Prosodic Phrase Break Prediction: Problems in the Evaluation of Models against a Gold Standard.

Claire Brierley - Eric Atwell

School of Computing, University of Leeds, Woodhouse Lane
Leeds, LS2 9JT, U.K., {claireb, eric}@comp.leeds.ac.uk

ABSTRACT. *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. This article recommends creating 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.*

RESUME. *L'objectif de la prédiction automatique des frontières entre syntagmes est d'identifier dans le texte les frontières prosodiques et syntaxiques qui correspondent à la manière dont un locuteur natif traiterait ou découperait ce texte en parlant. Ceci correspond à une tâche de classement en apprentissage automatique et les prédictions produites à partir des modèles de langage sont évaluées à l'aide d'un corpus de référence, c'est-à-dire un corpus de parole transcrite annoté manuellement par les frontières prosodiques entre syntagmes. Malgré l'utilisation de mesures rigoureuses comme la précision et le rappel, l'évaluation des modèles de frontières entre syntagmes reste problématique car la prosodie est intrinsèquement variable : l'analyse morphosyntaxique et les annotations prosodiques d'un texte donné ne sont pas représentatives de l'ensemble des stratégies d'analyse et de découpage possibles utilisées par les locuteurs natifs. Cet article recommande de générer automatiquement des variantes d'étiquetage morphosyntaxique et d'annotation prosodique d'un texte pour enrichir le corpus de référence et permettre une évaluation des modèles de langage plus robuste et tolérante au bruit.*

KEY WORDS: *Evaluation, prosody, supervised learning, statistical methods*

MOTS-CLES : *évaluation, prosodie, apprentissage supervisé, méthodes statistiques*

TAL. Volume 48 – n° 1/2007, pages 187 to 206

1. Introduction

Prosodic phrasing is a universal characteristic of language (Ladd, 1996) and refers to the way speakers of any given language process speech as a series of chunks: meaningful, stand-alone clusters of words which have some relationship to syntactic phrase structure, the 'natural joints' in sentences (Abney, 1995). The correlation and discrepancy between prosody and syntax is a continuing debate in the literature; but there does appear to be consensus on the fact that prosodic phrasing is simpler, shallower and flatter than syntactic structure. Abney (1992) proposes the unifying concept of *performance structure*, the way in which prosody and syntax interact in practice.

Performance structure in English is realised and perceived as a partnership between pitch accents and pauses which draws attention to these natural joints or boundaries in the speech stream. In text, prominent boundaries are marked by punctuation and it is second nature for us to associate different intonation and different degrees of pause with the various punctuation marks when reading that text aloud. Thus language models designed to predict prosodic phrase breaks from input text - for Text-to-Speech Synthesis applications, for example - will often use punctuation as a primary cue.

The goal of automatic phrase break prediction is, therefore, to identify natural joints in *text* which correspond naturally and intelligibly (these are the important criteria) to the way a native speaker might process or chunk that same text as speech. Once these boundaries have been discovered, the intervening text can be 'animated' with prosody - that is, given a suitable intonation contour. The shape of that synthetic contour and its faithfulness to fundamental frequency patterns in natural language will depend to some extent on how well accordance and anomaly at the prosody-syntax interface is understood.

Prosodic phrasing and intonation exhibit a dual purpose in speech: a chunking function to identify meaningful - and syntactically coherent - clusters of words and a highlighting function to emphasise salient items within clusters. In English, chunking and highlighting are often conflated (Peppe, 2006): prominent words tend to complete a phrase group and so occupy pre-boundary position. The convergence and *non*-convergence of these functions has consequences for the evaluation of language models that try to simulate them. How can a model distinguish between them when the empirical data from which the model is derived makes no such apparent distinction?

The incentive behind this question will be discussed in sections 4 and 5 of this paper, which begins with an overview of: phrase break prediction models; their derivation from prosodically annotated speech corpora, the 'gold standard' used for training and testing said models; and the metrics commonly used to evaluate model

performance. The proposals for future work in section 6 include the idea of generating parallel tiers annotated with different prosodies for an existing corpus which will then serve as a new gold standard for the task of phrase break prediction.

2. Automatic phrase break prediction: overview of process

Techniques for automated prediction of prosodic phrase boundaries in text, typically for Text-to-Speech Synthesis (TTS) applications, can be deterministic or probabilistic. In either case, the problem of phrase break prediction is treated as a classification task and outputs from the model, as in other Natural Language Processing (NLP) applications such as part-of-speech (POS) tagging, are evaluated against a human-labelled 'gold standard' corpus (Jurafsky and Martin, 2000 p.308), also known as a 'reference dataset' in the speech research community. For prosody, this gold standard is a test set where original transcriptions of recorded speech in the speech corpus include prosodic annotations by experts. Annotation systems commonly used for phrase break prediction are ToBI - Tones and Break Indices (Beckman & Ayers, 1997) - where the break index tier distinguishes 5 levels of juncture between words on a scale of 0 - 4, and the British system exemplified in SEC - the Spoken English Corpus (Taylor and Knowles, 1988) - which identifies 3 levels: no boundary, minor phrase boundary, major intonational phrase (IP) boundary. Minor and major boundaries are assigned the pipe symbols: < | > and < || > respectively, and map to break indices 3 and 4 in ToBI. In Roach (2000), these same symbols denote *tone unit boundary* < | > and *pause* < || >.

