Proceedings Paper:
Brierley, C and Atwell, ES (2007) Corpus-based evaluation of prosodic phrase break prediction using nltk_lite's chunk parser to detect prosodic phrase boundaries in the Aix-MARSEC corpus of spoken English. In: Davies, M, Rayson, P, Hunston, S and Danielsson, P, (eds.) Proceedings of CL'2007 Corpus Linguistics Conference. Corpus Linguistics 2007, 27-30 July 2007, University of Birmingham, UK. .

Corpus-Based Evaluation of Prosodic Phrase Break Prediction Using nltk_lite's Chunk Parser to Detect Prosodic Phrase Boundaries in the Aix-MARSEC Corpus of Spoken English[1]

Claire Brierley[2] and Eric Atwell[2]

1. Introduction

An automatic phrase break prediction system aims to identify prosodic-syntactic boundaries in text which correspond to the way a native speaker might process or chunk that same text as speech. In computational linguistics, Machine Learning from hand-annotated corpus data has become the de-facto standard approach to text annotation problems such as prosodic annotation. This is treated as a classification task in machine learning and output predictions from language models are evaluated against "gold standard" prosodic phrase break annotations in a speech corpus. Despite the application of rigorous metrics such as precision and recall, the evaluation of phrase break models is still problematic because prosody is inherently variable: a given linguist's set of morphosyntactic analysis and prosodic annotations for a given text may not be fully representative of the range of parsing and phrasing strategies available to, and exhibited by, native speakers. A fairer approach to evaluation requires POS tagged and prosodically annotated variants of a text to enrich the gold standard and enable more robust 'noise-tolerant' measurement of language models.

We report on experiments with the AIX-MARSEC spoken English corpus. This has already been richly annotated at several linguistic levels, allowing a range of features to be applied in Machine Learning of phrase break prediction. We have developed a rule-based prosodic phrase break predictor, which can be used to enrich the phrase-break mark-up, to expand from a single linguist's analysis to include a wider range of possible interpretations of the text. This allows for different predictions to both score well if the prosody is plausible, even if the predicted phrase breaks differ from the corpus linguist's analysis.

Prosodic phrasing is the means by which speakers of any given language break up an utterance into meaningful chunks. The term 'prosody' itself refers to the tune or intonation of an utterance and therefore prosodic phrases literally signal the end of one tune and the beginning of another. This study uses phrase break annotations in the Aix-MARSEC Corpus of spoken English as a "gold standard" for measuring the degree of correspondence between prosodic phrases and the discrete syntactic grouping of prepositional phrases, where the latter is defined via a chunk parse rule using nltk_lite's regular expression chunk parser. A three-way comparison is also introduced between "gold standard", chunk parse rule and human judgement in the form of intuitive predictions about phrasing. Results show that even with a discrete syntactic grouping and a small sample of text (around 1400 words), problems arise for this rule-based method due to uncategorical behaviour in parts of speech. Lack of correspondence between intuitive prosodic phrases and corpus annotations highlights

---

[2] School of Computing, University of Leeds
  e-mail: claireb@comp.leeds.ac.uk

the optional nature of certain boundary types. Finally, there are clear indications, supported by corpus annotations, that significant prosodic phrase boundaries occur within sentences and not just at full stops.

1.1 What are prosodic phrase boundaries?

Prosodic phrasing is a universal characteristic of language [1] and is the means by which speakers of any given language break up an utterance into meaningful chunks. One manifestation of this chunking function in English is the pause: there are perceptible stops and starts in the speech stream and this happens within as well as between utterances. The term 'prosody' refers to the tune or intonation of an utterance and therefore prosodic phrases literally signal the end of one tune and the beginning of another. In text, punctuation is traditionally used to mark such important pauses and the rules of syntax define what constitutes a sentence and thus govern the distribution of full stops. However, just as writers differ in the amount of punctuation used, so different speakers use pauses to a greater or lesser extent and therefore there is both consensus and divergence of opinion and practice at work in terms of the location of prosodic phrase boundaries, as evidenced in the literature and as this experimental study intends to demonstrate.

1.2 Corpus annotation of prosodic phrase boundaries

The standard model for prosodic annotation of machine-readable text is ToBI [2] which focuses on two types of event in the speech contour, namely pitch accents and prosodic phrase boundaries, via a discriminating set of labels for **To**(nes) and **B**(reak) **I**(ndices) as in the following example transcription [3]:

Tone Tier				L* H-		L* H-H%
Orthographic Tier	Will	You	have	marmalade,	or	jam?
Break Index Tier	**1**	**1**	**1**	**3**	**1**	**4**

Table 1: Example ToBI transcription from *Guidelines for ToBI Labelling* in [3].

The Break Index tier recognises four degrees of juncture between words in an utterance, with indices 3 and 4 locating intermediate and intonational phrases, junctures whose significance is marked by fluctuations in pitch: the phrase accent (break index 3) and the boundary tone (break index 4). These pitch accents are transcribed in the Tone tier; in the above example the word "marmalade" exhibits a low accent on the first syllable rising to a high phrase accent at the boundary site. Thus ToBI supports theories outlining a hierarchy of prosodic constituents; the existence of different boundary types is one aspect of this and will be discussed in the next two sections.

