The value of minimal prosodic information in spoken language corpora

Caroline Lyon and Jill Hewitt
Computer Science Department
University of Hertfordshire, UK

Abstract

This paper reports on an investigation into representing tone unit boundaries (pauses) as well as words in a corpus of spoken English. An analysis of data from MARSEC (Machine Readable Spoken English Corpus) shows that, for professional speakers, the inclusion of this minimal prosodic information will lower the perplexity of a language model. The analysis is based on information theoretic techniques, and an objective method of evaluation is provided by entropy indicators, which are explained. This result is of general interest, and supports the development of improved language models for many applications. The automated capture of pauses seems to be technically feasible, and warrants further investigation. The specific issue which prompted this investigation is a task in broadcasting technology: the semi-automated production of online subtitles for live television programmes. This task is described, and an approach to it using speech recognition technology is explained.

1 Introduction

There are a range of choices to be made when deciding on the contents of spoken language corpora. Useful information will be made available by including prosodic annotations but this is difficult to automate. This paper looks at the advantages that can be expected to accrue from the inclusion of minimal prosodic information, major and minor tone unit boundaries, or pauses. Capturing this information automatically does not present the same difficulties as producing a full prosodic annotation: a method is described by (Huckvale and Fang, 1996). A purpose of this paper is to show that the development of such methods warrants further investigation.

The task which prompted this investigation will use trained speakers in a controlled environment, as described below.

We investigate how much extra information is captured by representing major and minor pauses, as well as words, in a corpus of spoken English. We find that for speech such as broadcast news or talks the inclusion of this minimal prosodic annotation will lower the perplexity of a language model.

This result is of general interest, and supports the development of improved language models for many applications. However, the specific issue which we address is a task in broadcasting technology: the semi-automated production of online subtitles for live television programmes. Contemporaneous subtitling of television programmes for the hearing impaired is set to increase significantly, and in the UK there is a mandatory requirement placed on TV broadcasters to cover a certain proportion of live programmes. This skilled task is currently done by highly trained stenographers (the subtitlers), but it could be semi-automated. First, the subtitlers can transfer broadcast speech to written captions using automated speech recognition systems instead of specialist keyboards. A similar approach is being introduced for some Court reporting, where trained speakers in a controlled environment may replace traditional stenographers. Secondly, the display of the captions can be automated, which is the problem we address here. It is necessary to place line breaks in appropriate places so the subtitler can be relieved of this secondary task.

Based on previous work in this field (Lyon and Frank, 1997) a system will be developed to process the output of an ASR device. We need to collect examples of this output as corpora of training data, and have investigated the type

of information that it will be useful to obtain. Commercially available speech recognizers typically output only the words spoken by the user, but as an intermediate stage in producing subtitles we may want to use prosodic information that has been made explicitly available.

Information theoretic techniques can be used to determine how useful it is to capture minimal prosodic information, major and minor pauses. The experiments reported here take a prosodically marked corpus of spoken English, the Machine Readable Spoken English Corpus (MARSEC). Different representations are compared, in which the language model either omits or includes major and minor pauses as well as words.

We need to assess how much the structure of language is captured by the different methods of representation, and an objective method of evaluation is provided by entropy indicators, described below.

The relationship between prosody and syntax is well known (Arnfield, 1994; Fang and Huckvale, 1996; Ostendorf and Vielleux, 1994; Taylor and Black, 1998). Work in this field has typically focussed on the problem in speech synthesis of mapping a text onto natural sounding speech. Our work investigates the complementary problem of mapping speech onto segmented text.

It is not claimed that pauses are *only* produced as structural markers: hesitation phenomena perform a number of roles, particularly in spontaneous speech. However, it has been shown that the placement of pauses provide clues to syntactic structure. We introduce a statistical measure that can help indicate whether it is worth going to the trouble of capturing certain types of prosodic information for processing the output of trained speakers.

Contents of paper

This paper is organised in the following way. In Section 2 we describe the MARSEC corpus, which is the basis for the analysis. Section 3 describes the entropy metric, and explains the theory behind its application. Section 4 describes the experiments that were done, and gives the results. These indicate that it will be worthwhile to capture minimal prosodic information. In Section 5 we describe the subtitling task which prompted this investigation, and the paper concludes with Section 6 which puts our work into a wider context.