2.1. *Rule-based methods*

A standard rule-based method commonly used in TTS is to employ some form of 'chink-chunk' algorithm which inserts a boundary after punctuation and whenever the input string matches the sequence: open-class or content word (chunk) immediately followed by closed-class or function word (chink), based on the principle that chinks initiate new prosodic phrases. Bell Labs speech synthesizer uses this kind of rule to identify low-level phrasal units or f-groups (Abney, 2006). Variants of this algorithm may seek to shuffle parts-of-speech (POS) between open and closed-class groupings; the chink-chunk algorithm proper (Liberman and Church, 1992) treats tensed verb forms as chinks and object pronouns as chunks for more natural phrasing.

A more recent alternative rule-based method is described by Atterer (2002) and Atterer and Klein (2002); their model builds a hierarchical prosodic structure via a two-step process which uses the CASS chunk parser (Abney, 1991) to identify ϕ-phrases (f-groups) and then 'bundles' these minor phrases into intonational phrases. The algorithm uses a variable threshold figure (default setting 13) to limit the

number of syllables in an intonational phrase if there is no intervening punctuation.

2.2. *Statistical methods*

The leading study in the use of statistical methods for phrase break prediction is Taylor and Black's Markov model (1998), trained and tested on MARSEC, the Machine Readable Spoken English Corpus (Roach *et al*, 1993) and used in Edinburgh's Festival speech synthesis system (Black *et al*, 1999; Black, 2000). The training data for this supervised learning model is 'text' represented by a sequence of POS tags which include boundary tags. The model is structured such that states represent types of break - the desired classification outputs of break or non-break - and transitions represent likelihoods of phrase break sequences occurring. The model thus 'learns' the classification task by integrating two sets of information: the probability of a POS sequence, given juncture type, and the probability of a particular sequence of juncture types occurring. This extensive study actually goes on to compare the performance of both probabilistic *and* deterministic language models over six experimental settings, with a best score of 79% breaks-correct achieved with a higher order n-gram model and a more streamlined tagset obtained by post-mapping the output of the POS-tagger onto a smaller tagset of 23.

Busser *et al* (2001) compare the effectiveness of a Memory-Based Learning (MBL) approach to predicting phrase breaks in MARSEC to Taylor and Black's ('gold standard') use of HMMs for the same purpose. MBL is a supervised-learning approach where classification of data is made on the basis of maximum similarity to items in memory. In this study, the set of feature values descriptive of phrase break contexts, and used as input to train the classifier is: the orthographic form of the word in question; its POS tag; its CFP-value (status as content word, function word or punctuation mark); and an expanded tag which gives the word itself if it is a function word and the POS tag otherwise. A fixed-width feature vector of two words both to the left and right of the focus position in question supplies the context from which to extrapolate the 'minority' class 1 (break) or more frequent class 0 (non-break i.e. ordinary juncture). The study involves converting Taylor and Black's results over six experiments to the MBL metrics of precision, recall and F-score (see the discussion on performance measures in section 3) for the purposes of comparison and then experimenting with further optimization of these metrics, creating a different mix of information in the feature vectors via leave-one-out experiments and cross-validating against the training set. Busser *et al* report an improvement on the best HMM result for recall with a simple MBL algorithm which takes a limited context of one POS to the left and right of the focus position and assigns equal weighting to each of these positions.

Taylor and Black's use of a reduced tagset in their framework for assigning phrase breaks from POS information has been taken forward in a recent study by Read and Cox (2004). This presents a best first search algorithm (suitable for any tagset) for exploring and determining groupings of POS tags that will eventually

constitute a reduced, optimal tagset for phrase break prediction. Read and Cox use what they term a *flattened* prosodic phrase hierarchy classification of break/non-break on datasets from the Boston Radio News Corpus (Ostendorf *et al*, 1995) and MARSEC, and to evaluate their phrase break prediction model, use Taylor and Black's *junctures correct* measure, that is the percentage of non-breaks correctly predicted.

The statistical modelling technique known as CART (a Classification and Regression Tree) is used by Wang and Hirschberg (1991) to predict prosodic phrase boundaries from features that can be automatically generated from text. 'Learning' for this decision tree method includes training the splitting rules at each decision point in the tree to select the feature/value split which minimises prediction error rate in the training set. In this study, such features include: length of utterance in seconds and words; position of potential boundary site and distance from beginning and end of utterance; and syntactic constituents adjacent to the boundary site. An important additional feature used to compare the performance of the original model to an enhanced model which incorporates hand-labelled transcriptions in the data set (298 sentences of air travel information from DARPA, 1990) is accent status of $<w_i>$, where $<w_i , w_j>$ represents words either side of the boundary site. The best performing variable set included information from prosodic annotations of pitch accent and prior boundary location, giving a success rate of 90% boundaries correct and a streamlined tree with only 5 decision points.

A related and more recent study (Koehn *et al*., 2000) builds on an augmented version of the above feature set (Hirschberg and Prieto, 1996) by adding syntactic information from a high accuracy syntactic parser. The '1996' feature set consists of the following: a 4-word POS window and a 2-word accent window; the total number of words and syllables in the utterance; word distance from start and finish of the utterance in words, syllables and stressed syllables; distance from last punctuation mark and what punctuation, if any, follows the word; position of word in relation to, or within, a noun phrase; and finally, size and distance of word from start of noun phrase. The '2000' feature set builds on the intuition that prosodic phrase breaks occur between large syntactic units {NP, VB, PP, ADJP, ADVP} and incorporates binary flags indicating which words initiate a major phrase or a sub-clause. The study reports a 90.8% prediction rate for boundary detection which is cross-validated using other machine learning algorithms: a boosting algorithm, a rule learner, a boosted decision tree classifier and an alternating decision tree method.