1.3 Boundary annotations in the Aix-MARSEC corpus

The Aix-MARSEC corpus [4] originates from the Spoken English Corpus [5] and its machine-readable counterpart MARSEC [6] and consists of over 5 hours of BBC radio recordings of 53 different speakers in 11 different speech styles from the 1980s. In the Aix-MARSEC project, the original prosodic annotations made by Briony Williams and Gerry Knowles have been augmented in a series of multi-level annotation tiers which cover a range of segmental and suprasegmental linguistic features. This study, however, uses the original phrase break annotations for minor and major boundaries which equate to break indices 3 and 4 in the ToBI scheme. The following sample [7] from section A of the corpus (informal news commentary) illustrates the conventions used: a single pipe symbol for minor boundaries and double pipes for major boundaries. Juxtaposed against an ordinary transcribed version of the text, it also clearly shows that more boundaries are perceived than normal punctuation would suggest and that there is no simple mapping between punctuation marks and boundary type. A ball park figure based on the complete 619 word text from which the sample is taken reveals that phrase boundaries outnumber punctuation marks in the order of 2:1 (120 and 68 respectively).

Plain text version:

'…Athens is a favorite airport for hijackers. Beirut is another easy touch, but for different reasons. Given the state of lawlessness that exists in Lebanon the uninformed outsider might reasonably expect security at Beirut airport to be amongst the tightest in the world, but the opposite is true…'

Boundary annotations:

'…Athens is a favorite airport for hijackers || Beirut is another easy touch | but for different reasons || Given the state of lawlessness | that exists in Lebanon || the uninformed outsider might reasonably expect security | at Beirut airport || to be amongst the tightest in the world || but the opposite is true ||…'

1.3 Prosodic and syntactic phrase structure

The nature of the relationship between prosody and syntax has been a continuing debate in the literature since the 1960s, with the intriguing paradox that prosodic phrasing both reflects syntactic constituency but is 'somehow fundamentally simpler' [1] - shallower and flatter than syntactic structure. This is best illustrated by example. Intuitively, we might break the following sentence up into 2 or 3 prosodic phrases:

The two-phrase version:

In the popular mythology || the computer is a mathematics machine ||

The three-phrase version:

In the popular mythology || the computer | is a mathematics machine ||

It does not matter which version we choose: prosody, and here the distribution and classification of prosodic boundaries, is less clear cut than syntax; what matters is that each chunk is meaningful in its own right and that boundaries are not aberrant occurrences as in this next version:

Nonsensical phrasing:

In the popular | mythology the | computer is a mathematics | machine |

A full parse of the above sentence from Winograd **[8]** shows that while prosodic structure is linear, syntactic dependencies create a multi-layer structure, traditionally represented as a parse tree:

Figure 1: *phpSyntaxTree* is a web application available under GNU General Public License from sourceforge.net. One departure from convention in this parse tree is the use of Brown POS tags to identify parts of speech at terminal nodes. http://sourceforge.net/projects/phpsyntaxtree

This tree was constructed from the following labeled bracket notation and uses the Brown Corpus set of POS tags **[9]** to identify parts of speech (i.e. POS) mapped to terminal nodes:

[S [PP [IN In] [NP [AT the] [JJ popular] [NN mythology]]] [NP [AT the] [NN computer]] [VP [BEZ is] [NP [AT a] [NN mathematics] [NN machine.]]]]

The example suggests that prosodic phrase breaks equate to the nodes marked in red in this bracketed notation and that they occur between large syntactic units **{NP, VP, PP, ADJP, ADVP}**. This intuition is included in the selection of features used in a recent CART (Classification and Regression Tree) model for automatic phrase break prediction **[10]** which reports a 90.8% success rate in the detection of prosodic boundaries.

1.5 Chinks 'n' chunks

A highly successful rule-based method for determining prosodic boundaries is the chink chunk rule **[11]**, in effect the mainstay of the prosody module in a Text-to-Speech (TTS) Synthesis system because prosodic phrases must be identified before they can be given an appropriate tune. The algorithm defines a prosodic phrase as a sequence of chinks (the closed class of function words) followed by a sequence of chunks (the open class of content words) and inserts a boundary whenever a content

word immediately precedes a function word. The chink chunk rule would therefore correctly identify prosodic phrases in Winograd's sentence from fig. 1:

chink chink chunk chunk	chink chunk	chink chink chunk chunk
in the popular mythology ||	the computer |	is a mathematics machine ||

Table 2: Sentence showing classification of function words as chinks and content words as chunks.

… but would not be adequate for more complex prose such as:

'…where one found in continuous speech phonetic effects that would usually be found preceding or following a pause, the phonological element of juncture would be postulated…' [12]

The crucial phrase boundary between 'speech' and 'phonetic' would not be captured via the chink chunk method but would be captured by a model incorporating classification of major syntactic units, in this case a necessary distinction between the prepositional phrase 'in continuous speech' and the object noun phrase 'phonetic effects'.

2 Experimental aims

A number of questions emerge from the discussion so far and these are now raised and cross-referenced to sections in the introduction.

