 世 之 後, 莫 之 與 京。
Ba shi zhi hou, mo **zhi** [PP yu t_{zhi}] jing.
8 generation Gen after none 3.Obj than great
"After eight generations, there will be no one greater than him."
(左傳•莊公二十二年 5th c BC)

Returning to the underlying structure of *wh*-P, I propose that such inverse ordering is caused by individual raisings of the *wh*-element and the preposition: the DP moves to the specifier of the functional projection *Wh*P that is exclusively for *wh*-phrases, while the preposition moves to the head of *Wh*P. So these positionings generate a disparate relative order before the *wh*-constituent and preposition move. Taking (30b) as an example, the *wh*-word *he* "what" fronts to the specifier of *Wh*P, while the preposition *yi* moves to Wh^0. Consequently, the *wh*-element ends up in a position higher than the preposition.

There are two other potential explanations for the inverse order of *wh*-P: inversion with PP and mere *wh*-fronting (with the preposition stranded in its base position), both of which fail to account for the *wh*-P structure in LAC. In order to invalidate the approach of inversion within PP, P-NP is referred to here. As presented earlier, P-NP can be in the preverbal position, as shown in (31a), repeated as (36a). Furthermore, the canonical head-initial order can be inverted to NP-P, as in (31b), repeated as (36b). Nevertheless, the canonical P-NP order may appear in a postverbal position as well (36c), whereas the inverted NP-P order is restricted to the preverbal environment, namely, only when the PP itself precedes V (36b).

(36) a. 將 以 戈 擊 之 (左傳•昭公二十五年 5th c BC)
jiang yi ge ji zhi
will with spear attack 3.Obj
"(they) will attack him with a spear"

b. 弓 以 招 士 (左傳•昭公二十年 5th c BC)
gong [VP [PP yi t_{gong}] zhao shi]
bow with summon gentleman
"(he) summoned a gentleman with a bow"

c. 富 國 以 農 (韓非子•五蠹 3rd c BC)
fu guo yi nong
enrich state with agriculture
"enriching the state depending on agriculture"

Such distribution asymmetry between canonical head-initial form P-NP and marked NP-P structure indicates that the DP-P structure in LAC is not induced by an inversion within PP, otherwise D-P is expected to occur postverbally as well.

With respect to the other potential explanation for *wh*-P, it only entails *wh*-fronting, with the preposition stranded in its base position. This assumption implies that there can be constituents intervening between the *wh*-word and the preposition. First, when *wh*-P appears in its unmarked preverbal position,[23] negative/modal elements are expected to be allowed to intervene between the preposed *wh*-constituent and the stranded preposition. Nevertheless, there is a lack of such data validating intervening negative/modal elements, and example (30b) shows that the focalised *wh*-DP immediately precedes the preposition. Second, when *wh*-P appears in its marked postverbal position, it is predicted that the VP can intervene between the *wh*-constituent and preposition, generating *wh*-VP-P. When PP appears postverbally, the surface structure of VP-*wh*-P is generated via VP-fronting, so VP is also expected to move to a position lower than the *wh*-element while higher than the preposition. However, the *DP-VP-P pattern is unattested, and the *wh*-element and preposition are always attached directly to each other (as in example (32)). Therefore, given the wrong predictions made by this assumption in both contexts of preverbal and postverbal *wh*-P, the presumption of *wh*-fronting while P-stranding is ruled out.

Since either the approach of inversion with PP or mere *wh*-fronting can expound the inverse order of *wh*-P, I propose separate movements that the *wh*-element raises to the specifier of *Wh*P, while the preposition fronts to the head of *Wh*P. This argument accounts for three facts: 1) preposed *wh*-element is higher than the preposition in the tree; 2) there is no intervening constituent between the preposed *wh*-phrase and the preposition; and 3) there is a complementary distribution of fronting marker ZHI and prepositions. To be more specific, the reason why fronting marker ZHI only follows DP but never coexists with PP is because when the head node of *Wh*P is occupied by a fronted preposition, there is no position for any focus marker.

[23] The assertion of PP-VP being the basic order can be proved by its overwhelming proportion relative to the postverbal PP order.

5. Conclusion and Remaining Issues

This paper has explored the preverbal positioning of objects in the medial domain below TP and above *v*P in LAC. Based on the relative ordering of preposed non-*wh*-DPs and negation, I propose a high position and a low position. Fronted non-*wh*-objects in the higher position are consistent with a topical interpretation, whereas constituents in the lower position are consistent with a focal interpretation. Nominal and pronominal objects in LAC appear in both positions, and they occupy a specifier node of functional categories, followed by an optional topic/focus marker as the head of relevant functional projections. I also demonstrate that sentences involving modals of ability are not passive constructions. D-linked *which*-phrases in LAC are topical, therefore they occur in the high position. With respect to non-D-linked *wh*-DPs, although they are supposed to appear in the focus position below negatives, the Intervention Effect of negation triggers further *wh*-movement to an extra (focus) position between the topic position and negation. I also discuss the underlying structure of *wh*-P and illustrate that such inverse ordering is generated via separate raisings of *wh*-constituents and prepositions.

Of course, there are still remaining issues for future research: the motivation of *wh*-fronting/in-situ, the nature and motivation of pronoun fronting in the context of negation, the fact that pronouns in an identical environment sometimes undergo fronting, but sometimes do not, etc. These questions must be investigated in future research.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Peter Sells, for his inspiring guidance and instruction on this paper. I also would like to sincerely thank Professor Edith Aldridge for her invaluable and helpful comments and feedback on the original draft. I am also very grateful for the constructive suggestions from two anonymous reviewers.

References

Primary sources

CCL Corpus [Electronic Corpus of Chinese Texts]
http://ccl.pku.edu.cn:8080/ccl_corpus/
Center for Chinese Linguistics, Peking University, Beijing, China
Hanji Dianzi Wenxian [Electronic Corpus of Chinese Texts]
http://hanji.sinica.edu.tw/index.html
Institute of Linguistics, Academic Sinica, Taipei, Taiwan
The Sheffield Corpus of Chinese [Electronic Corpus of Chinese Texts]
http://www.hrionline.ac.uk/scc/db/scc/index.jsp
The University of Sheffield, Sheffield, UK

ALDRIDGE, E. 2006. "VP-internal Quantification in Old Chinese". In R. Djamouri & R. Sybesma (eds), *Chinese Linguistics in Budapest*, 1–15. Paris: Ecole des Hautes Etudes en Sciences Sociales, Centre de Recherches Linguistiques sur l'Asie Orientale.

ALDRIDGE, E. 2007. "*Wh*-indefinites and their Relation to *Wh*-in-Situ". In M. Elliott, J. Kirby, O. Sawada, E. Staraki, S. & Yoon (eds.), Proceedings of the 43rd Meeting of the Chicago Linguistics Society, vol. 2: The Panels, 139–153.

ALDRIDGE, E. 2010. "Clause-internal *Wh*-movement in Archaic Chinese". *Journal of East Asian Linguistics* 19.1, 1–36.

ALDRIDGE, E. 2011. "Survey of Chinese historical syntax". Ms., University of Washington.

ALDRIDGE, E. 2012a. "Focus and Archaic Chinese Word Order". In L. E. Clemens and C-M. L. Liu (eds.), *The Proceedings of the 22nd North American Conference of Chinese Linguistics* (NACCLS-22) and the 18th Annual Meeting of the International Association of Chinese Linguistics (IACL-18), vol. 2, 84–101.

ALDRIDGE, E. 2012b. "PPs and Applicatives in Late Archaic Chinese". *Studies in Chinese Linguistics* 33.3, 139–164.

ALDRIDGE, E. To appear. "Pronominal Object Shift in Archaic Chinese". In T. Biberauer and G. Walkden (eds.), *Syntax over Time: Lexical, morphological and information-structural interactions*. Oxford University Press.

BELLETTI, A. 2003. "Aspects of the low IP area". In L. Rizzi (ed.), *The structure of CP and IP: The cartography of syntactic structures*, vol. 2, 16–51. Oxford University Press.

CARDINALETTI, A. AND M. STARKE. 1996. "Deficient Pronouns: A View from Germanic. A Study in the Unified Description of Germanic and Romance". In *Studies in Comparative. Germanic Syntax*, Volume II. 21–65. Dordrecht: Kluwer.

CARDINALETTI, A. AND M. STARKE. 1999. "The typology of structural deficiency: A case study of the three classes of pronouns". In Riemsdijk H. van (ed.), *Clitics in the languages of Europe*, 145–233. New York: Mouton de Gruyter.

CHOU, F. 1963. "Stages in the development of the Chinese Language". In F. Chou (ed.), *Zhongguo yuwen luncong*, 432–438. Taipei: Zhengzhong Shuju.

DJAMOURI, R. 1991. "Pariticules de négation dans les inscriptions sur bronze de la dynastie des Zhou [Particles of negation in the bronze inscriptions of the Zhou dynasty]". *Cahiers de Linguistique—Asie Orientale* 20.1: 5–76.

DJAMOURI, R. 2000. "Preverbal Position of the Pronominal Object in Archaic Chinese". Paper presented at the 9th International Conference on Chinese Linguistics, The National University of Singapore.

DJAMOURI, R. 2001. "Markers of predication in Shang bone inscriptions". In H. Chappell (ed.), *Synchronic and Diachronic Perspectives of the Grammar of Sinitic Languages*, 143-171. Oxford : Oxford University Press.

DJAMOURI, R. AND W. PAUL. 2009. "Verb-to-preposition reanalysis in Chinese". In P. Crisma & G. Longobardi (eds.), *Historical syntax and linguistic theory*, 194–211. Oxford: Oxford University Press.

DJAMOURI ET AL. 2012. "Syntactic change in Chinese and the argument-adjunct asymmetry". In G. Cao, H. Chappell, R. Djamouri & T. Wiebusch (eds.), *Breaking down the barriers: Interdisciplinary studies in Chinese linguistics and beyond*. Taipei: Academia Sinica.

FENG, S. 1996. “Prosodically constrained syntactic changes in Early Archaic Chinese”. *Journal of East Asian Linguistics* 5, 323–371.

HORNSTEIN, N. 1999. Movement and Control. *Linguistic Inquiry* 30, 69–96.