3. Evaluation metrics used in studies

The previous section briefly discusses a range of machine learning methods applied in prosodic phrase break prediction. The evaluation metrics used in studies seem to fall into one of two groups, however. The first group (see Wang and Hirschberg, 1991; Atterer, 2002; Read and Cox, 2004;) select from the set of

accuracy and *error* measures discussed in Taylor and Black (1998) and presented in Tables 1 and 2 below.

% breaks-correct (*true positives*)	breaks correctly predicted / total number of breaks in test set	x 100
% non-breaks-correct (*true negatives*)	non-breaks correctly predicted / total number of non-breaks in test set	
% junctures-correct	(breaks + non-breaks) correctly predicted / total number of junctures in test set	

Table 1. *Accuracy measures for phrase break prediction*

% insertion errors (1) (*false positives*)	breaks retrieved by model / total number of breaks in test set	x 100
% insertion errors (2)	breaks retrieved by model / total number of junctures in test set	
% deletion errors (*false negatives*)	breaks missed by model / total number of breaks in test set	

Table 2. *Error measures for phrase break prediction*

Taylor and Black argue that *breaks-correct* is a better measure of algorithmic performance than *junctures-correct* because the latter includes *non-breaks* in the calculation and these are always more numerous.

The second group of evaluation metrics employed in statistical NLP (and for phrase break prediction see the aforementioned: Koehn *et al*, 2000; Busser *et al*, 2001; Atterer and Klein, 2002) are taken from the field of Information Retrieval and are known as *precision* and *recall*. The latter corresponds exactly to the *breaks-correct* measure, while the former equates to *positive predictive value*: in this case, the proportion of correct (relevant) predictions out of all the predictions made. In practice it is usual to combine precision and recall into a single overall performance measure or F-score which tends to maximise *true positives* (Manning and Schütze, 1999) - in this case *breaks-correct*. Table 3 shows how *precision*, *recall* and *F-score* are interpreted for the task of phrase break prediction.

Precision	breaks correctly predicted / number of breaks retrieved	x 100
Recall	breaks correctly predicted / total number of breaks in test set	
F-score	2 * precision * recall / precision + recall	

Table 3. *Information Retrieval measures used in phrase break prediction*

Phrase break prediction models are evaluated in terms of their ability to match boundary annotations in the test corpus. However, the long-term view is that the model will be able to generate intelligible and natural prosodic phrasing for *any* input text. It is hoped the model will have learnt the classification task well enough to make generalisations from the gold standard to the new domain. If it hasn't, it runs the risk of imposing a prosody template (one speaker, one realisation, one moment in time) on unsuspecting text. Some models over-predict; but how many of their false insertions or false positives are nevertheless valid in terms of performance structure? How many missed boundaries or false negatives in a given model are significant omissions?

4. A gold standard for prosodic phrasing

Two publications discussed in this paper raise questions about the practice of evaluating a prosodic phrase break model against a gold standard; in both cases the iconic prosodic annotations in versions of the Spoken English Corpus. Taylor and Black (1998) state that performance figures obtained in such experiments should be '…treated with caution…' because prosody itself is subjective: different speakers pause in different places; one speaker will vary their use of pauses; expert annotators differ in their perceptions. Similar comments about variability in human performance appear in Hirschberg (2002). Taylor and Black also point out that junctures differ in type: those junctures which coincide with weaker *syntactic* boundaries are more likely to be potential *prosodic* boundary sites (see also Abney, 1992; Abney, 1995). Atterer and Klein (2002) encapsulate all these reservations: '…the very notion of evaluating a phrase-break model against a gold standard is problematic as long as the gold standard only represents one out of the space of all acceptable phrasings…'

4.1. Inter-annotator agreement

The 'spaciousness' of acceptable prosody can be demonstrated straightaway by the gold standard itself in Figures 1 and 2, a sample from Section C in Aix-MARSEC (Auran *et al*, 2004), an augmented version of the Spoken English Corpus with multi-level annotation tiers covering a range of segmental and suprasegmental features. The extract comes from a Reith Lecture and is illustrative because, while there is only one speaker, there are two alternative phrasings: this is one of the overlapping sections of prosodic annotation from Briony Williams and Gerry Knowles (approximately 9% of the corpus).

The main difference between Knowles' and Williams' boundary annotations here seems to be one of perception. In the section marked in **bold**, Gerry Knowles 'hears' a more emphatic speaker than Briony Williams and inserts more pauses

overall (35 instead of 29). Both annotators insert a boundary at every punctuation mark in the original raw text transcript - another acceptable phrasing, perhaps?

for some people | this statement of orthodox economic doctrine | may appear | too unqualified || since it fails to mention explicitly | security of supply || often | though not always | the case for self sufficiency is argued | with reference to a country's need to ensure security | by minimising dependence | on foreign sources || the outside world is seen | at best | as unreliable | and subject to instability | at worst | as actively hostile || **from this fortress mentality | standpoint | autarchy | appears | to be common prudence ||** two sets of measures | then suggest themselves | one is to build up | domestic production of essentials | so as to reduce imports | to a minimum | the other | is to restrict exports | so as to ensure | that domestic supplies | are available | for domestic use ||