2.1 To what extent can prosodic phrase boundaries be located via a major syntactic grouping like prepositional phrases?

Intuitive phrasing of Terry Winograd's sentence in section 1.4 elicited a couple of options:

The two-phrase version:

In the popular mythology \|\| the computer is a mathematics machine \|\|

The three-phrase version:

In the popular mythology \|\| the computer \| is a mathematics machine \|\|

The contention here, based on cumulative, native speaker insight into the English language, is that the boundary separating the prepositional phrase 'in the popular mythology' from the main clause 'the computer is a mathematics machine' is more important than the optional boundary between subject and predicate. This is backed up by experimental evidence from the CART statistical model referred to in section 1.4. It was decided therefore to see how far the beginnings and ends of prepositional phrases coincided with boundary annotations by two expert linguists in extracts from the Aix-MARSEC corpus of spoken English.

2.2 To what extent does shallow parsing reflect prosodic phrasing?

Python's Natural Language Toolkit **[13]**, *nltk_lite version 0.6.5* **[14]**, includes a regular expression chunk parser, where the accompanying tutorial notes explain how chunk parsing creates flat '…structures of fixed depth (typically depth 2)…' **[15]** and why it is more robust than full parsing. This description ties in with the observation in Section 1.4 about the relative simplicity of prosodic structure and led to the realization that since this method uses regular expressions over POS tags to chunk non-overlapping linguistic groupings in text, it could be used to identify prosodic phrases. There is also the tradition of shallow parsing used to capture prosodic phrasing in the durable chinks 'n' chunks algorithm. It was decided therefore to use nltk_lite's chunk parser to set up a rule which specifies prepositional phrases as the node label for chunks and to run this over extracts from the corpus. Prepositional phrases play an important role as sentence modifiers and unlike other major syntactic units (see section 1.4) have the added advantage of always beginning with a chink.

2.3 Can any underlying principles be discovered governing the distribution of major and minor prosodic phrase boundaries?

The Aix-MARSEC corpus differentiates minor and major prosodic phrase boundaries (break indices 3 and 4) in an easily detectable, straightforward manner and facilitates comparison between expert annotators. It was anticipated that analysis of the planned chunk parsing experiment would naturally lead to close scrutiny of corpus annotations so that interesting correspondences between prepositional phrases and boundary type might be observed. The discovery of such linguistic patterns in speech corpora and the subsequent process of encoding that new knowledge as rules in a computational model of prosody is an example of what Huckvale advocates as the practice and goal of speech science **[16]**.

2.4 To what extent do people agree on prosodic phrasing?

This is an open-ended question. However, as part of this experiment, the plan was to compare the author's intuitive prosodic phrasing of extracts used to that of expert annotators'. To accomplish this, plain text versions of two complete informal news commentaries from Section A of the corpus were obtained **[7]** and **[17]**. The commentaries cover mid-1980s political issues in the Middle East (A08) and South Africa (A09).

3. Experimental work

Preparatory stages in this experimental work cover some of the natural language processing tasks essential to a Text-to-Speech synthesis system, in particular the task of morphosyntactic analysis: assigning part-of-speech tags to word tokens and imposing a hierarchical structure on sequences of POS tags. However, this hierarchical structure is not a full syntactic parse as in the tree diagram in Fig. 1 but a partial chunk parse which only seeks to identify one syntactic grouping: prepositional phrases. The experiment outlined below (Fig. 2) assesses the degree of

correspondence between the beginnings and ends of prepositional phrases retrieved via the chunk parse rule and "gold standard" prosodic boundary annotations in the Aix-MARSEC Corpus.
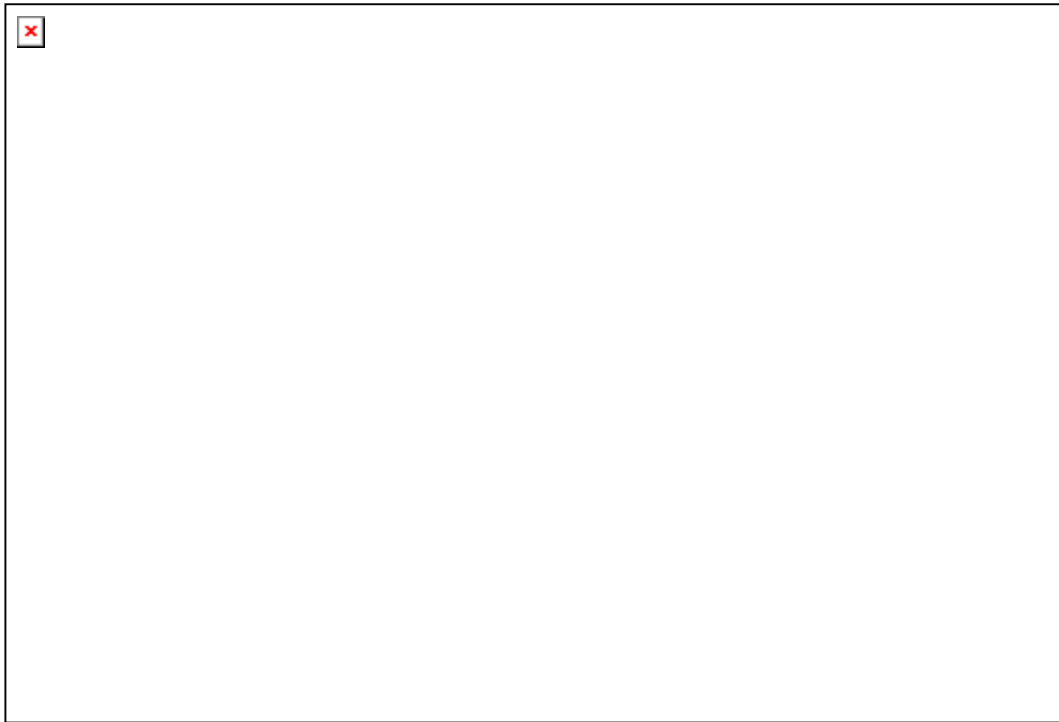

Figure 2: Experimental stages in semi-automatic POS tagging and partial chunk parsing of input text using nltk_lite.