HORNSTEIN, N. 2001. *Move! A Minimalist Theory of Construal*. Oxford: Blackwell.

KIM, S-S. 2002. “Intervention Effects Are Focus Effects”. In N. Akatsuka & S. Strauss (eds.), *Japanese/Korean Linguistics* 10, 615–628. Stanford: CSLI.

KISS, K. E. 1998. “Identificational focus versus information focus”. *Language* 74 (2), 245–273.

LA POLLA, R. 1993. “On the Change to Verb-Medial Word Order in Proto-Chinese: Evidence from Tibeto-Burman”. In H. Kitamura, T. Nishida & Y. Nagano (eds.), *Current Issues in Sino-Tibetan Linguistics*, 98–104. Osaka: National Museum of Ethnology.

LI, C. N. AND S. A. THOMPSON. 1974. “Co-verbs in Mandarin Chinese: Verbs or Prepositions?”. *Journal of Chinese Linguistics* 2.3, 257–278.

MEISTERERNST, B. 2008a. “Modal verbs in Han period Chinese Part I: The syntax and semantics of kě 可 and kě yǐ 可以”. *Cahiers de Linguistique—Asie Orientale* 37(1), 85–120.

MEISTERERNST, B. 2008b. “Modal verbs in Han period Chinese Part II: Negative Markers in combination with the modal auxiliary verbs *kě* 可 and *kěyǐ* 可以”. *Cahiers de Linguistique—Asie Orientale*, 37(2), 197–222.

MEISTERERNST, B. 2010. “Object Preposing in Classical and pre-Medieval Chinese”. *Journal of East Asian Linguistics*, 19.1, 75–102. 19.1 and Online publication doi: 10.1007/s10831-010-9056-x.

PAUL, W. 2002. Sentence-internal topics in Mandarin Chinese: the case of object preposing. *Language and Linguistics* [Academia Sinica, Taiwan] 3, 4: 695–714.

PAUL, W. 2005. Low IP area and left periphery in Mandarin Chinese. *Recherches linguistiques de Vincennes* 33, 111–134.

PESETSKY, D. 1987. “*Wh*-in-situ: movement and unselective binding”. In E. Reuland and A. Meulen (eds.), *The representation of (in)definiteness*. Cambridge, MA: MIT Press, 98–129.

PEYRAUBE, A. 1996. “Recent Issues in Chinese Historical Syntax”. In C.-T. J. Huang and Y.-H. A. Li (eds.), *New Horizons in Chinese Linguistics*. Dordrecht : Kluwer, 161–214.

PEYRAUBE, A. 1999. “The modal auxiliaries of possibility in classical Chinese”. In F. Tsao, S. Wang, and C. Lien (eds.), *Selected papers from the Fifth International Conference on Chinese Linguistics*. Taipei: The Crane Publishing Co., 27-52.

PEYRAUBE, A. 2001. “On the Modal Auxiliaries of Volition in Classical Chinese”. In H. Chappell (ed.), *Sinitic Grammar: Synchronic and Diachronic Perspectives*. Oxford: Oxford University Press, 172-188.

PULLEYBLANK, E. G. 1995. *Outline of Classical Chinese Grammar*. University of British Columbia: UBC Press.

WANG, L. 1958. *Hanyu shigao*. Reprinted in 2004. Beijing: Zhonghua Shuju.

XU, D. 2006. *Typological Change in Chinese Syntax*. Oxford: Oxford University Press.

Aiqing Wang
Department of Language and Linguistic Science
University of York
Heslington
York
YO10 5DD
United Kingdom
email: aiqing.wang@york.ac.uk

INFERENCE OF THREAT FROM NEUTRALLY-WORDED UTTERANCES IN FAMILIAR AND UNFAMILIAR LANGUAGES

DOMINIC WATT, SARAH KELLY & CARMEN LLAMAS
University of York

Abstract

Although verbal threats are a very common kind of language crime, the ways in which listeners interpret ostensibly 'neutral' utterances as threats are currently poorly understood. We present the results of an experiment in which monolingual English-speaking listeners were exposed to the same innocuously-worded phrase spoken in a 'neutral' or a 'threatening' way. They heard translations of the phrase in four unknown foreign languages as well as the original form in English. The listeners were asked to rate the utterances with respect to two perceived properties: (a) how threatening they thought the utterances sounded, and (b) how much 'intent' to carry out a harmful act the listeners inferred from the talker's speech. As predicted, the listeners assigned higher threat and intent ratings to the 'threatening' utterances in both English and the foreign languages than they did to the neutral ones. However, the listeners' ratings were considerably higher for both threatening and neutral utterances spoken in English than they were for the foreign language utterances. In the English condition there was also a much larger difference between the neutral and the threat utterances with respect to the overall perceived threat and intent ratings than there was for the foreign language utterances. This suggests that correctly interpreting the threatening utterances as threats is dependent upon familiarity with the language in which they were spoken. A gender effect was also found, whereby male listeners assigned higher threat and intent ratings than did women. It is suggested that men and women may respond differently to speech cues associated with threatening behaviour.

1. Introduction

Among the set of offences that are committed using only words – a set which, *inter alia*, includes blackmail, extortion, bribery, blasphemy, profanity, fraud, impersonation, perjury, defamation, and incitement to racial hatred (Shuy 2005; Tiersma & Solan 2012) – is making a verbal threat. A threat may loosely be defined as a form of expression that communicates that some undesirable state of affairs (physical injury, for instance) may or will befall the recipient or a third party as a consequence of another's purposeful actions. Existing research on this topic has tended to focus on the inference of threat from spoken utterances or written texts on the basis of their content: that is, the meanings of the words themselves, alone and in combination, and their illocutionary force: that is, what the speaker intended by uttering them (see Austin 1962; Fraser 1976, 1998; Gales 2012; Searle 1969; Shuy 1993).

Gales (2012) gives examples of three main kinds of threats: direct threats, conditional threats, and indirect threats. Direct threats assert that something bad will come about, and describe how, when and to whom it will happen, while conditional threats specify what will happen unless some condition is met. The third kind, the indirect threat, is the type of threat that is most difficult to identify as such, as it may masquerade as some other kind of speech act. It may simultaneously be interpretable as a warning, as helpful advice, or as a complaint, so its classification as a threat depends heavily upon how the listener reacts to it (Gingiss 1986; Napier & Mardigian 2003; Smith 2008; Storey 1995; Yamanaka 1995). Part of what makes

this sort of oblique utterance a threat is the supposition on the hearer's part that the speaker is potentially capable of performing some action that would materially disadvantage the hearer, or someone else. Its interpretation as a threat draws upon knowledge of relevant circumstances which is not encoded in the utterance. For instance, a comment such as 'you should be careful, you don't want to end up like person X', which might under some circumstances be interpreted as an altruistic warning, could reasonably be considered a threat if the addressee was aware that person X had recently been the victim of a violent beating. Whether the above comment is construable as threatening, and thereby has the desired effect of intimidating the recipient, is therefore contingent upon both speaker and recipient having a shared understanding of relevant background information (see further Austin 1962).

It can nonetheless be very difficult to prove that an indirect threat was *meant* as a threat, even if it was interpreted as such. The person accused of making the threat may very easily deny that he or she had any actual intention of harming someone else, and may say that no such implication was intended either. Even if the precise wording of the utterance is not in dispute, it could be claimed that it was meant in an advisory way, or simply as a joke. The wording of the utterance may in some cases look so neutral or harmless that the debate reduces to the level of irresolvable gainsaying. An illustration of this scenario is a British court case from 2012 that revolved around whether the phrase "When I get out of here [a police cell] I'm going to do something about this" constituted a reiteration of an earlier (but unrecorded) verbal threat to shoot a legal official. The defendant vociferously denied that the utterance, which was recorded by a CCTV camera in the cell and which was witnessed by the custody officer who gave evidence against the defendant, was a threat of any kind. Indeed, it is quite easy to imagine that the remark could perhaps have been a statement of the speaker's intent to lodge a formal complaint, or to write to his Member of Parliament. The custody officer was, however, adamant that because he (the officer) interpreted the remark as a reiteration of a threat it *was* therefore a threat, an assertion he defended by stating his impression that the defendant's behaviour at the time was agitated and belligerent, and that the tone of voice he used sounded aggressive.

It is rather obvious that what we might call the speaker's 'tone of voice' may play a key role in how ostensibly neutral utterances can be interpretable as threats to harm the addressee or a third person. Yet to the best of our knowledge no previous research whatever has been done on what we might call 'the phonetics of threat': the ways in which speakers convey menace, ill-will, intention to harm, and so forth, using speech cues that listeners interpret in the expected way. This gap in the research literature is surprising, given how frequently samples of allegedly threatening speech that are obtained via interception of telephone calls or from covertly-recorded conversations are encountered in forensic speech analysis casework. There is a considerable literature on the phonetic properties of other sorts of emotional speech, and that on 'angry' speech (Gobl & Ní Chasaide 2003; Polzehl *et al*. 2011; Xu & Kelly 2010) is probably most relevant here. However, we must avoid conflating angry speech with threatening speech. Wishing to threaten someone does not presuppose that the threatener is angry with the recipient. We should also distinguish between different sorts of anger (say, what are popularly known as 'hot' and 'cold' anger). However, it seems reasonable to expect some overlap in the clusters of phonetic cues that speakers might exhibit when angry and when communicating a threat, and we must allow in any event for the fact that listeners who believe that a threat is being made by a speaker might assume that the speaker is also angry; this heightened emotional state might be thought to be the reason for the threat to be made in the first place. All the same, in the perception experiment we describe in subsequent sections we aimed to keep the two areas distinct from one another, because while being angry with

another person is not an offence, threatening to harm him or her certainly is. In the next section we consider verbal threats from a legal perspective.

2. Legal aspects

So as to contextualise the experiment described below, it is important to consider further the legal implications of the use of verbal behaviour classifiable as 'threatening speech'. It should be stressed from the outset that for an offence to have been committed it is not necessary that the speaker *have* the capacity to carry out the actions that he or she implies or explicitly identifies, nor must it be demonstrated that the speaker has any *actual* intention of carrying them out. People very frequently make empty threats towards one another, after all (Shuy 1993). It is the act of producing a statement which leads an individual to believe that the words used encode a plausible intention to cause harm that is considered in and of itself to be damaging to the victim.