Figure 1. *This is a sample of Gerry Knowles' phrase break annotations for a BBC recording of a Reith Lecture from the 1980s.*

for some people | this statement of orthodox economic doctrine | may appear too unqualified || since it fails to mention explicitly | security of supply || often | though not always | the case for self sufficiency is argued | with reference to a country's need to ensure security | by minimising dependence on foreign sources || the outside world is seen at | best | as unreliable | and subject to instability | at worst | as actively hostile || **from this fortress mentality standpoint | autarchy appears to be common prudence ||** two sets of measures | then suggest themselves | one | is to build up domestic production of essentials | so as to reduce imports | to a minimum || the other | is to restrict exports | so as to ensure | that domestic supplies | are available for domestic use ||

Figure 2. *Briony Williams' phrase break annotations for the same sample as Fig. 1.*

Figure 3 shows both annotators largely in agreement on phrasing and on emphatic, bi-tonal accents (rise-falls) in a snapshot sentence from Figs. 1 and 2. The only area of dispute is whether or not to include a boundary after the word '…dependence…'

However, what if a new speaker took this same text and chunked it differently with the explicit intention of prioritising certain syntactic structures or constituents? What about the new phrasing in Fig. 4, for example, which differs from the original by deliberately highlighting intentions, movements, actions present in verb forms?

,often | though not **`/always |** the case for self sufficiency is **`/argued |** with reference to a country's need to ensure **se`/curity |** by minimising dependence | on foreign sources

Figure 3. *This is the corpus version, showing all prosodic phrase breaks noted by Knowles and Williams and the pitch accent annotations on words preceding boundaries where the experts are in agreement.*

often | though not always | the case for self sufficiency | is **argued |** with reference to a country's need | *to ensure security* | by **minimising |** dependence on foreign sources

Figure 4. *This alternative phrasing is largely achieved within the performance structure of the original - see discussion in section 4.2.*

4.2. The space of acceptable phrasings

The emphatic combination of (high) chunking accent and boundary in the matching annotations in **bold** in Fig. 3 is typical of English. An emphasis-boundary pattern has now been engineered in Fig. 4 for the participle **'…minimising |…'** (which gets a high level pitch accent from both annotators) and could enhance the infinitive construction '…*to ensure security*…' if a boundary were to be placed before the noun.

new instance:	[NP a country's need] | [VP to ensure | security]
new instance:	[VP to ensure | security] [PP by minimising dependence]
new instance:	[PP by minimising | dependence] [PP on foreign sources]

Figure 5. *Prosodic boundaries are shown in relation to the large syntactic units {NP, VB, PP, ADJP, ADVP} featured in Koehn et al, (2000).*

The difference between these new instances and the original template in Fig. 3 is that most of them occur *within* and not *between* discrete syntactic groupings – lower down the tree as it were. This is illustrated in Fig. 5. It is also worth noting that the only disruption these 'false insertions' make to the original phrasing surrounds the noun '…dependence…' where Knowles and Williams are not in agreement anyway (cf. Fig. 1 and Fig. 2).The new instances in Fig. 5 are not disfluencies (speaker hesitations); in fact, they evidence a coherent strategy on the part of the speaker to

emphasise ‘doing’. Furthermore, even though they would be classed as *false insertions* when compared to the corpus gold standard, they are definitely not wrong.

4.3. Different boundary types

Of course, that is not the end of the story. A new complication now arises in that these different types of boundaries - the *chunkers* higher up the syntax tree and the *highlighters* lower down the tree - are not differentiated in the corpus. It would be nice if they were analogous to major and minor boundary classifications and the symbols: < || > and < | >. This is not the case, however. Figures 6 and 7 show the same annotator (in this case Briony Williams) using different phrase break annotations to flag up major clause boundaries in a news bulletin and a lecture. The association of double pipes (ToBI’s break index 4) with major syntactic groupings, plus the use of pitch accent annotations *without boundary reinforcement* for highlighting in the first extract, seems much clearer.

there are ~two \,scanning machines || which give an `/X ray picture | on two tele`/vision *screens || of the _contents of `hand *baggage || when `/I've been through *Athens airport || and `that's about *two dozen `times in the past *two `/years || there's `/never been more than ~one se\,curity man on *duty || and ~he's \frequently reading a `newspaper || or ~chatting with _other `airport *staff ||

Figure 6. *This is a sample of Briony Williams’ annotations of informal news commentary from a BBC radio broadcast from the 1980s. It shows correspondence between major intonation unit boundaries and major clause boundaries.*

the `/history | of ~British nuclear ,power programmes | ~over the past thirty ,years | >pro_vides a de~pressing e\xample | of ~unreflecting _centralism in `action || `stoutly rein`forced | _I may /add | by `/other forms | of _DIY`E || `one aspect of this ,centralism | is the **i`/dea | which** has been em~braced by su*ccessive British _governments of `/both **parties | that** a ,choice | `has to be made | at `/Cabinet level | of ~one par,ticular re`/actor system | for _future nuclear `power stations | in \Britain ||

Figure 7. *Another sample annotation from Briony Williams shows minor intonation unit boundaries being used to demarcate major clause boundaries.*

5. Developing a syntax-driven model

The chink chunk rule inserts a prosodic phrase boundary after a punctuation mark and between a content word and a function word. The authors have been

experimenting with syntax-driven, rule-based models using **nltk_lite's** regular expression chunk parser (Bird and Loper, 2006). The first model uses the discrete syntactic grouping of prepositional phrases to locate prosodic boundaries (see Brierley and Atwell, 2007 for a full account of experimental work here) and the current prototype recognises potential boundary sites via POS tag oppositions (in effect unweighted bigrams) observed from empirical data in Aix-MARSEC. Sample predictions from this prototype rule are illustrated in Figs. 8 and 9 for two reasons: to demonstrate some of the problems encountered when interpreting and evaluating outputs from prosodic phrase break models (even *developmental* models) against a corpus gold standard; and for interested readers familiar with the Natural Language Toolkit (NLTK) and the *chunking* tutorial in particular (Bird *et al*, 2007). The outputs themselves are similar to those of other CFP algorithms in that they capture low level phrasal units; but the rule is also able to match corpus phrasing which discriminates *between* function words (see **bold** items in Fig. 8), one objective of model design being to explore the conventional mapping of function words to chinks.