3.1 The first step: POS tagging

The chunk parsing experiment and the comparative study of intuitive prosodic phrasing versus boundary annotations in the corpus have both been run using unpunctuated text i.e. no **{ . , : ; ? () }** as well as plain text versions with just the full stops restored. To obtain selected transcripts, the 'TextTier' was extracted from the following Notepad files in Aix-MARSEC, available in TextGrid format ready for use with Praat [18]: A0801B to A0805B, annotated by Briony Williams and totalling 619 words, plus A0901G to A0906G, annotated by Gerry Knowles and totalling 789 words. Changes to A08 in preparation for POS tagging with the Brown corpus tagset were as follows:

• **'tee double u ay'** was changed to **TWA** aircraft;

• hyphens were inserted for **'x-ray'**, **'x-rayed'** and **'check-in'**;

• enclitics such as **'that's'** and **'they've'** were restored and all apostrophes checked and left in place e.g. **'Shi'ite'** and **'hero's'**;

• subject-verb agreement was corrected in the following context: **'…hijackings from Ben Gurion…are unknown…'**

7

There are no changes to report for A09, except to say that all apostrophes were checked and left in place e.g. **'nobody else's'**.

Plain text versions of A08 and A09 were POS tagged using a composite tagger similar to the one outlined in the nltk_lite tutorial on categorizing and tagging words **[19]**. This takes the form of a bigram tagger trained on tagged extracts from the Brown corpus as "gold standard" (genres A and B, Press Reportage and Press Editorial respectively); the bigram tagger backs off to a unigram tagger trained on the same genres, which in turn backs off to a default tagger that tags everything as NN, a singular noun. Sample code listing for this, only slightly modified from the original nltk_lite tutorial notes in **[19]**, is given below and demonstrates the degree to which this toolkit is customised to NLP tasks. Here, the toolkit provides a tokenize() function, various classes of tagger and an associated train() method to facilitate the process of POS-tagging any input text.
text = sourcefile.readlines()

```
# the next line stores the input text as a list of word tokens
# in the variable: tokens
tokens = list(tokenize.whitespace(text))
my_tagger = tag.Default('nn')
unigram_tagger = tag.Unigram(backoff=my_tagger)
train_sents = list(brown.tagged(['a', 'b']))
unigram_tagger.train(train_sents)
bigram_tagger = tag.Bigram(backoff=unigram_tagger)
# the next line trains the tagger on "gold standard" tagged text
# from the Brown Corpus
bigram_tagger.train(train_sents)
# the next line stores a new version of the input text
# as a list of ('token', 'tag') tuples in the variable: tagged
tagged = list(bigram_tagger.tag(tokens))
```

The combined tagger correctly tagged 86.13% of word tokens for Aix-MARSEC A08, and 87.07% of word tokens for A09. The tagged versions of Aix-MARSEC were then hand-corrected and all the tags were capitalised ready for the chunk parser. Roughly half the tagging errors resulted from the default tagger (e.g. 'past' tagged as NN in the following phrase 'in the past two years'). Significantly, 16.28% of tagging errors in A08 and 21.57% of tagging errors in A09 were due to the word class of prepositions which could be tagged **<IN>, <RP>, <RB>, <CS>** (preposition, adverb particle, adverb or subordinating conjunction). This had repercussions for the chunk parse rule which specifies a preposition <IN> as chunk node; and it is often difficult to determine whether there is an error or not e.g. 'on' in '…Pretoria's hold on the mineral rich territory…' tagged as <RP>. This will be further discussed in Section 5.

3.2 Developing the chunk parse rule

The chunk parse rule used in this experiment was developed over several iterations on a complex test sentence of 77 words **[20]**. I have called this the imported rule. Though still a prototype, this rudimentary, catch-all formula attempts to specify the syntactic constituents of any prepositional phrase via a tag pattern, a regular expression pattern over strings of tags delimited by angled brackets **[15]** and is evidently transferable from one context to another with very little intervention. The only significant changes between the imported rule and versions A08 and A09 are that:

- coordinating conjunctions **<CC>** have been removed from the rule because they interfere with boundary prediction (see discussion in Section 5);

- as a stop-gap measure, **<PP$>** (personal pronoun: possessive) has been replaced by **<POSS>** (a made-up tag) simply because the chunk parser does not recognize the dollar symbol.