2 MARSEC : Machine Readable Spoken English Corpus

Since trained speakers will be producing the subtitles a corpus of speech from professional broadcasters is appropriate for this initial investigation. The MARSEC corpus has been mainly collected from the BBC, and is available free on the web (see references). It is marked with prosodic annotations but not with POS tags. We have used part of the corpus, just over 26,000 words, comprising the 4 categories of news commentary (A), news broadcasts (B), lectures aimed at a general audience (C) and lectures aimed at a restricted audience (D).

The prosodic markup has been done manually, by two annotators. Some passages have been done by both, and we see that there is a general consensus, but some differing decisions. Interpreting the speech data has an element of subjectivity. In Table 1 we show some sample data as we used it, in which only the major and minor tone unit boundaries are retained. When passages were marked up twice, we chose one in an arbitrary way, so that each annotator was chosen about equally.

2.1 Comparison of automated and manual markup of pauses

We suggest that the production of this type of data may be technically feasible for a trained speaker using an ASR device, and is worth investigating further.

(Huckvale and Fang, 1996) describe their method of automatically capturing pause information for the PROSICE corpus. The detection of major pauses is technically straightforward: they find regions of the signal that fall below a certain energy threshold (60Db) for at least 250ms. Minor pauses are more difficult to find, since they can occur within words, and their detection is integrated into the signal / word alignment process.

We find that in that in the manually annotated MARSEC corpus, the ratio of words to major pauses is approximately 17.8, to minor pauses 5.4, or 4.1 if both type of pause are taken together. (Huckvale and Fang, 1996) quote figures that work out at 7.7 for major pauses, 30.8 for minor ones, or 6.15 taken together. This suggests that there is some discrepancy between

Key:	
‖ is major pause	| is minor pause
annotator 1	annotator 2
we	we
heard	heard
automatic	automatic
fire	fire
|	|
a	a
few	few
yards	yards
away	away
|	‖
we	we
drove	drove
on	on
‖	‖
a	a
jet	jet
appeared	appeared

Table 1: Example of MARSEC corpus with minimal prosodic annotations

what is considered a major or minor pause. However, taking both together the results from the automated system is not out of line with the manual one on this statistical measure.

3 Entropy indicators

The entropy is a measure, in a certain sense, of the degree of unpredictability. If one representation captures more of the structure of language than another, then the entropy measures should decline.

If H represents the entropy of a sequence and P the perplexity, then

$$P = 2^H$$

P can be seen as the branching factor, or number of choices.

3.1 Definition of entropy

Let $\mathcal{A}$ be an alphabet, and X be a discrete random variable. The probability mass function is then $p(x)$, such that

$$p(x) = probability(X = x), x \in \mathcal{A}$$

For an initial investigation into the entropy of letter sequences the x's would be the 26 letters of the standard alphabet.

The entropy $H(X)$ is defined as

$$H(X) = -\sum_{x \in \mathcal{A}} p(x) * log\ p(x)$$

If logs to base 2 are used, the entropy measures the minimum number of bits needed on average to represent X: the wider the choice the more bits will be needed to describe it.

We talk loosely of the entropy of a sequence, but more precisely consider a sequence of symbols X_i which are outputs of a stochastic process. We estimate the entropy of the distribution of which the observed outcome is typical.

Further references are (Bell et al., 1990; Cover and Thomas, 1991), or, for an introduction, (Charniak, 1993, Chapter 2).

3.1.1 Shannon's work

Though we are investigating sequences of words, the subject is introduced by recalling Shannon's well known work on the entropy of letter sequences (Shannon, 1951). He demonstrated that the entropy will decline if a representation is found that captures (i) the context and (ii) the structure of the sequence.

Shannon produced a series of approximations to the entropy H of written English, which successively take more of the statistics of the lan-

guage into account. H_0 represents the average number of bits required to determine a letter with no statistical information. Thus, for an alphabet of 16 symbols $H_0 = 4.0$.

H_1 is calculated with information on single letter probabilities. If we knew, for example, that letter e had probability of 20% of occurring while z had 1% we could code the alphabet with, on average, fewer bits than we could without this information. Thus H_1 would be lower than H_0.

H_2 uses information on the probability of 2 letters occurring together; H_n, called the n-gram entropy, measures the amount of entropy with information extending over n adjacent letters of text, [1] and $H_n \leq H_{n-1}$. As n increases from 0 to 3, the n-gram entropy declines: the degree of predictability is increased as information from more adjacent letters is taken into account. This fact is exploited in games where the contestants have to guess letters in words, such as the "Shannon game" or "Hangman" (Jelinek, 1990).