… cast/VBN their/POSS spell/NN | not/XNOT only/RB | on/IN our/POSS eminent/JJ professional/JJ colleague/NN Dr/NPT FitzGerald/NP | but/CC **also/RB | on/IN** Mr/'NPT Howell/NP | who/WP **himself/PPL | has/HVZ** a/AT First/OD Class/NNP degree/NN | in/IN Economics/NNP…

Figure 8. *A representation of phrase break predictions from a syntax-driven rule which finds boundaries in the corpus that occur between function words.*

Raw predictions in Fig. 9 below show the rule working quite well on a sentence from the Reith Lecture transcript; the annotator here is Briony Williams. True positives (boundaries correct) are marked ☑ and false positives (false insertions) marked ☒. Commas were deliberately stripped from input text and comma sites retrieved by the rule at major chunking boundaries are therefore given in bold.

NLTK's chunking tutorial referred to above recommends '…several rounds…' of rule development and testing in order to create a good chunker. The data in Figs. 8 and 9 was obtained by running a simple chunking algorithm on part of the Reith Lecture transcript annotated by Briony Williams (1463 tokens); manually examining outputs and refining the rule; and running the revised rule on the same section and finally on a previously unexamined section - i.e. the remainder of the Reith Lecture transcript annotated by Gerry Knowles (2445 tokens). Scores were recorded as shown in Table 4.

('one', 'CD')
(INSERT_BOUNDARY: ('aspect', 'NN') ('of', 'IN')) ☒
('this', 'DT')
(INSERT_BOUNDARY: ('centralism', 'NN') ('is', 'BEZ')) ☑

('the', 'ATI')	
(INSERT_BOUNDARY: **('idea', 'NN') ('which', 'WP'))**	☑
('has', 'HVZ')	
('been', 'BEN')	
(INSERT_BOUNDARY: ('embraced', 'VBN') ('by', 'IN'))	☒
('successive', 'JJ')	
('British', 'JNP')	
(INSERT_BOUNDARY: ('governments', 'NNS') ('of', 'IN'))	☒
('both', 'ABX')	
(INSERT_BOUNDARY: **('parties', 'NNS') ('that', 'CS'))**	☑
('a', 'AT')	
(INSERT_BOUNDARY: ('choice', 'NN') ('has', 'HVZ'))	☑
('to', 'TO')	
('be', 'BE')	
(INSERT_BOUNDARY: ('made', 'VBN') ('at', 'IN'))	☑
('Cabinet', 'NP')	
(INSERT_BOUNDARY: ('level', 'NN') ('of', 'IN'))	☑
('one', 'CD1')	
('particular', 'JJ')	
('reactor', 'NN')	
(INSERT_BOUNDARY: ('system', 'NN') ('for', 'IN'))	☑
('future', 'NN')	
('nuclear', 'JJ')	
('power', 'NN')	
(INSERT_BOUNDARY: ('stations', 'NNS') ('in', 'IN'))	☑
('Britain', 'NP')	
('.', '.')	

Figure 9. *Phrase break predictions from a rudimentary syntax-driven rule retrieve sites of commas at major clause boundaries.*

	Annotator	**Precision**	**Recall**	**F-score**
Test 1	BW	65.19%	59.03%	61.96%
Test 2	BW	66.76%	71.35%	68.98%
Test 3	GK	70.85%	61.04%	65.58%

Table 4. *Sample P, R and F-scores from development tests on chunk parse phrase break rule based on unweighted bigrams*

While it is encouraging to measure the performance of this prototype phrase break rule in terms of P, R and F-score, the authors believe that such early results are

not so enlightening as close scrutiny of, and reflection on, raw outputs from the model as represented in Figs. 8 and 9 and discussed in sections 5.1 to 5.3 below. Many experimental accounts do not cover such details; and yet might it not be the case that these details give us insights into the nature of prosodic variability so that we can design better models?

5.1. Chunking versus highlighting

The examples in section 4 demonstrate how denser prosodic phrasing (highlighting) can be inserted into the existing chunk structure of a sentence. The rest of section 5 covers instances where prosody redistributes prominence, first by ignoring, and second by shifting chunk boundaries.

True positives in Fig. 9 above evidence a reliable rule-of-thumb when a major clause boundary and comma-site is retrieved before a subordinating conjunction:

'…the idea | which has been embraced | by successive British governments | of both ***parties*** **|** ***that*** a choice | has to be made…'

This is generally a POS context where prosody, performance structure and syntax (Abney, 1992) are in agreement: a prosodic boundary generally occurs with a major clause boundary. Nevertheless, the mismatch between prediction and empirical evidence in Fig. 10 below shows the speaker making a different chunking choice for this POS context - glossing over a major syntactic boundary and favouring the highlighting over the chunking function of prosody by placing adjectives **'…important…self-sufficient…'** in phrase-final position. Consequently, predictions-by-rule quickly get out of sync with empirical phrasing (though not out of sync with naturalness) because they each start to take a different processing route through the sentence. As a final twist, however, predicted phrasing manages to regain contact with the original after coverage of the *theme* (everything before the copula) is complete (see **bold** items in Fig. 10).