Imported rule version:

The tag pattern and description string for this rule instruct the parser to begin the chunk with a word token tagged as a preposition, and to include in that chunk any combination in any order of tokens tagged as follows: another preposition; determiner/pronoun (singular); determiner/pronoun (singular or plural); article; personal pronoun (object); nominal pronoun; determiner/personal pronoun (possessive); adjective; coordinating conjunction; noun (singular); noun (plural).

```
parse.ChunkRule('<IN><IN|DT|DTI|AT|PPO|PN|PP$|JJ|CC|NN|NNS>+',
"Chunk IN with sequences of IN, DT, DTI, AT, PPO, PN, PP$, JJ, CC, NN,
NNS")
```

A08 version:

This rule removes <CC> (coordinating conjunctions), replaces <PP$> with <POSS>, and adds the following constituents: determiner/pronoun or post determiner; cardinal number; superlative adjective; proper noun.

```
parse.ChunkRule('<IN><IN|DT|DTI|AT|AP|CD|PPO|PN|POSS|JJ|JJT|NP|NN|NNS>+'
, "Chunk IN with sequences of IN, DT, DTI, AT, AP, CD, PPO, PN, POSS,
JJ, JJT, NP, NN, NNS")
```

A09 version:

This rule incorporates the following additions: ordinal numbers and semantically superlative   adjectives.

```
parse.ChunkRule('<IN><IN|DT|DTI|AT|AP|CD|OD|PPO|PN|POSS|JJ|JJT|JJS|NP|NN
|NNS>+',
"Chunk IN with sequences of IN, DT, DTI, AT, AP, CD, OD, PPO, PN, POSS,
JJ, JJT, JJS, NP, NN, NNS")
```

3.3 Intuitive prosodic phrasing

A further aspect of this experimental work, and a means of familiarisation with the corpus, was to compare the first-named author's intuitive prosodic phrasing to that of expert annotators' and to mark out longer prosodic phrases in response to Liberman and Church's own criticism of the chink chunk rule in their original paper [11]. They consider the prosodic phrases or 'function word groups' captured by the rule to be too small to accommodate sufficient variation in prosody and are interested in discovering how these smaller units '…combine hierarchically to form sentence-sized units…' The procedure followed in the current study was to assign major and minor boundaries with the same pipe symbol notation as the corpus, using unpunctuated text versions of A08 and A09 (i.e. no commas or full stops etc) and without reference to

the original recordings. Intuitive boundary locations and types were then compared to corpus annotations (see table 3). An example of these intuitive predictions is given below and set alongside corpus annotations in a short extract from A08 where the phrasing is quite dense – more so in the intuitive version than the original. The intuitive phrasing version also arranges the text so that what are considered to be the most important boundaries, those giving rise to longer prosodic phrases, appear at the end of the line:

Intuitive phrasing:

Given the state of lawlessness that exists in Lebanon ||
the uninformed outsider | might reasonably expect | security | at Beirut airport |
to be amongst the tightest in the world ||
but the opposite is true ||

Corpus annotations:

Given the state of lawlessness that exists in Lebanon ||
the uninformed outsider might reasonably expect security | at Beirut airport ||
to be amongst the tightest in the world ||
but the opposite is true ||

4. Results

4.1 The chunk parse rule

The chunk parser's rule-based identification of prosodic phrases via retrieval of prepositional phrases, plus the author's intuitive predictions were compared to "gold standard" boundary annotations of extracts A08 and A09 in the Aix-MARSEC corpus by two expert linguists. An overview of how many boundaries of both types (major and minor) were correctly located by rule and by human judgement is presented in this section, while the discussion of error types – deletions (missed boundaries) and false insertions – plus overall performance of the chunk parser is reserved for the following section.

	GK A09 "gold standard"	Chunk Parse 1	Chunk Parse 2	Intuitive phrasing
Total number of boundaries (minor + major)	200	131	135	156
Total number of boundaries (minor + major) correct	-	81	87	139
Total number of major boundaries	31	-	-	52
Total number of major boundaries correctly located	-	9	18	31
Total number of minor boundaries	169	-	-	104
Total number of minor boundaries correctly located	-	72	69	83
Total number of full stops	24	-	-	-
Total number of full stops correctly located	-	7	15	23

	BW A08 "gold standard"	Chunk Parse 1	Chunk Parse 2	Intuitive phrasing
Total number of boundaries (minor + major)	120	not run	110	93
Total number of boundaries (minor + major) correct	-	-	56	85
Total number of major boundaries	67	-	-	60
Total number of major boundaries correctly located	-	-	33	45
Total number of minor boundaries	53	-	-	33
Total number of minor boundaries correctly located	-	-	23	12
Total number of full stops	33	-	-	33
Total number of full stops correctly located	-	-	-	32

Table 3: Raw counts of prosodic boundaries discovered via the chunk parse rule and by intuitive predictions as compared to corpus annotations in Aix-MARSEC extracts A08 and A09.

In evaluating the effectiveness of the chunk parse rule and the intuitive phrasing approach, 3 different measures have been used: total number of boundary positions correctly located; number of major and minor boundary types correctly located; and number of full stops correctly located. The first measure does not distinguish between major and minor boundaries; so as long as boundary site was correctly identified, an

exact match between position and boundary type was not looked for. Chunk parse 1 took as input text without full stops or commas etc (as did the author when making intuitive predictions) but this did not locate boundaries where constituents included in the rule spanned the boundary as in:

'…some form {of local government ‖ at a news conference}…the party leaders…'

This approach was therefore abandoned, with an overall success rate of 40.50% boundary positions correctly located in A09. For chunk parse 2, full stops only were restored and this gave marginally better performance: 43.50% boundary positions correct for A09 and 46.66% correct for A08. Obviously, detection could be improved with fuller punctuation but as already pointed out, punctuation is partly a matter of style and the idea behind this experiment was to create a catch-all rule, independent of text domain.