In England and Wales, the Public Order Act (1986) provides definitions and guidance relating to the status of the verbal threat as a criminal offence. In particular, Section 4.1 of Chapter 64 of the Act ('Fear or provocation of violence') specifies the following:

> (1) A person is guilty of an offence if he—
>
> (a) uses towards another person threatening, abusive or insulting words or behaviour,
>
> or
>
> (b) distributes or displays to another person any writing, sign or other visible representation which is threatening, abusive or insulting,
>
> with intent to cause that person to believe that immediate unlawful violence will be used against him or another by any person, or to provoke the immediate use of unlawful violence by that person or another, or whereby that person is likely to believe that such violence will be used or it is likely that such violence will be provoked. (Public Order Act 1986, Ch. 64, Section 4.1).

The subsequent section relating to 'Intentional harassment, alarm or distress' (4A) states that an offence has been committed when, 'with intent to cause a person harassment, alarm or distress', the offender 'uses threatening, abusive or insulting words or behaviour… thereby causing that or another person harassment, alarm or distress' (Public Order Act 1986, Ch. 64, Section 4A).

Demonstrating that there was intent on the speaker's part to cause psychological or physical harm to another person when the threat was made is crucial if the utterance is to qualify as a criminal offence under Section 4 or 4A ('A person is guilty of an offence under section 4 only if he intends his words or behaviour, or the writing, sign or other visible representation, to be threatening, abusive or insulting, or is aware that it may be threatening, abusive or insulting' (Public Order Act 1986, Ch. 64, Section 6.3). However, Section 5 of Chapter 64 of the Act also allows for the fact that the accused may not necessarily have directed the offending language at the person who believes the speaker to intend him or her harassment, alarm or distress: the linguistic material in question simply has to be produced within the hearing or sight of the recipient.

We cannot engage here very fully with the ways in which the wording of these sections of the Act might be interpreted generally, let alone in any specific case. These are tasks best handled by specialists in legal language, and linguists who concern themselves with the

pragmatic functions of different sorts of speech acts, including forensic linguists (see further Austin 1962; Shavell 1993; Rothchild 1998; Searle 1969, 1975, 1979; Smith 2008; Storey 1995; Salgueiro 2010; Tiersma & Solan 2012). However, we deal briefly below with some of the principal themes of earlier research on threatening language, which has largely dwelt upon discussion of what lends a linguistic construction the status of a threat and whether it is to be treated as a 'legal' threat versus an illegal one. Not all utterances that are classifiable as threats violate any law (e.g. 'if you don't pass your exam, then I'll remove you from this course of study'), but they still count as threats insofar as they are statements formulated in such a way as to imply that the person uttering them has at least the will, and perhaps also the capacity, to bring about a state of the world that is unfavourable to the recipient (Fraser 1976). Fraser also contends that when a statement is considered by the recipient to be unfavourable to him or her (e.g. to that person's health or safety) it thereby acquires the status of a 'successful' threat, irrespective of whether the person who produced the communication is the same individual who intends to carry out the unfavourable act.

In spite of the lucidity of the definitions of threats and threatening verbalisations provided by Fraser and other researchers working in this area, there are still extensive grey areas in relation to what kinds of words or speech acts might be construed as threatening, abusive, insulting, harassing, alarming or distressing. In the real world, it seems safe to say that direct, explicit threats produced 'performatively' (e.g. 'I'm threatening you with violence') will be hugely outnumbered by those worded less directly. Conversely, utterances which look like threats may not be intended as such. An ostensibly threateningly-worded utterance such as 'I'm going to kill you' might on the one hand be used as a playful rebuke between friends during a board game, but under other circumstances (a whispered telephone call received in the dead of night, for example) it could cause the person hearing it feelings of intense fear, dread, and distress. Though the same words in the same order are being used, the circumstances in which they are uttered are pivotal; as mentioned earlier, the offence is committed when justifiable alarm, distress or fear for personal safety is experienced on the part of the addressee. Decisions about whether an utterance constitutes a threat or not must therefore rely to a considerable degree on the ability of the recipient of the threat, a police officer, or a legal professional to judge whether any intent to harm was insinuated by the speaker (Gales 2012; Smith 2008). Whether the speaker is in a position to carry out the action(s) implied by the threat – i.e., whether Austin's 'ability condition' is satisfied (Austin 1962) – is also relevant.

It can also be difficult to distinguish objectively between what counts as a threat versus an insult or verbal abuse, but under some circumstances the distinction is unimportant: the target of insulting language may feel very threatened by it, even if the abuser has not used words or phrases that would usually be interpreted to denote desire or intent to cause the recipient harm. As before, insulting or abusive language serves as a threat if it can be demonstrated that there was intent on the speaker's part to cause alarm or emotional distress to another person. Abusive or insulting terms accompanying a more explicit threat may also affect how severely or seriously the threat is perceived.

Individuals making verbal threats towards others often wish to hide their identities by delivering their messages over the telephone. They may know, however, that there is a possibility that their speech is being recorded, and they may also be aware of the penalties associated with being found guilty of threatening behaviour. It is therefore in the interests of the threatening party to phrase the threat as indirectly as possible, by trying to strike a balance between framing the threatening message in such a way that on one level it looks innocuous, while on another level the recipient will interpret the threat as the speaker intended it. This leaves the threatener, if challenged, with the escape route of claiming that the recipient

simply misinterpreted the words that were used, as per the scenario in the case described in Section 1. A veiled threat may in some circumstances therefore be the most useful type from the threatener's point of view: written down, the message might look neutral or harmless, but if delivered in spoken form in the right way, it might have precisely the intended impact on the listener.

In the following sections, we consider what this 'right way' might be. We do not attempt in the present paper to say what kinds of phonetic strategies a speaker might use when trying to make an innocuously-worded phrase sound threatening to a listener, nor do we seek to identify specific cues in the talker's speech that listeners might respond to such that they infer a greater or lesser level of threat or intent to carry out some harmful action. Rather, we draw our initial focus on whether listeners act consistently in how they assign threat and intent ratings to a set of controlled utterances, in line with whether the utterances were produced by talkers in a neutral way versus a way that was intended to sound threatening.

3. Background to the experiment

To help us to gain some more clarity with respect to the above issue, we use an experimental approach to examine listeners' reactions to sentences read aloud by talkers who were instructed to adopt a threatening tone of voice while reading. We label these productions 'induced-threat' sentences. The sentences were worded so as to seem innocuous if taken at face value, such that if listeners were to infer threat from them when hearing recordings of them being spoken aloud, their inferences would, in principle, be drawn on the basis of the way the sentences were spoken, instead of the semantic properties of the words and phrases of which the sentences were composed.

We are not seeking here to investigate the acoustic properties of these phonetic cues themselves, though we recognise that pursuing this objective would be a useful contribution to our knowledge in this area. Rather, our experiment is designed to gather some initial observations about listeners' judgments of speech samples produced by talkers who had been asked to adopt a 'threatening tone of voice', in respect of how listeners chose to rate the samples for 'level of threat' and perceptions of what we might informally call the speaker's 'sincerity', i.e. the extent to which listeners believe that the speaker might actually carry out some (unspecified) harmful action.[1] Since the classification of a verbal threat as an offence in UK and US legislation depends so heavily upon the intimidatory effect that the perpetrator's utterance has on the listener, we thought it appropriate to focus principally upon how listeners react to verbal stimuli designed as threats. That is, as experimenters we did not wish to assume that our speakers' utterances would be heard as threatening by our impartial listeners simply because we had instructed the speakers to adopt a threatening tone when reading them out loud; the key thing was to establish how the listeners reacted to the samples. Moreover, as a way of drawing a sharper focus on how the sentences were delivered, we circumvented any interference from semantic content by presenting listeners with stimuli spoken in four foreign languages which were unknown to them as well as stimuli spoken in English. By including utterances in multiple languages our results might lend support to claims made by, for

[1] It should be stressed that in this research we are not seeking to attempt to identify phonetic cues to sincerity, or anything of that kind; we are conscious that unless we are careful to make this point explicit we might run the risk of being seen to contravene item 9 of the 2004 Code of Practice of the International Association for Forensic Phonetics and Acoustics (www.iafpa.net/code), viz., 'Members should not attempt to do psychological profiles or assessments of the sincerity of speakers'. Our goal here is simply to report our listeners' subjective judgments in a way that is relevant to the questions at hand.

example, Bolinger (1989), Gussenhoven (2004) or Xu *et al*. (2013) about the universality of associations between prosodic cues and social/emotional affect.

We also set out to obtain responses from both men and women so as to allow us to test whether there is a consistent effect for listener sex, in line with results reported in earlier literature which indicate that men and women process and respond to emotional speech differently (Bonebright *et al*. 1996; Schirmer *et al*. 2004, 2005; Schirmer & Kotz 2003, 2006). Though threatening speech has not previously been investigated, it is conceivable that sensitivity to threat cues in speech might be higher in one sex than the other. While we would hesitate to classify threatening speech as a form of 'emotional speech' in any straightforward way, sex-dependent differences between listeners in the present study would accord with the findings of existing research on vocal correlates of emotion, which has often included anger among the set of emotions tested for. For instance, Bonebright *et al*. (1996), who investigated sex differences in the perception of vocal affect using recordings of acted portrayals of fear, anger, happiness, sadness and a 'neutral' emotional state, found that 'male actors were better at portraying anger than females, and… anger was the only emotion which males were better at identifying than females' (1996: 440). They account for this asymmetry by pointing to differences in the socialisation of children, asserting that 'males are encouraged to express anger and to control the display of other emotional states', and citing anthropological research on hunter-gatherer societies in which 'males were taught to express threats and show no fear' (Miller 1928, cited in Bonebright *et al*. 1996: 441).