Corpus phrasing:

'…The idea that it's **important |** for developing countries to become **self-sufficient** | in food | **is** widely | and uncritically accepted | not just in Brussels; | but from the orthodox economic standpoint | it's without foundation…'

Predicted phrasing:

'…The **idea |** that it's important for developing **countries |** to become self-sufficient in food | **is** widely | and uncritically accepted | not just in Brussels | but from the orthodox economic standpoint | it's without foundation…'

Figure 10. *Predicted phrasing matches the corpus once the theme (everything before the copula* ***'…is…')*** *is established.*

5.2. Prepositions versus verb particles

The prototype rule inserts a boundary before true prepositions, POS-tagged <IN>. This accounts for false inserts - but legitimate, if somewhat emphatic ('Tony Blair style') prosodic phrasing - in the following sentence fragment in Fig. 11.

Corpus phrasing:

'…the idea | which has been embraced by successive British governments of both parties | that a choice | has to be made…'

Predicted phrasing:

'…the idea | which has been ***embraced*** **|** ***by*** successive British governments | of both parties | that a choice | has to be made…'

Figure 11. *Predicted phrasing abides by the gold standard POS tagged version of this sentence which classifies the function word '…by…' as a preposition.*

It will be noted from Fig. 9 that there are four empirically verified (true positive) phrase boundaries before prepositions in the section as a whole. Moreover, since the POS-tagged version of this text is itself a gold standard, and since this version classifies '…embraced by…' as <VBN><IN> (a past participle followed by a preposition), we have a situation where two equally valid gold standards - tagged text versus prosodic annotation - are in conflict. This arises because the same speaker in this particular instance has realised '…embraced by…' as one unit and, *via prosody*, has in effect tagged the preposition as a verb particle: <VBN><RP>. This rules out an intervening *chunking* prosodic phrase boundary and significant *chunking* accent on '…embraced…' Corpus annotation on the verb testifies to this: **em~braced** is a level accent.

5.3. ***A conflict of standards?***

Abney (1991) raises the thorny issue of prepositional phrase attachment, '…the most explosive source of ambiguity in parsing…' The POS identity of '…embraced by…' (see section 5.2 and Fig 12 below) is a case in point: is it <VBN><RP> or is it <VBN><IN>? If the function word ***by*** is tagged <RP>, it falls within the subcategorisation frame of the verb and is classed as an *argument*; whereas if it is tagged <IN>, its attachment is to the ensuing noun ('…**by** successive British **governments**…') and its behaviour is that of an adjunct - see Merlo *et al* (2006) for recent discussion of argument/adjunct distinction for prepositions.

(1) [NP the idea] | [VP which has been embraced by] [NP successive governments]

(2) [NP the idea] | [VP which has been embraced] | [PP by successive governments]

Figure 12. *Alternative 'chunk' parsing strategies for sentence fragment.*

The 'blended category' POS status (Manning and Schutze, 1999) of ***by*** in this instance is an opportunistic moment for the speaker to run with one of two different prosodies and two different parsing strategies as shown in Fig.12. Strategy (1) is the corpus version and strategy (2) is the version created by the POS-tagger. Since both versions are *inherent in the plain text* and both are equally valid, then perhaps such 'conflicts' can be resolved by generating POS tagged and prosodically annotated *variants* for a given text? These parallel prosodic-syntactic realisations will then enrich the gold standard and enable more robust, i.e. 'noise-tolerant', evaluation of language models and contribute to our understanding of linguistic phenomena, the goal of 'speech science' as defined by Huckvale (2002). Moreover, the idea of including variant annotations in a gold standard has been proposed and/or adopted in other areas of computational linguistics. It is well-established that two or more linguists may disagree on the analysis/annotation of a given sample of data (Shriberg and Lof 1991, Carletta 1996, Bayerl and Paul 2007); and sometimes both analyses can be legitimate. The MorphoChallenge2005 gold standard for evaluation of morphological analysis programs entered for the contest (Kurimo et al 2006) included occasional variant morphological segmentations; for example: **pitchers** can legitimately be analysed as **pitch er s,** OR **pitcher s**. Part-of-Speech taggers are normally expected to predict a single unambiguous PoS-tag for each word, but the gold standard Penn Treebank does allow for rare occasions when the Part of Speech is genuinely ambiguous (Santorini 1990, Marcus et al 1994, Atwell 2007); for example: **The duchess was entertaining last night**, the word **entertaining** is tagged **JJ|VBG** - Adjective OR Present Participle Verb. Similarly, a Multitreebank or collection of variant syntactic analyses of sentences can be used for comparative evaluation of rival parsing programs (Atwell, 1996), corpus linguists' parsing schemes (Atwell *et al*, 2000), and unsupervised machine learning Grammatical Inference systems (van Zaanen et al 2004).