The formula for calculating the entropy of a sequence is given in (Lyon and Brown, 1997). An account of the process is also given in (Cover and Thomas, 1991, chapter2) and (Shannon, 1951).

3.2 Entropy and structure

The entropy can also be reduced if some of the structure of the letter strings is captured. As Shannon says "a word is a cohesive group of letters with strong internal statistical influences" so the introduction of the space character to separate words should lower the entropy H_2 and H_3. With an extra symbol in the alphabet H_0 will rise. There will be more potential pairs and triples, so H_2 and H_3 could rise. However, as the space symbol will prevent "irregular" letter sequences between words, and thus reduce the unpredictability H_2 and H_3 do in fact decline. For instance, for the words

COOKING CHOCOLATE

the trigrams "N-G-C" and "G-C-H" will be replaced by "N-G-space", "G-space-C" and "space-C-H"..

---

[1] This notation is derived from that used by Shannon. It differs from that used by (Bell et al., 1990).

3.3 The entropy of ASCII data

For other representations too, the insertion of boundary markers that capture the structure of a sequence will reduce the entropy. Gull and Skilling (1987) report on an experiment with a string of 32,768 zeroes and ones that are known to be ASCII data organised in patterns of 8 as bytes, but with the byte boundary marker missing. By comparing the entropy of the sequence with the marker in different positions the boundary of the data is "determined to a quite astronomical significance level".

3.4 The entropy of word sequences

This method of analysis can also be applied to strings of words. The entropy indicator will show if a sequence of words can be decomposed into segments, so that some of the structure is captured. Our current work investigates whether pauses in spoken English perform this role.

Previously we showed how the entropy of text mapped onto part-of-speech tags could be reduced if clauses and phrases were explicitly marked (Lyon and Brown, 1997). Syntactic markers can be considered analogous to spaces between words in letter sequence analysis. They are virtual punctuation marks.

Consider, for example, how subordinate clauses are discerned. There may be an explicit opening marker, such as a 'wh' word, but often there is no mark to show the end of the clause. If markers are inserted and treated as virtual punctuation some of the structure is captured and the entropy declines. A sentence without opening or closing clause boundary markers, like

The shirt he wants is in the wash.

can be represented as

The shirt { he wants } is in the wash.

This sentence can be given part-of-speech tags, with two of the classes in the tagset representing the symbols '{' (virtual-tag1) and '}' (virtual-tag2). The ordinary part-of-speech tags have probabilistic relationships with the virtual tags in the same way that they do with each other. The pairs and triples generated by the second string exclude (noun, pronoun), (noun, pronoun, verb) but include, for instance, (noun, virtual-tag1), (noun, virtual-tag1, pronoun)

43

Using this representation, the entropy, H_2 and H_3, with virtual tags explicitly marking some constituents is lower than that without the virtual tags. In a similar way the words from a speech signal can be segmented into groups, with periodic pauses.

4 Results from analysis of the MARSEC corpus

We can measure the entropy H_0, H_1, H_2 and H_3 for the corpus with and without prosodic markers for major and minor pauses. However, rather than use words themselves we map them onto part-of-speech tags. This reduces an indefinite number of words to a limited number of tags, and makes the investigation computationally feasible. We expect

- H_0 will be higher with a marker, since the alphabet size increases

- H_1, which takes into account the single element probabilities, will increase or decrease depending on the frequency of the new symbol.

- H_2 and H_3 should fall if the symbols representing prosodic markers capture some of the language structure. We expect H_3 to show this more than H_2.

- If instead of the real pause markers mock ones are inserted in an arbitrary fashion, we expect H to rise in all cases.

To conduct this investigation the MARSEC corpus was taken off the web, and pre-processed to leave the words plus major and minor tone unit boundaries, or pauses. Then it was automatically tagged, using a version of the Claws tagger[2]. These tags were mapped onto a smaller tagset with 26 classes, 28 including the major and minor pauses. The tagset is given in the appendix. Random inspection indicated about 96% words correctly tagged.

Then the entropy of part of the corpus was calculated (i) for words only (ii) with minor pauses represented (iii) with major pauses represented and (iv) with major and minor pauses represented. Results are shown in Table 2, and in Figure 1.

[2]Claws4, supplied by the University of Lancaster, described by Garside (1987)

H_3 is calculated in two different ways. First, the sequence of tags is taken as an uninterrupted string (column H_3 (1) in Table 2). Secondly, we take the major pauses as equivalent to sentence ends, points of segmentation, and omit any triple that spans 2 sentences (column H_3 (2)). In practice, this will be a sensible approach.