Neurophysiological studies reviewed by Schirmer & Kotz (2006) provide evidence of sex differences in the degree of automaticity with which listeners respond to 'emotional-prosodic information' in speech processing, leading them to conclude that '[b]ased on these findings one can assume that emotional expressions and thus social interactions are of greater significance to women than to men' (2005: 27), and that women's 'greater interest in affiliation might make [them] more dependent than men on the emotional state of others, which might in turn enhance their perception of emotional cues' (2005: 27). On the behavioural side, it is argued by Lochman *et al*. (2006) that females exhibit more of a tendency than do males towards 'relational aggression', defined as '[harming] others not through the use of physical violence, but instead through acts that damage peer relationships or threaten to do so' (2006: 116). Thus, we might reasonably predict that in the present experiment the responses gathered from men and women will differ, and will do so in ways that could be accounted for in terms of gender socialisation factors (e.g. that from a young age boys engage in physically threatening behaviour more often than girls do and are thus more highly sensitised to the signals associated with an impending physical attack, or that girls are more attuned to emotional-prosodic information than boys are, and can read these cues more accurately).

With these considerations in mind, we proceed to a description of the experiment itself, in which the above variables and others we controlled for are discussed in more detail.

4. Methods

4.1. Materials

The recordings used for the experiment were collected from 12 male speakers (8 native British English speakers and 4 native or near-native speakers of Norwegian, Finland-Swedish, Hebrew, and Arabic). The sentences read by the speakers of the languages other

than English were direct translations of the English versions (see below). No attempt was made to control for habitual voice quality across the talkers.

Owing to the heavy demands that having to listen to a full set of stimuli would make on our participants, we did not have every speaker perform every reading task. Table 1 shows which reading tasks were carried out by which speaker. The 'threatening script' mentioned in the rightmost columns in Table 1 is described in the text below the table.

Speaker	**Language**	**Task**			
		A	**B**	**C**	**D**
		Say the phrase 'I know where you live' (or foreign language equivalent)	Say the phrase 'I know where you live' (or foreign language equivalent) in a threatening way	Read extract from a non-threatening script (or foreign language equivalent)	Read extract from threatening script (or foreign language equivalent)
1	English	x	x		
2	English	x	x		
3	English	x	x		
4	English	x	x		
5	English			x	x
6	English			x	x
7	English			x	x
8	English			x	x
9	Finland-Swedish	x	x		
10	Norwegian	x	x		
11	Hebrew			x	x
12	Arabic			x	x

Table 1: Reading tasks performed by each of 12 speakers (8 English speakers, 4 speakers of other languages).

The phrase 'I know where you live' (and its translated equivalents) was chosen as although it would be hard to argue that it intrinsically constitutes a threat *per se*, it nevertheless has ample potential as an indirect threat, under certain circumstances. Listening to a speaker reading the sentence on its own in a 'neutral' way (Task A) ought not to suggest to listeners that they are hearing the speaker threatening someone. A higher threat rating for the same isolated sentence read in a threatening way (Task B) would therefore indicate that participants had responded as expected to changes in the speaker's articulation of the utterance. Similarly, a consistent difference between listener's ratings for the neutral and

induced-threat versions of the test sentence in some or all of the other four languages would suggest that it is changes in the phonetic form – i.e., the segmental and prosodic features of the utterances – that are responsible for changes in the way the utterances are rated.

As shown in Table 1, half of the speakers (four of the English speakers, the Finland-Swedish speaker and the Norwegian speaker) were given the instructions *Say the phrase 'I know where you live'* and *Say the phrase 'I know where you live' in a threatening way*, in that order. These speakers read from 'cold': they were offered no other explicit directions, and were free to read the phrases as they saw fit. They had not been informed in advance that the purpose of the study was to investigate perceptions of threatening speech, so had no reason to assume that the first version of the phrase should be read in anything other than a neutral or default manner. Giving more guidance on what was meant by 'a threatening way' might also have been leading, and it was left to the speaker to decide how to produce the phrase as though intended as a threat.

The other six speakers (i.e. the other four English speakers, the Hebrew speaker and the Arabic speaker) were presented in Tasks C and D with an imaginary backstory and then asked to read aloud the sentences (1)-(3) in each task, as follows:

Task C (neutral)
Answerphone message/voicemail in which you and some friends are inviting another friend for an afternoon at the park. The friend's house is situated on the way there, so you offer him/her a lift in your car.

(1) We're going to the park for a picnic, should we come and get you? [pause]
(2) I know where you live [pause]
(3) So we'll come and pick you up

Task D (induced-threat)
Answerphone message/voicemail threatening someone who owes you a large sum of money, and refuses to pay it back to you.

(1) You better watch out [pause]
(2) I know where you live [pause]
(3) If you don't pay me back, I'll know where to find you

For the foreign language samples, the scripts were translated by native or highly-proficient users of the language in question. The three sentences in Task C were designed to be produced by the speaker in a non-threatening manner, without any suggestion of animosity. By contrast, while there is nothing in the wording of Task D that directly encodes a threat of violence, the speaker had been encouraged by the backstory to develop a sense of being justified in leaving the threatening message, as though the money had been 'stolen' from him. At no point was the reader of Task D instructed to try to sound angry. Both scenarios were designed to represent realistic situations that participants could imagine themselves being involved in.

The sentence *I know where you live*, and its foreign language equivalents, were extracted from the Tasks C and D readings, such that listeners would hear the sentence in isolation from its accompanying context. Extraction of the sentence was facilitated by the pauses left by speakers on either side of it.

The purpose of implementing these two differing strategies for obtaining the speech samples was to establish whether embedding the target sentence in a script would result in more consistent perceptions of threat in the induced-threat sentences (Task D) by listeners than in the parallel case where the sentences were read without an accompanying context. Providing speakers with scenarios in which the sentence might plausibly be spoken under real circumstances would, it was thought, enable them to draw upon their own experiences and perhaps also media depictions of threatening behaviour, and to produce the utterances in a more naturalistic way.

The recordings were made using the audio capture capabilities of a Toshiba Satellite C850–10C personal computer, with the speaker positioned approximately 60cm from the microphone. Because recordings in forensic casework are generally of poor technical quality it was in the interests of realism not considered a priority to obtain the best possible quality recordings, but all the same the recording sessions took place in environments with low levels of background noise. The recorded sentences were then embedded in an online questionnaire via which participants' responses were logged.

4.2. Participants

A total of 30 participants (13 male, 17 female) were approached via online social media to complete an online survey. All were native English speakers aged 18 years or over with little or no proficiency in a language other than English, and no reported hearing difficulties. The majority of participants had no formal linguistic or phonetic training.

4.3. Procedure

Participants were asked to complete a questionnaire in which they would rate the following: (a) the degree to which a sample of speech seemed threatening (the 'threat' rating), and (b) the degree to which they believed the speaker intended harm (the 'intent' rating). They responded using an on-screen horizontal sliding scale running between 0 (for parameter (a), 0 represented 'not at all a threat') and 100 ('very likely a threat')), allowing participants a degree of precision in their responses. By default, the slider was set at 0 when presented to the participant. Each listener rated 24 sentences apiece (16 English, 8 foreign language), yielding responses for a total of 720 sentences.

Participants were also offered the opportunity to give a description of the voice they had heard. As the majority of participants had had no previous linguistic training, these responses were expected to resemble typical lay-listener responses. For the foreign language utterances, participants were also asked whether they knew what language had been spoken, and if so, to name it.

Foreign language samples were played in randomised order before listeners heard any English tokens (also randomised), in an attempt to ensure that participants' responses to the foreign language stimuli were not biased by their having heard the English material beforehand.

The results collected from the online survey were subjected to ANOVA testing in order to establish the strength of the effects of the independent variables (e.g. language spoken) on the threat and intent ratings. Pearson's product-moment correlation coefficients were also calculated to gauge whether listeners' threat and intent scores were correlated. Although it is recognised that individual respondents may have used the scale in different ways (e.g. some

may have avoided giving responses at the extreme ends of the scale), we did not attempt to normalise the responses in any way, for example by expressing listeners' individual response values relative to the range of values each listener used. Correcting for individual variation in this way may yield a different picture and it is worth considering applying such a method in future research on this topic.

5. Research hypotheses

The hypotheses we are seeking to test in the current research are as follows:

1. Induced-threat utterances will be identified as such – that is, listeners' ratings for these sentences on the threat scale (and probably also the intent scale) will be higher than those for their 'neutral' counterparts.
2. Male and female listeners will rate the utterances differently, in line with previous experimental literature showing gender differences in the perception of emotional speech.
3. Utterances produced by speakers who were provided with the scenario plus three-sentence script to read from will yield consistently higher threat and intent ratings than will induced-threat utterances that were read from cold, i.e. without any supporting backstory/context.
4. Foreign language utterances will be rated by listeners for threat and intent in the same way as their English equivalents are, because the set of phonetic cues used to express threat will be approximately the same across the five languages. Put another way, the linguistic content of the utterance being rated for threat and intent is immaterial: the listener will rate the utterance on the two scales solely on the basis of its phonetic form.
5. The threat and intent ratings will be correlated, because the more threatening a utterance is perceived to be, the more listeners will believe that the speaker intends to carry out the threat.

In the following sections we discuss the level of support given to each of these hypotheses by the data collected for the purpose.

6. Results

Data collected from the online survey were separated into two groups: *English* and *foreign language*. We look first at the differences that can be observed among the threat and intent ratings for the neutral and induced-threat utterances in the *English* condition.

Figure 1 summarises the results for the English utterances where listeners' responses on the threat and intent scales are presented by listener sex and by production mode ('neutral' – Tasks A and C readings pooled – versus 'induced-threat' (Tasks B and D) sentences pooled). It can immediately be seen that the average scores are all rather low on the 0–100 scale, showing that listeners were, overall, not strongly inclined to hear the utterances as threatening. This is to be expected for the neutral utterances produced for Tasks A and C, though more surprisingly they were still rated some way above zero ('not at all a threat'). Another conspicuous trend is that, in both the neutral and the induced-threat conditions, male listeners awarded higher threat and intent ratings than did females, more particularly in the induced-threat condition. This difference is supported by the result of ANOVA testing, which reveals there to be a highly significant effect for listener sex ($F = 20.289$; $p < .0001$). In all cases, the average scores for threat are marginally higher than those for intent, except in the case of the male listeners' responses to the induced-threat utterances, for which the average

level of perceived threat is more than 6% higher than that for intent. The male listeners' average score for threat among the English induced-threat utterances attains nearly 50%, indicating that the male listeners appear to have perceived these utterances to be considerably more threatening than did the female listeners ($F = 55.527$; $p < .0001$). Listener sex is shown also to have a strong effect upon intent scores ($F = 10.642$; $p = .001$), and for intent the effect of production mode (neutral vs. induced-threat) is also highly significant ($F = 31.381$; $p < .0001$).