6. Conclusions and further work

The utility of a phrase break prediction model, like any other language model in computational linguistics, is evaluated '…against a quantifiable measure of success at some task…' (Abney, 2002). For prosodic phrasing, the task is to recapture original boundaries stripped from the corpus test set; and seminal papers discussed in section 2 show this is often achieved by training the model on POS contexts in which boundaries are likely to occur. Outputs from the model - its predictions - are compared against a gold standard, in effect, against *human* performance,

encapsulated in the prosodic annotations in the corpus. The quantifiable measures of success are then expressed as: boundaries-correct, recall, precision and f-score.

It must be remembered, however, that what we really need from a phrase break prediction model is the ability to distinguish different *types* of boundary. One set of boundary types has something to do with length: there are major intonational phrases which are then made up of lower-level units, variously termed f-groups and ϕ-phrases. Another set of boundary types may be described via the simplifying concepts of *chunking* and *highlighting*. Chunking boundaries may be said to correspond to syntactic joints and may be manifest at different levels in the syntax tree. It might be hypothesised that highlighting boundaries occur much lower down the tree and that they can even split up the individual constituents of low-level syntactic units. A further characteristic of English is that it likes to chunk and highlight at the same time.

Boundary annotations in the corpus gold standard (the ultimate performance measure) do not consistently make these distinctions between boundary types. The few samples discussed in this paper have already shown that the annotation symbol for minor intonation phrase units (< | >) is used variously at major syntactic clause boundaries, f-group boundaries and within f-groups. A machine learner which uses the accent-boundary association as part of its context may therefore give better performance. The co-occurrence of falling accent and new clause initiated by a coordinating conjunction, for example, is a useful pattern signifying a major intonational phrase boundary in English. Performance might also be enhanced by incorporating features from parallel annotation tiers such as those provided by the Aix-MARSEC Project.

The wider problem when comparing predicted versus annotated boundaries is that the corpus gold standard embodies all the inconsistencies of human performance. One such inconsistency is that no two prosodies are alike and different speakers will employ different chunking and highlighting strategies for the same string. An individual speaker will also vary their treatment of the same POS context. Despite the fact that speech corpora contain recordings of utterances from a range of speakers, each annotated transcript only represents one in the space of possible performance strategies for a given text. Phrasing models will miss some boundaries simply because they were not designed to recognise that type of boundary in the first place. They will also predict alternative phrasing to that of the corpus, resulting in so-called false insertions. Again, examination of the gold standard in section 4.2 of this paper shows that boundaries can be inserted without disrupting the speaker's original processing route through a sentence.

To moderate the process of evaluating a language model against one prosodic template, and to better explore prosody itself, it would be useful to create a series of templates for use as training, development and test sets. These could be manually produced by expert annotators; however, such variants could also be generated automatically by training the phrase break model on a completely *different* corpus

and then running it on the *target* corpus. Alternative predictions emerging from this process in the form of extra or erased boundaries would then need to be verified for naturalness and intelligibility by human subjects.

Finally, sections 4 and 5 of this paper suggest that alternative prosodies might be explored and generated by incorporating the following in the design of models: differentiation between chunking and highlighting boundaries; differentiation between major and minor chunking boundaries; strategic modelling of different prosodies by targeting certain parts of speech - a *verbs* version, perhaps, as well as the more usual bias towards nouns (see section 4.1 in particular). Design-orientation of such phrase break models would then need to be factored into the per cent correct score. Overall, automatically-generated phrasing variants would be a natural extension to a multi-tiered corpus such as Aix-MARSEC; and passages like the one discussed in Fig. 6, which support these kinds of distinctions, would seem a good place to start. There are precedents for allowing alternative legitimate analyses in a gold standard corpus in other levels of linguistic research, for example in morphosyntactic analysis (Kurimo et al 2006), Part-of-Speech tagging (Santorini 1990, Atwell 2007), parsing (Atwell 1996), Grammatical Inference (van Zaanen et al 2004). So, we should also allow for encoding of genuine ambiguity in a gold standard corpus for evaluation of prosodic phrase break prediction models.

7. References

Abney S., "Parsing by Chunks." In: *Principle-Based Parsing: Computation and Psycholinguistics* Berwick R.C., Abney S., Tenny, C., (eds.), Kluwer Academic Publishers, Dordrecht (1991)

Abney S., "Prosodic Structure, Performance Structure and Phrase Structure." In: *Proceedings, Speech and Natural Language Workshop,* pp.425-428. Morgan Kaufmann Publishers, San Mateo, CA. (1992)

Abney S., "Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax." In: *Computational Linguistics and the Foundations of Linguistic Theory.* CSLI. (1995)

Abney S., "Statistical Methods." *Encyclopedia of Cognitive Science.*" Nature Publishing Group, Macmillian. (2002)
[Accessed December, 2006: http://www.vinartus.net/spa/publications.html]

Abney S., "Introduction to Computational Linguistics: Chunk Parsing." PowerPoint presentation (2006)
[Accessed December 2006: www.cs.um.edu.mt/~mros/csa2050/**ppt**/chunking.**ppt**]

Atterer M., "Assigning Prosodic Structure for Speech Synthesis: a Rule-Based Approach." In *Proceedings of the First International Conference in Speech Prosody, SP2002* Aix-en-Provence, France (2002)

Atterer M., Klein E., "Integrating Linguistic and Performance-Based Constraints for Assigning Phrase Breaks." In *Proceedings of Coling 2002* pp.29-35, (2002)

Atwell, E. "Comparative evaluation of grammatical annotation models" In Sutcliffe, R, Koch, H & McElligott, A (editors) *Industrial Parsing of Software Manuals*, pp. 25-46 Rodopi. (1996)