Syntactic contexts in which the chunk parse rule does seem to approach natural phrasing include consecutive prepositional phrases, for example:

'…{near the top of the political agenda of the major Western powers}…'

One could argue for a boundary after the word 'agenda'; equally, one could get by quite comfortably without it. The chink chunk rule would create a surplus of boundaries here – 3 in all. This example does raise one issue, however, about the status of the preposition 'of' which seems to have a weaker semantic identity than other prepositions and which is reliant on neighboring nouns. Here, the word 'of' marks degrees of proximity to a desired target: the TOP of a particular agenda. Its link-up role can be illustrated by a further example where a boundary is invoked at the point where 'of' re-establishes contact between target and tributary nouns in the pattern '…a picture of..:'

'…an x-ray picture | on two TV screens | of the contents of hand baggage…'

Corpus annotations indicate the boundary after 'screens' is stronger than the boundary after 'picture'.

4.2 Reflections on intuitive prosodic phrasing

Perhaps the most interesting result of this three-way comparison of predicted and perceived prosodic phrasing is within-sentence allocation of major boundaries by the author and by Knowles and Williams. Raw data from table 3 can be reworked as follows:

	% major boundaries not accounted for by full stops			
	GK	CB	BW	CB
A09	22.58%	53.85%	-	-
A08	-	-	50.75%	45%

Table 4: Percentage distribution of major intonational phrase boundaries *within* sentences by expert annotators GK (Gerry Knowles) and BW (Briony Williams), and also by author (CB).

The further point of interest is the performance of this rather crude chunk parse rule relative to human judgement. The former gets between 43 and 47 per cent of boundaries correct for A09 and A08 respectively, while the latter scores between 69 and 71 per cent. The rule-based method actually performs better than the author when discovering minor phrase boundaries in A08.

5. Discussion

The table in figure 4 summarizes error types thrown up by the chunk parsing experiments on extracts A08 and A09, where missed boundaries are classified as deletion errors and boundaries not in sync with corpus annotations are classified as insertion errors. A standard textbook on statistical natural language processing [21] discusses ambiguity caused by non-categorical behaviour of parts of speech: individual words can be POS-tagged differently in different syntactic contexts and, though allocated a particular POS tag in a particular context, may retain and exhibit simultaneous behaviours. Such ambiguity is evident from table 5 in that there are arguments for and against the inclusion of certain parts-of-speech within the chunk rule and because the class of prepositions is associated with a range of POS tags.

SYNTAX		EXAMPLE IN CONTEXT		ERROR TYPE	
POS TAG	CONSTRUCTION			DELETION ERRORS	INSERTION ERRORS
VBG	collapsed relative clause	1	|on top of a hill| **overlooking Windhoek}**	X	-
VBG	GERUND (-ing form as noun)	2	mistakes they had made |in their} **handling** | of the Algerian people|	-	X
VBG	PARTCIPLE heading verb phrase	3	left to fly back |to South Africa| **leaving** those internal leaders	no error here	
VBN	PAST PARTICPLE as noun premodifier	4	to make way |for an} **unchecked** SWAPO government |in Windhoek|	-	X
NN	consecutive noun phrases	5	given the state |of lawlessness| that exists |in **Lebanon} the uninformed outsider**| might reasonably expect	X	-
CC	conjunction needed within rule	6	recent operations |in Angola} **and** Botswana	X	-
CC	conjunction NOT wanted within rule	7	need their weapons |on board| **and** getting them through	no error here	
RP and CC	two examples of noise	8	|on aeroplanes| flying |**around** the Middle East} **and** the Mediterranean	-	X
RB and (RP or IN)	adverbial overlap and noisy tags	9	Pretoria's hold |**on** the mineral rich territory| replaced |by a} **possibly** Marxist government	-	X
RB	RB needed in rule	10	at Heathrow} **once**	-	X
RB	RB NOT wanted within rule	11	gathered together |under one roof| **hence** its name	no error here	

Table 5: Classification of error type in the chunk parsing experiment, where **pipes indicate boundaries correct** and **squigs indicate a deletion or insertion error**; errors are then attributed to particular words and POS tags.

The first 3 examples here involve words tagged as <VBG>, the verb form ending in 'ing'. Words tagged with this part of speech can function as verbs or as nouns but the tag itself does not make this distinction. Resolving the problem in example 2 would be a straightforward case of re-tagging the word 'handling' as a gerund or verbal noun [22] and including this tag in the rule. However, examples 1 and 3 could not be resolved so easily. In (1) we understand '…a hill which overlooks or which is overlooking…' a place; in (3) we understand that someone did 2 kinds of leaving: they left for home and left a group of people behind to sort things out – strangely, a present participle is being used to refer to a past event! Moreover, in (1) we want <VBG> in the rule, whereas in (3) we don't because here the tagged entity initiates a new chunk in the sentence and has nothing to do with the prepositional phrase.