6.1 English stimuli condition

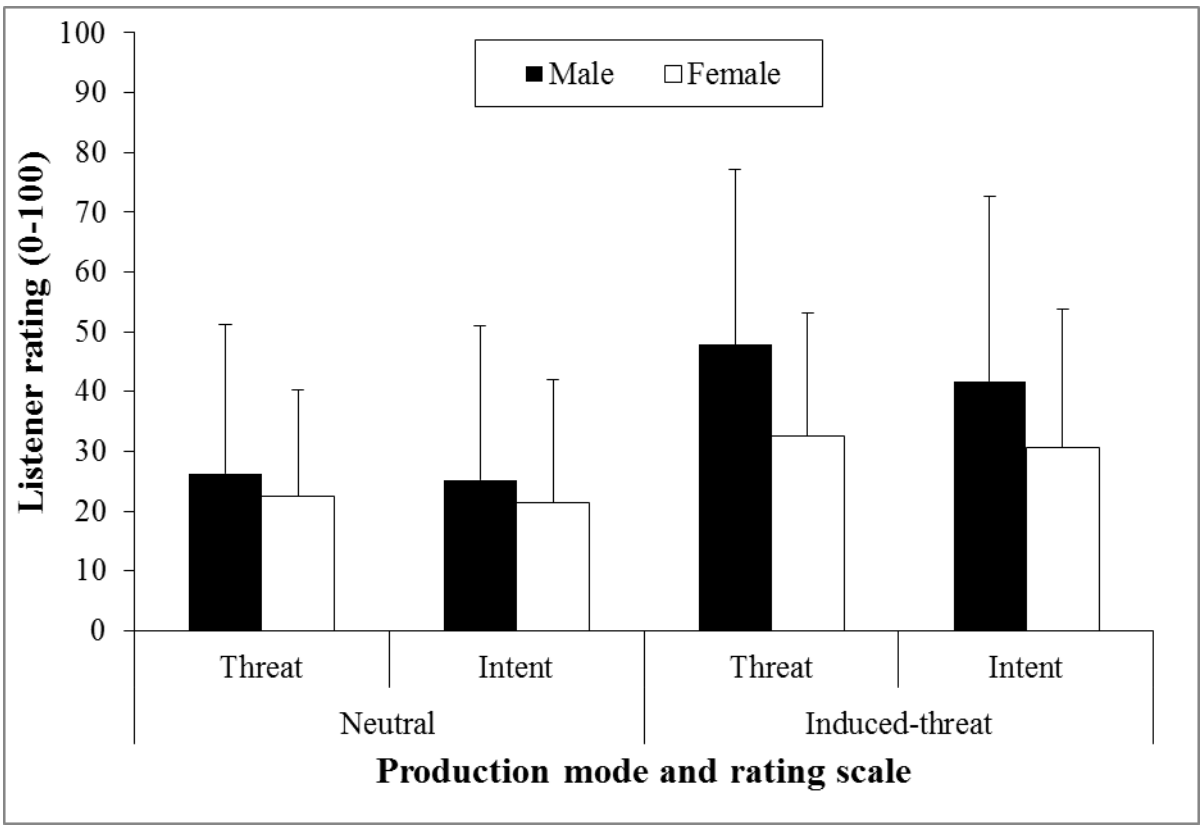


Figure 1: Threat and intent ratings (%) for Task A-D sentences in the English condition, by listener sex. Error bars represent one standard deviation.

The error bars in Figure 1 indicate that there is considerable variation in the individual scores, and a good deal of overlap between the ratings given by the men and the women is also evident. However, two of our hypotheses have been supported by the English data: it appears that listeners can distinguish neutral from induced-threat utterances, and that male listeners consistently give higher ratings (albeit often just marginally higher ones) for both threat and intent than females do, even if the sentences they were rating are not intended to sound threatening.

6.1.1 The influence of scripted context

In this section we investigate whether listeners gave different ratings on the threat scale for the same target sentence *I know where you live* when the sentence is read cold (Tasks A and

B) versus when it is embedded in a three-sentence script accompanied by a contextualising backstory (Tasks C and D; see Section 4.1).

Four of the eight English-speaking readers were given the Task C and D materials to prepare for the recordings. It was thought that being given brief details of a plausible scenario which might prompt a person to say the target sentence would help the speaker to produce the neutral and induced-threat utterances more 'naturally'.

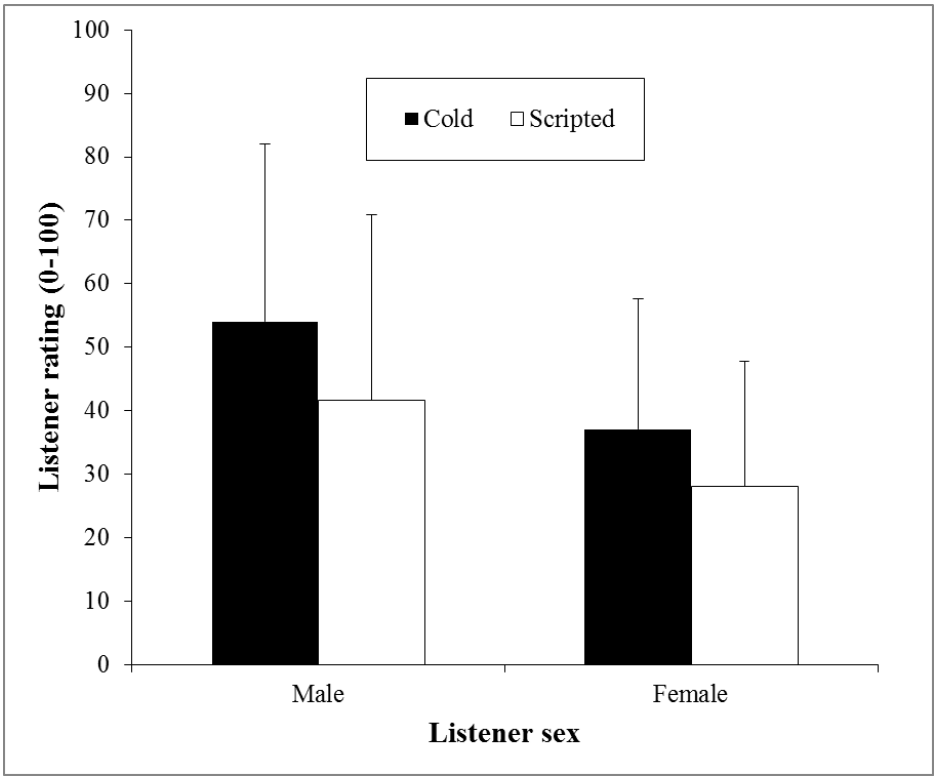


Figure 2: Average perceived threat ratings (%) for the induced-threat target sentence spoken 'cold' (Task B) and in a scripted context (Task D), by listener sex (English condition). Error bars represent one standard deviation.

An ANOVA test on the listeners' threat ratings of the English sentences showed there to be a significant effect for scripted context versus cold reading ($F(8.541) = 1$, $p < .004$). At first glance, this might appear to bear out our third hypothesis. However, the sentences produced by speakers who performed Task B (cold reading, induced-threat) were on average rated higher on the threat scale than those spoken by Task D readers (scripted context, induced-threat), as can be seen in Figure 2.

Owing to the subjective nature of the listening task we cannot identify an obvious explanation for why the pattern in Figure 2 is the reverse of that expected, but it is possible that when reading the sentence in the absence of accompanying context the Task B speakers exaggerated the phonetic cues they used to indicate that the sentence was intended as a threat, whereas in Task D they relied more upon the context to communicate to the recipient that *I know where you live* was meant to represent an indirect threat to harm him or her (note that speakers were not told that the second of the three sentences would be excised from the recording and played to listeners in isolation, so may have assumed that listeners would hear all three sentences). It is also possible that the readers felt more constrained by the context in

Task D, such that they felt less inclined to adopt what they would consider a threatening tone of voice.

When the data are viewed as shown in Figure 2 we can again see that male listeners tended to give higher threat ratings than did female listeners. For the Task B 'cold' readings, the male listeners' average score exceeded the 50% mark, indicating that they believed the utterances to merit ratings closer to the 'very likely to be a threat' end of the scale than to the 'not at all a threat' end.

We next evaluate the listeners' judgments of the foreign language sentences, to see whether the data give support to hypothesis (4), which asserts that in the current experiment the language in which translations of the neutral and induced-threat utterances are delivered will have no influence on listeners' threat and intent ratings.

6.2 Foreign language condition

The foreign language judgments were initially aggregated into a single group on the grounds that as the listeners had had little foreign language experience, there was no *a priori* reason to assume that the foreign language sentences would be treated differently (other than that they were spoken by different individuals). Therefore, they were considered collectively for the purposes of comparison with the English samples.