This experiment shows how the entropy H_3 declines when information on pauses is explicitly represented. Though there is not a transparent mapping from prosody to structure, there is a relationship between them which can be exploited. These experiments indicate that English language used by professional speakers can be coded more efficiently when pauses are represented.

4.1 Comparison with arbitrarily placed pauses

Compare these results to those of another experiment where the corpora of words only were taken and pauses inserted in an arbitrary manner. Major pauses were inserted every 19 words, minor pauses every 7 words, except where there is a clash with a major pause. The numbers of major and minor pauses are comparable to those in the real data. Results are shown in Table 3. H_2 and H_3 are higher than the comparable entropy levels for speech with pauses inserted as they were actually spoken. Moreover, the entropy levels are higher than for speech without any pauses: the arbitrary insertion has disrupted the underlying structure, and raised the unpredictability of the sequence.

4.2 Entropy and corpus size

Note that we are interested in *comparative* entropies. The entropy converges slowly to its asymptotic value as the size of the corpora increases, and this is an upper bound on entropy values for smaller corpora. Ignoring this may give misleading results (Farach and et al., 1995). The reason why entropy is underestimated for small corpora comes from the fact that we approximate probabilities by frequency counts, and for small corpora these may be poor approximations. The count of pairs and triples is the basis for the probability estimates, and with small corpora many of the triples in particular that will show later have not occurred. Thus the entropy and perplexity measures un-
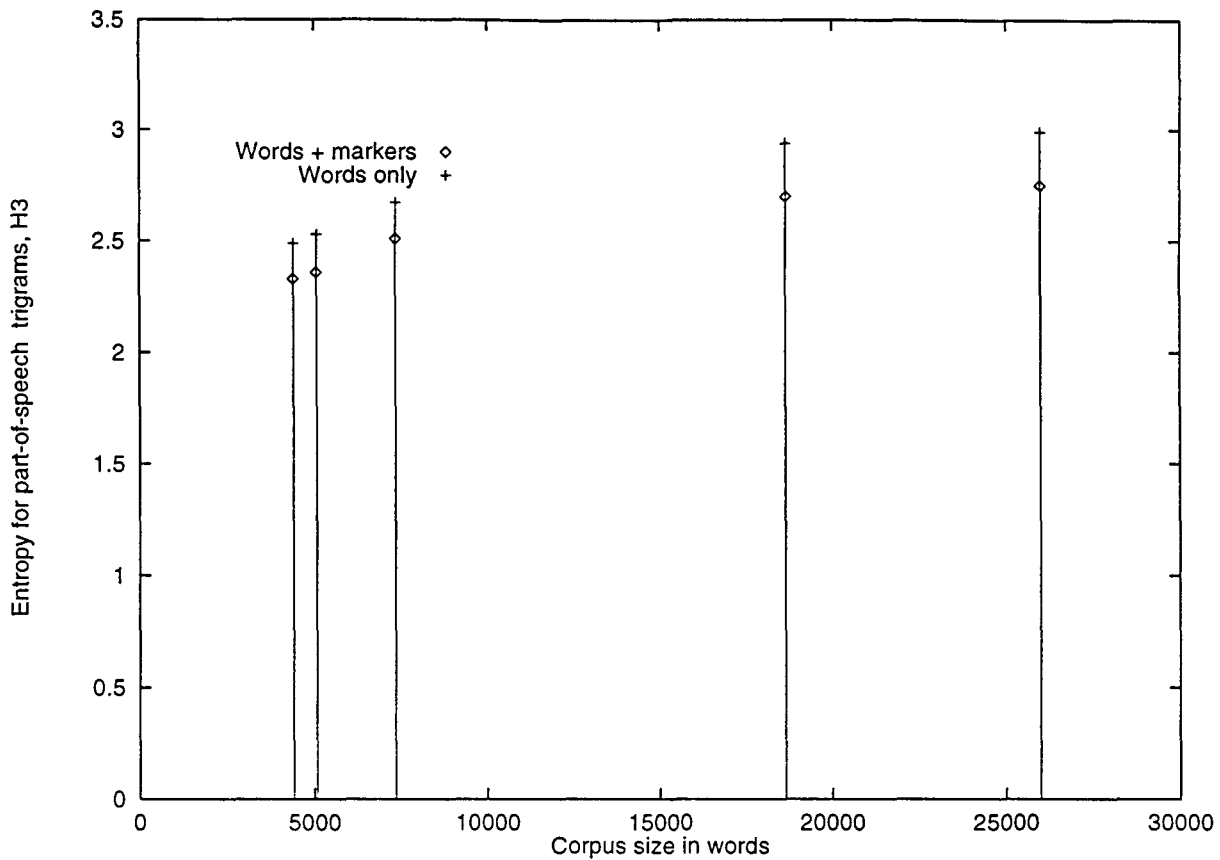
Figure 1: Comparison of trigram part-of-speech entropy for sections of the MARSEC corpus, (i) with both major and minor pauses marked (ii) without either. The tagset size is 28 with the pauses represented, 26 without them. The entropy is calculated with trigrams spanning a major pause omitted, as in Table 2 column H_3 (2)