Atwell, E., Demetriou, G., Hughes, J., Schriffin, A., Souter, C., Wilcock, S., "A comparative evaluation of modern English corpus grammatical annotation schemes". *ICAME Journal*, vol. 24, pp. 7-23. (2000)

Atwell, E. "Development of tag sets for part-of-speech tagging" In: Lüdeling, A., Kytö, M., (editors) *Corpus Linguistics: An International Handbook* Mouton de Gruyter (2007)

Auran C., Bouzon C., Hirst D., "The Aix-MARSEC Project: An Evolutive Database of Spoken English." Presented at *Speech Prosody 2004, International Conference*; Nara, Japan, March 23-26, 2004, Bel B., Marlien I., (eds) ISCA Archive (2004)

Bayerl P., Paul, K., "Identifying sources of disagreement: Generalizability theory in manual annotation studies". *Computational Linguistics*, Vol. 33 No. 1 pp.3-8 (2007)

Beckman M.E., Ayers G.M., "Guidelines for ToBI Labelling." Department of Linguistics, Ohio State University. (1997) [Accessed September, 2006: http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/ToBI/ToBI.1.html]

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

Bird S., Curran J., Klein E., Loper E., Chunking (2007) [Accessed March, 2007: http://nltk.sourceforge.net/lite/doc/en/chunk.html]

Black A.W., Taylor, P., Caley R., "The Festival Speech Synthesis System: System Documentation Festival" version 1.4 (1999) [Accessed December, 2006: http://www.cstr.ed.ac.uk/projects/festival/manual/festival_17.html]

Black A.W., Taylor, P., Caley R., "Speech Synthesis in Festival: A Practical Course in Making Computers Talk" Festival version 2.0 (2000) [Accessed December, 2006: http://festvox.org/festtut/notes/festtut_toc.html

Brierley C., Atwell E., "Using nltk_lite's chunk parser to detect prosodic phrase boundaries in the Aix-MARSEC corpus of spoken English" Research Report 2007.02, School of Computing, University of Leeds (2007)

Busser B., Daelemans W., van den Bosch A., "Predicting phrase breaks with memory-based learning." 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Edinburgh (2001)

Carletta, J., "Assessing agreement on classification tasks: the kappa statistic." *Computational Linguistics* Vol. 22, No. 2, pp.249-254 (1996)

Hirschberg J., Prieto P., "Training intonational phrasing rules automatically for English and Spanish text-to-speech." In *Speech Communication*, Volume 18, Number 3, May 1996, pp. 281-290(10) (1996)

Hirschberg J., "Communication and Prosody: The Functional Aspects of Prosody." In *Speech Communication* Volume 36, Number 1, January 2002, pp. 31-43(13) Elsevier Science (2002)

Huckvale, M., "Speech Synthesis, Speech Simulation and Speech Science", In *Proc. International Conference on Speech and Language Processing*, Denver, pp1261-1264 (2002)

Koehn P., Abney S., Hirschberg J., Collins, M., "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 (2000)

Kurimo M., Creutz M., Varjokallio M., Arisoy E., Saraclar M., "Unsupervised segmentation of words into morphemes - Challenge 2005: An Introduction and Evaluation Report" In: Kurimo, M, Creutz, M & Lagus, K (editors) *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*. (2006)

Ladd, R. *Intonational Phonology* Cambridge, Cambridge University Press. (1996)

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

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

Marcus M.P., Santorini B., Marcinkiewicz M.A., "Building a Large Annotated Corpus of English: The Penn Treebank" *Computational Linguistics*, Vol. 19, No.2, Pages 313-330 (1994)

Merlo P., Ferrer E.E., "The Notion of Argument in Prepositional Phrase Attachment" *Computational Linguistics*, September 2006, Vol. 32, No. 3, Pages 341-378 (2006)

Ostendorf M., Price P.J., Shattuck-Hufnagel S., "The Boston University Radio News Corpus" Boston University, Technical Report ECS-95-001 (1995)

Peppe, S., Private Correspondence (2006)

Read I., Cox S., "Using part-of-speech for predicting phrase breaks." In *INTERSPEECH-2004* pp. 741-744 (2004)

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

Roach P., *English Phonetics and Phonology: A Practical Course* (3rd. edition) Cambridge, Cambridge University Press (2000)

Shriberg, L., Lof, G., "Reliability studies in broad and narrow phonetic transcription", *Clinical Linguistics and Phonetics*, Vol. 5, No. 3, pp.225-279 (1991)

Taylor P., Black A.W., "Assigning Phrase Breaks from Part-of-Speech Sequences." In *Computer Speech and Languag*e, 12(2) pp. 99-117 (1998)

Taylor L.J., Knowles G., *Manual of Information to Accompany the SEC Corpus: The machine readable corpus of spoken English* University of Lancaster (1988) [Accessed September, 2006: http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM]

Santorini, B., *Part-of-Speech tagging guidelines for the Penn Treebank Project.* Technical report MS-CIS-90-47, University of Pennsylvania (1990)

Van Zaanen M., Roberts, A., Atwell, E., "A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation" In Kranias, L., Calzolari, N., Thurmair, G., Wilks, Y., Hovy, E., Magnusdottir, G., Samiotou, A., Choukri, K., (editors) *Proceedings of LREC'04 Workshop on The Amazing Utility of Parallel and Comparable Corpora*, pp. 58-61 European Language Resources Association (2004)

Wang M.Q., Hirschberg J., "Predicting intonational phrasing from text." ACL'91, (1991)