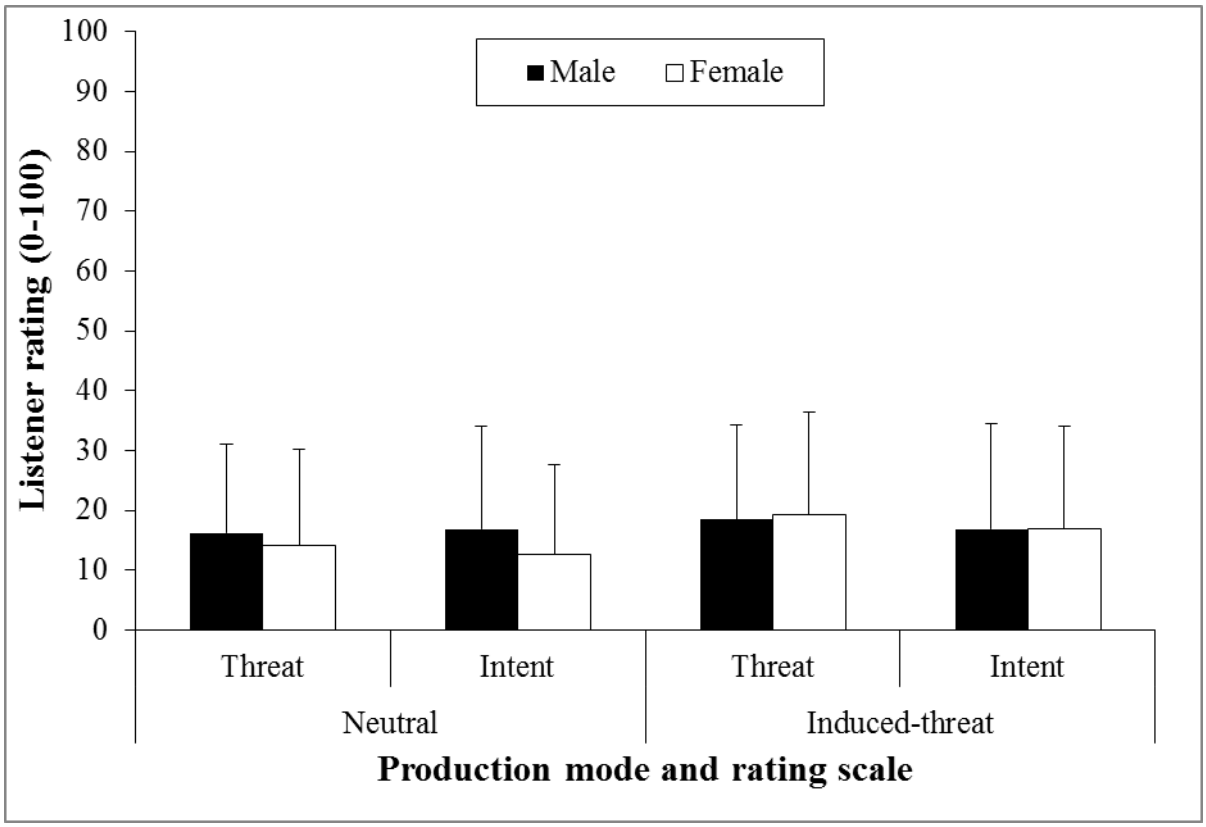


Figure 3: Threat and intent ratings (%) for Task A-D sentences in the foreign language condition (results for all 4 languages pooled), by listener sex. Error bars represent one standard deviation.

Figure 3 shows there to be practically no difference in the listeners' threat and intent ratings across the neutral and induced-threat conditions. The ratings are low throughout; no average rating exceeds 20%, and the error bars imply that even the highest ratings for the induced-threat utterances did not often surpass 50% on either the threat or the intent scales. Listeners, in other words, apparently did not hear much difference between the neutral and the induced threat utterances, though such differences as there are go in the expected direction for perceived threat in the induced-threat condition. Males give slightly higher average ratings for females in the neutral condition, but the reverse is true in the induced-threat condition.

6.2.1 Influence of scripted context

As we saw in Figure 3, the threat ratings for the induced-threat utterances (those spoken in Tasks B and D) in the foreign languages are generally very low. It appears to make practically no difference to listeners' ratings whether the test sentences are read 'cold' in isolation or embedded in a script when read aloud (Figure 4; note that the Task B readings were for Norwegian and Finland-Swedish only, while those for Task D were in Hebrew and Arabic only).

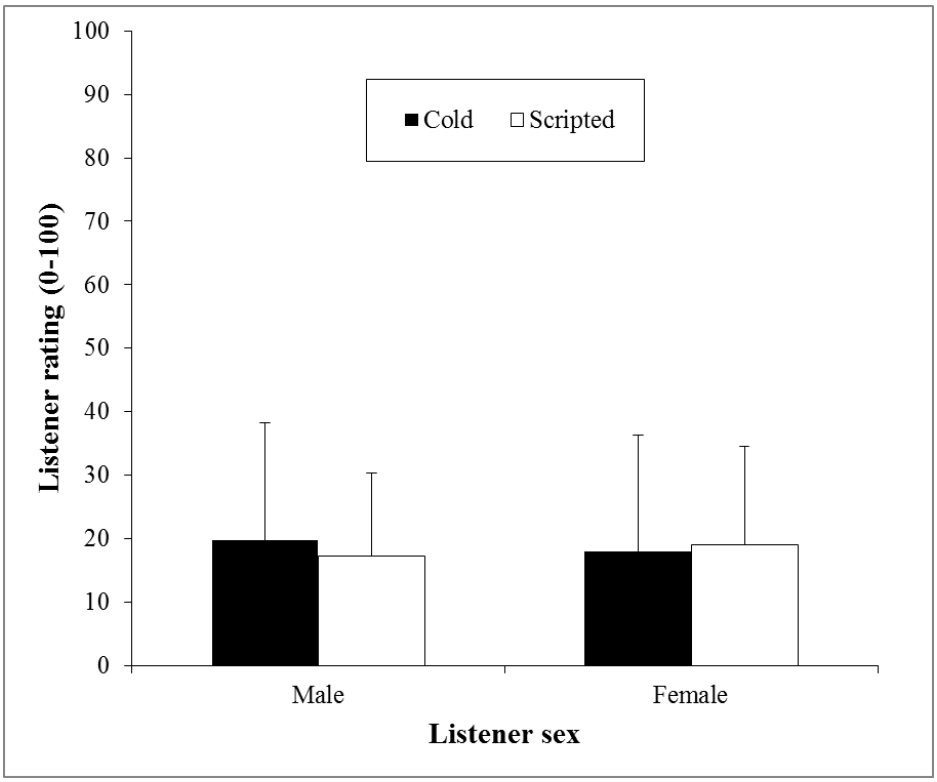


Figure 4. Average perceived threat ratings (%) for the induced-threat target sentences spoken 'cold' (Task B; Norwegian and Finland-Swedish only) and in scripted contexts (Task D; Arabic and Hebrew only), by listener sex.

The pattern for the male listeners, whereby the cold reading is rated higher for perceived threat than its scripted equivalent, recalls that seen for the English language data. However, the difference is only marginal. That for the female listeners goes in the opposite (but initially expected) direction. Neither difference is found to be statistically significant, and the male listeners' average scores are not significantly different from the females' scores either.

The general tendency overall, then, is for listeners to treat the foreign language utterances as lacking in perceived threat. Perceived intent is correspondingly low. The latter is understandable, because it is hard to imagine a scenario in which intent to harm would consistently be rated higher than perceived threat for any given utterance or set of utterances. It seems probable, nonetheless, that scores on the two scales will be correlated: an utterance which is perceived to be low for threat will tend to be rated low for intent, while utterances judged to sound highly threatening to listeners will be more likely to yield high ratings for intent. We examine the evidence for this assumption – our hypothesis (5) – in the following section.

6.3 Correlations between threat and intent

There was found to be a very strong positive correlation between the scores on the threat and intent scales overall (Pearson's $r = 0.883$, df = 718, $p < .0001$),[2] with the great majority of threat ratings (73.9%) being either equal to or larger than the corresponding intent rating for the same utterance rated by the same listener.

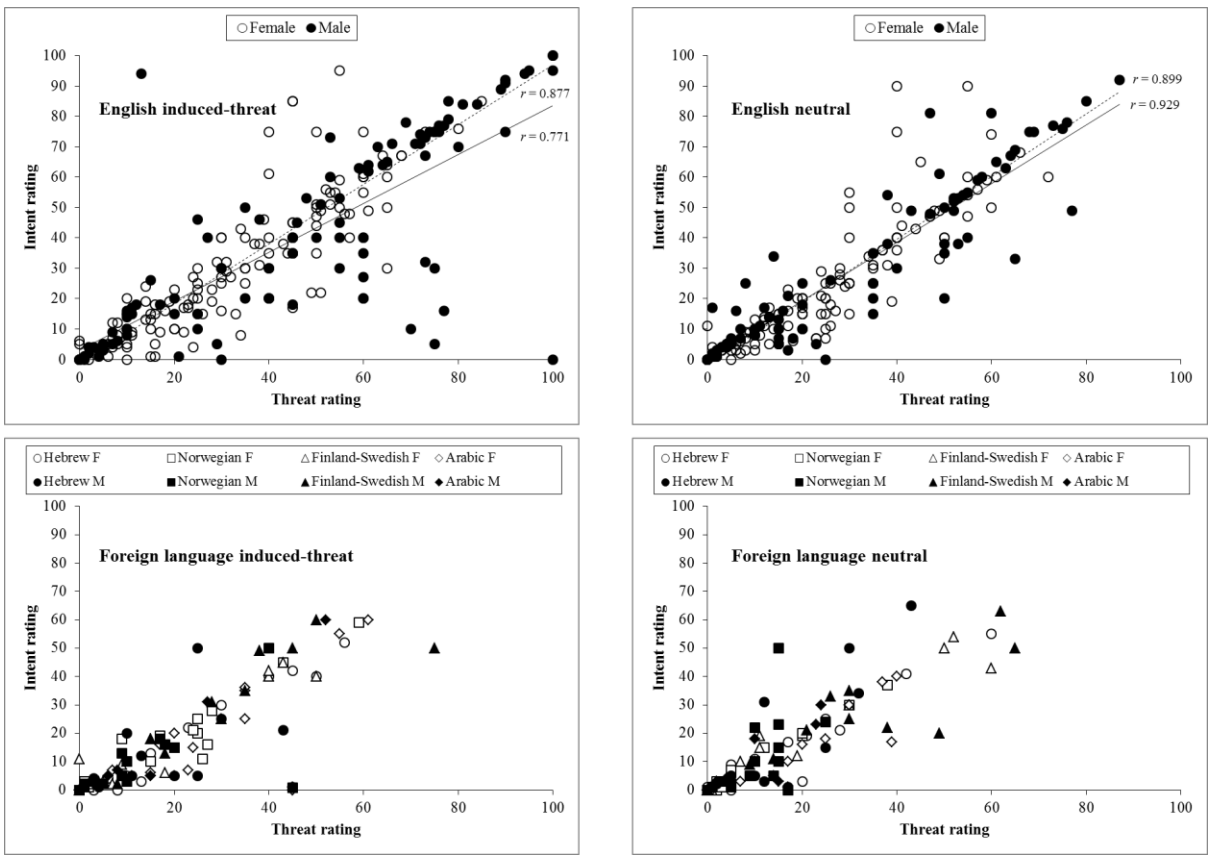


Figure 5. Correlations of threat and intent ratings for induced-threat and neutral utterances in English and four foreign languages, by listener sex. Solid trendlines in the English condition are for male listeners, dashed lines are for female listeners.