Examples 1 to 3 demonstrate the notion of 'category blends' [21], words simultaneously functioning as 2 or more parts of speech – in this case, 'ing' forms blurring the distinction between nouns and verbs. Example (4) is another instance of this, where the past participle <VBN> is functioning as an adjective and as such should be included in the rule. Working through the list of errors presented, example (5) is evidence that the linearity of the chunk parse rule is both good and bad for prosody. It defines a chunk quite flexibly through an exclusive set of tags but is not able in its present form to differentiate between immediately adjacent chunks which present an unbroken sequence of POS tags belonging to the prepositional phrase set. Examples 6 to 8 again present the catch-22 situation of whether to include a tag in the rule or not. Since <CC> stands for a powerful set of words, whose very title of 'coordinating conjunctions' alerts us to their role as linking devices between chunks, this tag was banished from the rule.

The remaining examples (9 to 11) demonstrate a major problem for this rule which requires the tag <IN> (preposition) to initiate a chunk. It was reported in Section 3.1 that round about a fifth of tagging errors were caused by multiple tags associated with prepositions: <IN>, <RP>, <RB>, <CS>. Examples (8) and (9) highlight the difficulty of discriminating between prepositions and verb particles, while examples (10) and (11) present conflicting instances of adverbials inside and outside the rule. Though not reported in fig. 3, the initial POS tagging of A08 provided several instances of the prepositions 'before' and 'for' being tagged as subordinating conjunctions <CS>; this was inappropriate for the context in which they appeared.

6. Conclusions and further work

Prepositional phrases constitute a powerful linguistic grouping as sentence modifiers and this initial study confirms that there is a degree of correspondence between the edges of these syntactic units and prosodic phrase boundaries. The study also confirms the principle that prosodic phrases can be successfully identified via a shallow chunk parse. However, the chunk parse rule devised to isolate prepositional phrases here is still incomplete. It could be supported by a more discriminating tagset (different tags for present participles and gerunds, for example) but this would not resolve instances where the same tag, and thus same part of speech, appears legitimately inside and outside the rule. The fact that such a small sample of text poses conundrums of this kind is telling. Furthermore, prepositional phrases are not the only syntactic grouping which corresponds to prosodic phrases. Evidence here suggests that there is a useful distinction to be made for this rule-based method

between prepositions heading a phrase and prepositions occurring within noun phrases, particularly object noun phrases, and this is one area where the chunk parse rule will be developed. The comparison of intuitive prosodic phrasing to corpus annotations illustrates, first, that major prosodic boundaries (break index 4) are being used and perceived *within* sentences and not just in sentence-final position. What also emerges is the optional nature of minor boundaries and minor boundary positions, particularly when, in one extract, the crude chunk parse rule outperformed human judgement in securing a boundaries-correct result. Nevertheless, to discover whether certain minor boundary positions are more essential than others, it will be necessary to investigate accent-boundary combinations, a significant feature included in **[10]**, and to use the full range of prosodic annotations in the Aix-MARSEC Corpus to look at occurrences of minor boundaries marked by pitch accents versus minor boundaries preceded simply by tonic stress marks. The accent-boundary relationship will also be an essential feature to include in the study of within-sentence major boundary positions. In this case, pitch accent type prior to a major boundary will be important to see whether choice of accent is indeed indicating the end of a tune.

The goal of automatic phrase break prediction is to identify prosodic-syntactic boundaries in text which correspond to the way a native speaker might process or chunk that same text as speech. This is treated as a classification task in machine learning and output predictions from language models are evaluated against a 'gold standard': human-labelled prosodic phrase break annotations in transcriptions of recorded speech - the speech corpus. Despite the introduction of rigorous metrics such as precision and recall, the evaluation of phrase break models is still problematic because prosody is inherently variable; morphosyntactic analysis and prosodic annotations for a given text are not representative of the range of parsing and phrasing strategies available to, and exhibited by, native speakers. We conclude that we need to create automatically-generated POS tagged and prosodically annotated variants of a text to enrich the gold standard and enable more robust 'noise-tolerant' evaluation of language models.

The next stage in our research is to build a prosody lexicon, for prosodic markup of further corpora, and also for other NLP applications. We are developing a prototype prosody lexicon for English, derived from the computer-usable dictionary CUVPlus, in conjunction with CELEX-2 and the Carnegie-Mellon Pronouncing Dictionary. The lexicon is designed for the target application of prosodic phrase break prediction but is also relevant to other machine learning tasks. It currently incorporates new fields for Penn Treebank part-of-speech tags mapped to default closed and open-class word categories, plus information about the rhythmic structure of each word in the form of a lexical stress pattern, interpreted as a potential 'text-based' feature for automatic phrase break classifiers. The lexicon comes as a textfile, with entries in multi-field format; this lends itself to exploitation as a Python dictionary, with multiple values associated with each wordform 'key'. In its final form, the lexicon will be made freely available for corpus-based research in speech synthesis and speech recognition.

This research is another step towards a better understanding of the interaction between grammar and prosody **[23]**. Its practical application is in improving prosody in speech synthesis used in text-to-speech systems; this could make speech systems much more widely acceptable as a general computing and internet interface **[24]**. Prosody is also a challenge for learners of English as a foreign language **[25]**, so prosody analysis and prediction should be useful in advanced English language teaching **[26]**.