derestimate their true values.

5 The subtitling task

We now show how the investigation described here is relevant to the subtitling task. Trained subtitlers are employed to output real time captions for some TV programmes, currently as a stream of type written text. In future this may be done with an ASR system. In either case, the production of the caption is followed by the task of displaying the text so that line breaks occur in natural positions. The type of programmes for which this is needed include contemporaneous commentary on on news events, sports, studio discussions, and chat shows.

Caption format is limited by line length, and there are usually at most 2 lines per caption. Some examples of subtitles that have been

displayed are taken from the broadcast commentary on Princess Diana's funeral, with line breaks as shown:

```
As I said the great tenor bell is
half muffled with
```

```
a piece of leather around its
clapper
```

```
They now bear the coffin of the
Princess
```

```
of Wales into Westminster Abbey.
```

An example from a chat show is:

```
Who told you that you resemble Mr
Hague?
```

45

Speech representation	Number of minor pauses	Number of major pauses	H_0	H_1	H_2	H_3 (1)	H_3 (2)
Words only	0	0	4.70	4.11	3.29	2.94	2.94
Words + minor pauses	3454	0	4.75	4.09	3.18	2.84	2.84
Words + major pauses	0	1029	4.75	4.19	3.32	2.94	2.84
Words + both pauses	3454	1029	4.81	4.17	3.16	2.82	2.70

Table 2: Entropy measures for 18655 words of the MARSEC corpus, (sections A, B, C concatenated) with and without major and minor pauses. H_3 (2) measures entropy without triples spanning a major pause (see text).

Speech representation	Number of minor pauses	Number of major pauses	H_0	H_1	H_2	H_3 (1)
Words + pauses in arbitrary positions	3109	1209	4.81	4.19	3.63	3.05

Table 3: Entropy measures for the same part of the MARSEC corpus with pauses in arbitrary positions : a major pause every 19 words, minor pause every 7 words (except for clashes with major)

```
I work at a golf club and we have
lots

of societies and groups come in.
```

The quality of the subtitles can be improved by placing the line breaks and caption breaks in places where the text would be naturally segmented. Though this is partially a subjective process, a style book can be produced that gives agreed guidelines.

Some of the poor line breaks can be readily corrected, but the production of a high quality display overall is not a trivial task. The pauses in speech do not map straight onto suitable line breaks, but they are a significant source of information. In this work we have been considering the output of trained speakers, or the recording of rehearsed speech. This differs from ordinary, spontaneous speech, where hesitation phenomena may have a number of causes. However, in the type of speech we are processing we have shown that the use of pauses captures some syntactic structure. An example given by (Ostendorf and Vielleux, 1994) is

```
Mary was amazed Ann Dewey was
angry.
```

which was produced as

```
Mary was amazed || Ann Dewey was
angry.
```

To illustrate a problem of text segmentation consider how conjunctions should be handled. Now conjunctions join like with like: verb with verb, noun with noun, clause with clause, and so on. If a conjunction joins two single words, such as "black and blue" we do not want it to trigger a line break. However, it may be a reasonable break point if it joins two longer components. Consider the following example from the MARSEC corpus:

```
it held its trajectory for one
minute | flashes burst | from
its wings | and rockets
exploded | safely behind us ||
```

The word "and" without the pause marked is part of a trigram "noun conjunction noun" which would typically stick together. In fact, it actually joins two sentences, and would be a good point for a break. By including the pause marker we can identify this.