Figure 5 shows all 720 individual pairs of ratings divided by language (English in the top two panes, and the foreign languages in the lower two panes), by listener sex (where male

[2] Correlation coefficients were calculated using Wessa (2012).

listeners are represented by black markers, and female listeners by white ones). The two left-hand panes represent the results for the induced-threat conditions, and those on the right show the data in the neutral conditions. The results for the four foreign languages are displayed separately in the lower two panes. Regression trendlines and correlation coefficients (Pearson's *r*) are shown in the upper two panes (English conditions) for the male and the female listeners separately.

Positive correlations are evident in all of the data sets represented in Figure 5. For the English data (top panes) the match between the ratings for threat and intent appears to be tightest for high and low values, with more variability in the middle of the range. In the English induced-threat condition, it tends to be male listeners who award high ratings on both scales. For given utterances in this condition, male listeners tend to give higher ratings for threat than for intent in the mid-range area, while for female listeners it is the other way round. In the English neutral condition, the scores given tend to fall lower on each scale (cf. Figure 1), and there is greater agreement within each listener group, as reflected by the higher correlation coefficients. The correlations between the sex-group scores for threat and intent in both the induced-threat and neutral conditions were for English very highly significant (induced-threat, male listeners, $r = 0.771$, df = 102, $p < .0001$; induced-threat, female listeners, $r = 0.877$, df = 134, $p < .0001$; neutral, male listeners, $r = 0.929$, df = 102, $p < .0001$; neutral, female listeners, $r = 0.899$, df = 134; $p < .0001$).

Similarly, when the foreign language data are pooled the correlations between threat and intent scores for the male and female listeners groups were very highly significant (induced-threat, male listeners, $r = 0.858$, df = 50, $p < .0001$; induced-threat, female listeners, $r = 0.96$, df = 66, $p < .0001$; neutral, male listeners, $r = 0.79$, df = 50, $p < .0001$; neutral, female listeners, $r = 0.959$, df = 66, $p < .0001$). As we saw in Figure 3, the foreign language sentences were overall rated lower than their English counterparts for both threat and intent, but for individual utterances listeners still consistently gave scores on each scale that agreed with one another quite closely. Owing to a lack of space it is not possible to show regression lines for each of the languages in the plots themselves, but the correlations for all language/listener sex group/production mode pairings of threat and intent ratings (e.g. female listeners' ratings for the Norwegian neutral utterances) were shown to achieve significance at the 5% level. The loosest correlation was for the male listeners' threat and intent ratings for the Hebrew induced-threat utterances ($r = 0.578$, df = 11, $p = .039$). The responses for this particular condition show that listeners mostly rated the utterances low for threat, and that in the main the listeners' ratings on the threat scale were higher than those on the intent scale. While listeners tended to perceive threat in these utterances, then, we might say that they did not perceive them to be delivered with much conviction.

7. Discussion

The results discussed in the preceding section bear out the majority of our initial research hypotheses. The first of these, which proposed that listeners will show that they can distinguish induced-threat utterances from their neutral counterparts by assigning utterances of the former type higher ratings for threat and intent than those they gave to the latter, is clearly supported by the English data. The differences in the average scores for the English data are, however, not as big as we might have expected, for two reasons: the neutral utterances were perceived as threatening and intentful to a greater degree than anticipated, and the induced-threat utterances were apparently not perceived to be as threatening and intentful as we had thought they would be. Male listeners returned higher scores on both

scales than female ones did, on average, particularly for those sentences produced in the induced-threat mode.

The corresponding scores for the foreign language utterances, by contrast, exhibit practically no differences of this kind. The induced-threat utterances are rated only fractionally higher on the two scales than the neutral ones, and the scores are in any case very low. The responses for the two listener sex groups are closely comparable with one another. It appears, then, that English-speaking listeners hearing target sentences spoken in an unfamiliar language cannot decide whether they are hearing a neutral or a threatening utterance. The average scores on the two scales for utterances in both production modes are in the region of 15–20%, so it is not the case that they are heard as completely lacking in threat or intent, but the scores may just reflect the fact that listeners were aware of the nature of the experiment and so were more strongly inclined to assign threat and intent ratings to the stimuli than they might have been had it not been suggested to them that the criteria of interest related to the making of verbal threats.

Our second hypothesis, that men and women would respond differently to the stimuli when rating the utterances for perceived threat and intent, is also supported by the English data, but not to anything like the same degree by the foreign language data. These patterns are clearly visible in Figures 1–4, and to a lesser extent in Figure 5. We might tentatively conclude that male listeners are more sensitive than women to threatening content in speech produced by other males, in line with the arguments about sex differences in the perception of emotional speech put forward by Bonebright *et al*. (1996), Schirmer & Kotz (2005), Lochman *et al*. (2006), and others. Males are more likely than females to be the targets of physical attacks (Archer 2004) and on balance it is therefore in their interests to be alert, whether consciously or subconsciously, to signals of hostility or potential harm expressed by others, whether verbally or otherwise. It is notable that following the recording session, several of the speakers recruited for that part of the experiment reported that they had produced the target utterances as though they were addressing another man. Whether the phonetic strategies they adopted to signal threat would have differed had they been asked specifically to speak as though the target of the threat were a woman rather than a man would present an interesting line of inquiry. Future research in this area may also cast light on whether any parallel disparity in sensitivity to same-sex utterances exists, whereby female listeners might respond more strongly than male ones to sentences produced in the two modes by female talkers.

When the threat scores were partitioned according to whether the English sentence *I know where you live* or its foreign language equivalents (induced-threat mode only) were produced in isolation (‘cold’) or embedded in a script with an accompanying backstory (‘scripted’), the pattern predicted by hypothesis (3) was not observed. Rather, we found that – with the exception of the female listeners’ ratings for the pooled foreign language utterances – higher threat scores were returned for the unscripted sentences. This suggests that in the ‘cold’ readings the English speakers may have produced the target sentence using a greater number of phonetic cues that the listeners associated with threatening speech, and/or using more extreme deviations from their ‘neutral’ pronunciation setting.[3] They may have believed that the wording of the context sentences flanking the target sentence in the induced-threat mode (Task D) would make it clear that the target sentence was meant as a threat. They had not been told that the listeners would only hear the target sentence from Task D. Alternatively, the speakers may have chosen to focus more effort on making the first or third sentences of the three in the script sound threatening, such that the second was closer to a neutral reading for that speaker than it would have been when the sentence was read in isolation.

[3] We intend to investigate the nature of the phonetic cues themselves in a future study.

In the English condition, the male listeners returned significantly higher average scores on the threat scale for both cold and scripted induced-threat sentences than did the females. This pattern was not observable in the foreign language condition, however, nor was there any difference between the average scores for the cold and scripted utterances. It should be remembered, on the other hand, that the data for the four foreign languages were pooled, and that averaging the scores in this way might disguise effects that obtained for one language but not others. This question awaits further investigation.

The foregoing discussion should by now have made it apparent that the responses on the threat and intent scales differ markedly according to whether the language heard is English (the listeners' native language) or an unfamiliar foreign language. This finding runs counter to hypothesis (4). It seems that perceiving threat and intent to harm in a talker's speech depends on being able to understand the words being spoken, and is not solely a matter of attuning to a set of phonetic cues that listeners might associate with threatening speech. As before, one might argue that rolling the four foreign languages together in this fashion ignores the differences between the languages and the talkers who spoke in them. In Figure 5, one can see that in the two lower panes (representing the individual responses for foreign language induced-threat and neutral utterances) the Finland-Swedish speaker, denoted by filled and open triangles, has a tendency to be rated relatively high on both the threat and the intent scales. In the same way, we have not in the current paper paid regard to differences between the individual English speakers, although some are consistently perceived to sound more threatening and intentful than others. Confidence in reading aloud for a sound recording, experience with acting, and differing levels of exposure to real and simulated threat in the media (violent films and TV programmes, computer games, etc.) may all have had an influence on the way in which the utterances were produced in the induced-threat mode. Space constraints preclude a more nuanced treatment that would reveal how far the languages chosen and the phonetic behaviour of individual talkers contribute to the differences in listener scores, but we believe that it is important to investigate these issues in subsequent research.

The fifth of our hypotheses, which concerns the correlation between listeners' ratings for threat and intent, is strongly supported by the present data. There is indeed a close relationship between the level of threat listeners report perceiving in the test utterances, and the strength of their belief that the speaker intends to cause the recipient harm. It may be, of course, that the participants did not distinguish very clearly between threat and intent. For a threat to be perceived as such may require an assumption on the hearer's part that the speaker at least potentially wishes to harm the target of the threat, so to that extent one could for a given utterance always reasonably expect to see a value on one scale that is commensurate with a value on the other. And as noted earlier, it would also be peculiar if individual utterances were frequently rated higher for intent than for threat. However, listeners were free to choose values on the two scales independently for each stimulus they heard, and in many cases they exercised this freedom, so it is not the case that the two scales are so inextricably linked to one another that a rating for threat invariably predicts an approximately equivalent rating for intent.

8. Conclusions

In this paper we set out to investigate how English-speaking listeners respond to neutrally-worded test sentences produced by speakers of five different languages (English, Hebrew, Norwegian, Finland-Swedish and Arabic), where the perceptual criteria being tested for are

the level of perceived threat, and the level of intent on the part of the speaker to perform an act that would be harmful to the target of the utterance.

The results show, first and foremost, that the listeners identify utterances designed as threats as such much more accurately if they are spoken in English than in an unfamiliar foreign language. Given that making a threat constitutes an offence, we might conclude that in real-world scenarios we should be especially careful when evaluating witness testimony in which it is claimed that a threat was being made, where the language used is one that is unfamiliar or unknown to the witness.[4] The results of this study also suggest that we should also remain open-minded in respect of the factors that underpin a witness's claim that a threat was made where the language being used is English, in light of the fact that there is so much variability in how the individual English utterances tested here were rated for threat and intent (Figure 5).

The second key finding of the current research is that male and female listeners respond differently to the stimuli, with men tending to perceive a substantially higher level of threat and intent in the English utterances. This listener sex difference is neutralised in the foreign language results. This finding confirms some of the generalisations made in earlier research on sex differences in the perception of vocal expressions of emotion, and it links the perception of threat – which has to our knowledge not been investigated using this kind of experimental approach before – more closely to the perception of related emotions, in particular anger.