References

[1] Ladd, R. (1996) Intonational Phonology Cambridge, Cambridge University Press

[2] Pitrelli, J., Beckmann, M. and Hirschberg, J. (1994) ToBI (Tones and Break Indices) Proceedings of the 1994 International Conference on Spoken Language Processing, 18–22 September, Yokohama
http://www1.cs.columbia.edu/%7Ejulia/research.html

[3] Beckman, M.E., Ayers, G.M. (1997) Guidelines for ToBI Labelling, Department of Linguistics, Ohio State University
http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/ToBI/ToBI.1.html

[4] Auran, C., Bouzon, C. and Hirst, D. (2004) The Aix-MARSEC Project: An Evolutive Database of Spoken English Presented at Speech Prosody 2004, International Conference; Nara, Japan, March 23–26, 2004, ed. by Bernard Bel and Isabelle Marlien, ISCA Archive
http://www.isca-speech.org/archive/sp2004/sp04_561.html

[5] Taylor, L.J. and Knowles, G. (1988) Manual of Information to Accompany the SEC Corpus: The machine readable corpus of spoken English. University of Lancaster
http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM

[6] Roach, P., Knowles, G., Varadi, T. and Arnfield, S. (1993) "Marsec: A machine-readable spoken English corpus" Journal of the International Phonetic Association, vol. 23, no. 1, pp. 47–53

[7] Spoken English Corpus text A08, Speaker: Keith Graves Broadcast notes: BBC Radio 4, 11.30 a.m., 22nd June, 1985

[8] Winograd, T. (1984) Computer Software for Working with Language in Scientific American 251: 31–45

[9] Francis, W.N., and Kucera, H., (1979) Brown Corpus Manual (Revised and Amplified), Department of Linguistics, Brown University
http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM

[10] Koehn, P., Abney, S., Hirschberg, J., and Collins, M. (2000) Improving Intonational Phrasing with Syntactic Information In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol 3, pp. 1289-1290, Istanbul, June, 2000
http://citeseer.ist.psu.edu/koehn00improving.html

[11] Liberman, M.Y., and Church, K.W. (1992) Text Analysis and Word Pronunciation in Text-to-Speech Synthesis In Furui, S., and Sondhi, M.M., (eds) (1992) Advances in Speech Signal Processing New York, Marcel Dekker, Inc.

[12] Roach, P. in Arnfield, S. (1994) Prosody and Syntax in Corpus Based Analysis of Spoken English PhD Thesis, University of Leeds

[13] Bird, S. and Loper, E. (2004) NLTK: The Natural Language Toolkit In Proceedings 42nd. Meeting of the Association for Computational Linguistics (Demonstration Track) pp. 214–17, Barcelona, Spain
http://citeseer.ist.psu.edu/loper02nltk.html

[14] Bird, S. and Loper, E. (2006) nltk_lite v. 0.6.5
http://nltk.sourceforge.net/lite/doc/api/nltk_lite-module.html

[15] Bird, S., Curran, J., Klein, E., and Loper, E. (2006) Chunk Parsing
http://nltk.sourceforge.net/lite/doc/en/chunk.html

[16] Huckvale, M. (2002) Speech Synthesis, Speech Simulation and Speech Science, Proc. International Conference on Speech and Language Processing, Denver, 2002, pp. 1261–64
http://www.phon.ucl.ac.uk/home/mark/

[17] Spoken English Corpus text A09, Speaker: Graham Leach Broadcast notes: BBC Radio 4, 11.30 a.m., 22nd June, 1985

[18] Boersma, P. and Weenink, D. (2006): Praat: doing phonetics by computer (Version 4.4.26) [Computer program] http://www.praat.org/

[19] Bird, S., Curran, J., Klein, E., and Loper, E. (2006) Tagging
http://nltk.sourceforge.net/lite/doc/en/tag.html

[20] Paulin, T. (2003) Spirit of the Age In The Guardian, Saturday 5 April, 2003
http://books.guardian.co.uk/review/story/0,12084,929528,00.html

[21] Manning, C.D., and Schutze, H. (1999) Foundations of Statistical Natural Language Processing Cambridge, Massachusetts The Massachusetts Institute of Technology

[22] Gerund http://en.wikipedia.org/wiki/Gerund

[23] Arnfield, S. and Atwell, E. (1993) A syntax based grammar of stress sequences. In: Lucas, S (editor) Grammatical Inference: Theory, Applications and Alternatives, pp. 71–77, IEE Colloquium Proceedings no.1993/092.

[24] Atwell, E. (2005) Web chatbots: the next generation of speech systems? European CEO, November-December, pp. 142–44.

[25] Atwell, E., Howarth, P., and Souter, C. (2003) The ISLE corpus: Italian and German Spoken Learner's English. ICAME Journal, vol. 27, pp. 5–18.
http://icame.uib.no/ij27/index.html

[26] Oba, T. and Atwell, E. (2003) Using the HTK speech recogniser to anlayse prosody in a corpus of German spoken learner's English. In: Archer, D, Rayson, P, Wilson, A and McEnery, T (editors) Proceedings of CL2003: International Conference on Corpus Linguistics, pp. 591–98 Lancaster University
http://www.comp.leeds.ac.uk/eric/cl2003/ObaAtwell.doc