The proposed system for finding line breaks will integrate rule based and data driven components. This approach is derived from earlier work in a related field in which a partial parser has been developed (Lyon and Frank, 1997). It will be based on a part-of-speech trigram model combined with lexical information. We will be able to develop a better language model if we explicitly include a representation for major and minor pauses.

6 Role of pauses in speech in a wider context

As hypothesized the entropy, H_3, declines as the major and minor boundary markers are inserted. This indicates that it will be worthwhile to capture the prosodic information on major and minor pauses from the ASR, in addition to the usual transcription of the words themselves.

Our investigation was prompted by a specific task in which the output of trained speakers is transcribed automatically. However, it is of wider interest. We show that representing pauses as well as words helps determine the structure of language, and thus contribute to the quality of a language model.

It is many years since Mandelbrot investigated the way in which the statistical structure of language is best adapted to coding words (Mandelbrot, 1952). He suggested that language is "intentionally if not consciously produced in order to be decoded word by word in the easiest possible fashion." If we accept his suggestion we would expect that naturally occurring events, such as pauses in speech, are utilised to facilitate the transfer of information. Patterns of speech segmentation are likely to emerge to produce an efficient coding (Lyon, 1998).

References

S Arnfield. 1994. *Prosody and Syntax in Corpus Based Analysis of Spoken English.* Ph.D. thesis, University of Leeds.

T C Bell, J G Cleary, and I H Witten. 1990. *Text Compression.* Prentice Hall.

E Charniak. 1993. *Statistical Language Learning.* MIT Press.

T M Cover and J A Thomas. 1991. *Elements of Information Theory.* John Wiley and Sons Inc.

Alex Chengyu Fang and Mark Huckvale. 1996. Synchronising syntax with speech signals. In V.Hazan, M.Holland, and S.Rosen, editors, *Speech, Hearing and Language.* University College London.

M Farach and M Noordewier et al. 1995. On the entropy of dna. In *Symposium on Discrete Algorithms.*

R Garside. 1987. The CLAWS word-tagging system. In R Garside, G Leech, and G Sampson, editors, *The Computational Analysis of English: a corpus based approach,* pages 30–41. Longman.

S Gull and J Skilling. 1987. Recent developments at cambridge. In C Ray Smith and Gary Erickson, editors, *Maximum -Entropy and Bayesian Spectral Analysis and Estimation Problems.*

M Huckvale and A C Fang. 1996. PROSICE: A spoken English database for prosody research. In S Greenbaum, editor, *Comparing English Worldwide: The International Corpus of English.* O U P.

F Jelinek. 1990. Self-organized language modeling for speech recognition. In A Waibel and K F Lee, editors, *Readings in Speech Recognition,* pages 450–503. Morgan Kaufmann. IBM T.J.Watson Research Centre.

C Lyon and S Brown. 1997. Evaluating Parsing Schemes with Entropy Indicators. In *MOL5, 5th Meeting on the Mathematics of Language.*

C Lyon and R Frank. 1997. Using Single Layer Networks for Discrete, Sequential Data: an Example from Natural Language Processing. *Neural Computing Applications,* 5 (4).

C Lyon. 1998. Language evolution: survival of the fittest in the statistical environment. Technical report, Computer Science Department, University of Hertfordshire, June.

B Mandelbrot. 1952. An informational theory of the statistical structure of language. In *Symposium on Applications of Communication Theory.* Butterworth.

M Ostendorf and N Vielleux. 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics,* 20(1).

C E Shannon. 1951. Prediction and Entropy of Printed English. *Bell System Technical Journal,* pages 50–64.

P Taylor and A Black. 1998. Assigning phrase breaks from part-of-speech sequences.

Appendix

Description of the Tagset

The tagset used in these experiments is derived from CLAWS4, mapped onto a smaller set of classes. They are as follows

- article or determiner - singular
- article or determiner - plural
- predeterminer e.g. "all"
- pronomial determiner e.g. "some"
- pronomial determiner - singular
- proper noun
- noun - singular
- noun - plural
- pronoun - singular
- pronoun - plural
- relative pronoun
- possessive pronoun
- verb - singular
- verb - plural
- auxiliary verb - singular
- auxiliary verb - plural
- existential "here" or "there"
- present participle
- past participle
- infinitive "to"
- preposition
- conjunction
- adjective
- singular number "one"
- adverb
- exceptions

Two extra tag classes are added for the analysis of tone unit boundaries:

- minor pause
- major pause

The tagging process includes the identification of common phrases or idioms, which are then treated as single lexical items. For instance, "of course" is tagged as an adverb.