From the forensic viewpoint we might also learn usefully from the fact that the perception of threat in a given utterance seems to correlate so tightly with the perception of intent to harm. On the one hand this is not at all surprising, but it is easy to conceive of scenarios in which a defendant accused of having threatened a second party might find it difficult to claim that s/he had only meant the utterance as an empty threat. For example, if jurors are allowed to listen to a recording of the incriminating utterance and believe it to sound highly threatening, they are more likely to perceive it to signal a high level of intent on the speaker's part to harm the addressee, thereby increasing the probability of the imposition of a harsher penalty.

There is of course plentiful scope for expanding and improving upon the design of the current study. Varying the test utterance beyond a single sentence per language would be one way in which to do this. For example, it would be possible to gain a sense of how much the semantic content of the utterance contributes to its perception as a threat, regardless of its phonetic characteristics. It would be uncontroversial to say that a direct threat like 'I'm going to stab you with this knife' could be interpreted as a threat no matter how it was spoken, but a sentence as neutrally worded as 'I put the book on the shelf' would depend on some other source of information to elevate it to the status of a threat. Use of professional actors rather than untrained volunteers might also yield greater consistency between speakers and listeners in terms of how the test utterances were produced and perceived, though there might be a risk that the utterances would be performed in an excessively stylised way that would detract

[4] Interestingly, some of the listeners reported hearing foreign language utterances *as* English ones, when they were asked if they were familiar with the language being used. Eight of 17 listeners believed that the Hebrew utterances were English or 'Scottish', with three transcribing the Hebrew target sentence /ʔani jodeʔa ʔefo ʔata ɡaʁ/ as the English phrase 'And you're there for the girl'. Four listeners heard the Arabic utterance as English. One of these, who had also heard the Hebrew sentence as English, offered the transcription 'An ad for the secretary' for the Arabic sentence /ʔana ʕa:rif inta sa:kin fe:n/).

from their forensic realism (cf. Roberts 2011). The majority of people who threaten one another with harm are not trained actors, after all.

Benefit might also be gained by using foreign languages which are sufficiently familiar to listeners that stereotypes of the language or of people who speak it would come into play. Additionally or alternatively, languages which listeners might have difficulty identifying by their spoken forms alone might be named to one half of the listener group, to see whether knowing the name of the language biases listeners towards higher or lower threat/intent ratings. Showing pictures of individuals from different ethnic or religious groups during exposure to the auditory stimuli might provoke similar changes in threat perception, while a greater range of regional and social accents of English would also no doubt provide insights into how some sociolects are associated more strongly with criminality and other anti-social tendencies (e.g. Dixon *et al*. 2002).

The subjectivity of emotion perception makes it a difficult field in which to draw firm conclusions, but without systematically-collected data on which to ground a more extended programme of research we cannot progress from the situation we are currently in, whereby the classification of utterances as indirect threats in courts of law is entirely a matter of one person's opinion against another. It is important to reiterate that we do not wish to claim that we yet have any objective grasp of what phonetic cues turn an innocuous, neutral statement into an utterance that makes the recipient fear for his or her safety. Rather, we see the results reported in this paper as a first step towards gaining a better, less simplistic understanding of how the interplay of intentions, actions and responses on the part of speakers and listeners conspire to create this particular kind of verbal crime.

References

ARCHER, JOHN. 2004. "Sex differences in aggression in real-world settings: A meta-analytic review". *Review of General Psychology* 8(4): 291–332.

AUSTIN, JOHN L. 1962. *How to Do Things with Words*. Cambridge, MA: Harvard University Press.

BOLINGER, DWIGHT L. 1989. *Intonation and its Uses: Melody in Grammar and Discourse*. Stanford: Stanford University Press.

BONEBRIGHT, TERRI L., THOMPSON, JERI L. & LEGER, DANIEL W. 1996. "Gender stereotypes in the expression and perception of vocal affect". *Sex Roles* 34(5/6): 429–445.

DIXON, JOHN A., MAHONEY, BERENICE & COCKS, ROGER. 2002. "Accents of guilt: Effects of regional accent, race and crime type on attributions of guilt". *Journal of Language and Social Psychology* 21(2): 162–168.

FRASER, BRUCE. 1976. "Warning and threatening". *Centrum* 3(2): 169–180.

FRASER, BRUCE. 1998. "Threatening revisited". *International Journal of Speech, Language and the Law* (formerly *Forensic Linguistics*) 5(2): 159–173.

GALES, TAMMY. 2012. "Linguistic analysis of disputed meanings: Threats". In: Carol A. Chapelle (ed), *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell. Online resource. DOI: 10.1002/9781405198431.wbeal0711

GINGISS, PETER. 1986. "Indirect threats". *Word* 17(3): 155–158.

GOBL, CHRISTER & NÍ CHASAIDE, AILBHE. 2003. "The role of voice quality in communicating emotion, mood and attitude". *Speech Communication* 40: 189–212.

GUSSENHOVEN, CARLOS. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

LOCHMAN, JOHN E., POWELL, NICOLE R., CLANTON, NANCY & MCELROY, HEATHER K. 2006. "Anger and aggression". In: George G. Bear & Kathleen M. Minke (eds), *Children's Needs III: Development, Prevention, and Intervention*. Washington, DC: National Association of School Psychologists. pp. 115–133.

MILLER, NATHAN. 1928. *The Child in Primitive Society*. New York: Brentano.

NAPIER, MICHAEL R. & MARDIGIAN, R. STEPHEN. 2003. "Threatening messages: The essence of analyzing communicated threats". *Public Venue Security*, September/October, pp. 16–19.

POLZEHL, TIM, SCHMITT, ALEXANDER, METZE, FLORIAN & WAGNER, MICHAEL. 2011. "Anger recognition in speech using acoustic and linguistic cues". *Speech Communication* 53(9-10): 1198–1209.

PUBLIC ORDER ACT. 1986. London: Her Majesty's Stationery Office. Online resource: http://www.legislation.gov.uk/ukpga/1986/64/pdfs/ukpga_19860064_en.pdf

ROBERTS, LISA. 2011. "Acoustic effects of authentic and acted distress on fundamental frequency and vowel quality". *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, August, pp. 1694–1697.

ROTHCHILD, JOHN. 1998. "Menacing speech and the First Amendment: A functional approach to incitement that threatens". *Texas Journal of Women and the Law* 8: 207–223.

SALGUEIRO, ANTONIO B. 2010. "Promises, threats, and the foundations of Speech Act Theory". *Pragmatics* 20(2): 213–228.

SCHIRMER, ANNETT & KOTZ, SONJA A. 2003. "ERP evidence for a gender specific Stroop effect in emotional speech". *Journal of Cognitive Neuroscience* 15(8): 1135–1148.

SCHIRMER, ANNETT & KOTZ, SONJA A. 2006. "Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing". *Trends in Cognitive Sciences* 10(1): 24–30.

SCHIRMER, ANNETT, ZYSSET, STEFAN, KOTZ, SONJA A. & VON CRAMON, D. YVES. 2004. "Gender differences in the activation of inferior frontal cortex during emotional speech perception". *Neuroimage* 21: 1114–1123.

SCHIRMER, ANNETT, KOTZ, SONJA A. & FREDERICI, ANGELA D. 2005. "On the role of attention for the processing of emotions in speech: Sex differences revisited". *Cognitive Brain Research* 24(3): 442–452.

SEARLE, JOHN R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

SEARLE, JOHN R. 1975. "Indirect speech acts". In: Peter Cole & Jerry L. Morgan (eds), *Syntax and Semantics, vol. 3: Speech Acts*. New York: Academic Press. pp. 59–82.

SEARLE, JOHN R. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge: Cambridge University Press.

SHAVELL, STEVEN. 1993. "An economic analysis of threats and their illegality: Blackmail, extortion and robbery". *University of Philadelphia Law Review* 141: 1850–1903.

SHUY, ROGER W. 1993. *Language Crimes: The Use and Abuse of Language Evidence in the Courtroom*. Oxford: Blackwell.

SHUY, ROGER W. 2005. *Creating Language Crimes: How Law Enforcement Uses (and Misuses) Language*. Oxford: Oxford University Press.

SMITH, SHARON S. 2008. "From violent words to violent deeds? Assessing risk from FBI threatening communication cases". In: Meloy, J.R., Sheridan, L. & Hoffmann, J. (eds.), *Stalking, Threatening, and Attacking Public Figures: A Psychological and Behavioral Analysis*. Oxford: Oxford University Press. pp. 435–455.

STOREY, KATE. 1995. "The language of threats". *International Journal of Speech, Language and the Law* (formerly *Forensic Linguistics*) 2(1): 74–80.
TIERSMA, PETER M. & SOLAN, LARRY M. 2012. "The language of crime". In: Peter M. Tiersma & Larry M. Solan (eds), *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press. pp. 340–353.
WESSA, PATRICK. 2012. *Pearson Correlation (v1.0.6) in Free Statistics Software (v1.1.23-r7), Office for Research Development and Education*. Online resource: http://www.wessa.net/rwasp_correlation.wasp
XU, YI & KELLY, ANDREW. 2010. "Perception of anger and happiness from resynthesized speech with size-related manipulations". *Proceedings of Speech Prosody 2010*, Chicago, 100027: 1–4.
XU, YI, KELLY, ANDREW & SMILLIE, CAMERON. 2013. "Emotional expressions as communicative signals". In: Sylvie Hancil & Daniel Hirst (eds), *Prosody and Iconicity*. Amsterdam: Benjamins. pp. 33–60.
YAMANAKA, NOBUHIKO. 1995. "On indirect threats". *International Journal for the Semiotics of Law* 8(2): 37–52.

Dominic Watt
Department of Language and Linguistic Science
University of York
Heslington
York
United Kingdom
YO10 5DD
email: dominic.watt@york.ac.uk

Sarah Kelly
Department of Language and Linguistic Science
University of York
Heslington
York
United Kingdom
YO10 5DD
email: sk720@york.ac.uk

Carmen Llamas
Department of Language and Linguistic Science
University of York
Heslington
York
YO10 5DD
United Kingdom
email: carmen.llamas@york.ac.uk