ProPOSEC: A Prosody and PoS Annotated Spoken English Corpus

Claire Brierley[1 2], Eric Atwell[1]
[1]School of Computing, University of Leeds, LEEDS LS2 9JT, UK
[2]School of Business and Creative Technologies, University of Bolton, Deane Rd., BOLTON BL3 5AB, UK
E-mail: scscb@leeds.ac.uk; C.Brierley@bolton.ac.uk; eric@comp.leeds.ac.uk

Abstract

We have previously reported on ProPOSEL, a purpose-built Prosody and PoS English Lexicon compatible with the Python Natural Language ToolKit. ProPOSEC is a new corpus research resource built using this lexicon, intended for distribution with the Aix-MARSEC dataset. ProPOSEC comprises multi-level parallel annotations, juxtaposing prosodic and syntactic information from different versions of the Spoken English Corpus, with canonical dictionary forms, in a query format optimized for Perl, Python, and text processing programs. The order and content of fields in the text file is as follows: (1) Aix-MARSEC file number; (2) word; (3) LOB PoS-tag; (4) C5 PoS-tag; (5) Aix SAM-PA phonetic transcription; (6) SAM-PA phonetic transcription from ProPOSEL; (7) syllable count; (8) lexical stress pattern; (9) default content or function word tag; (10) DISC stressed and syllabified phonetic transcription; (11) alternative DISC representation, incorporating lexical stress pattern; (12) nested arrays of phonemes and tonic stress marks from Aix. As an experimental dataset, ProPOSEC can be used to study correlations between these annotation tiers, where significant findings are then expressed as additional features for phrasing models integral to Text-to-Speech and Speech Recognition. As a training set, ProPOSEC can be used for machine learning tasks in Information Retrieval and Speech Understanding systems.

1. Introduction

The authors have previously reported on ProPOSEL, a purpose-built ***pro***sody and ***PoS*** ***E***nglish ***L***exicon (Brierley and Atwell, 2008a; 2008b), compatible with the Python-based Natural Language ToolKit, and widely-used text corpora, for automatic annotation of text with real-world knowledge of prosody and syntax, and for exploring subtle linguistic features of text which may enhance the performance of classifiers traditionally trained on syntactic and graphemic features. Here, we report on ProPOSEC, a dataset built using this lexicon: namely, a version of Section A (Commentary) in SEC, the ***S***poken ***E***nglish ***C***orpus (Taylor and Knowles, 1988) with multi-level parallel annotations juxtaposing linguistic information from different versions of the corpus with canonical dictionary forms, in a format optimized for query with Perl or Python and other text processing programs. We first describe the contents of this resource, and how ProPOSEL was used to create it. We then describe how this dataset may be used in studies of correlations between these multi-level annotation tiers, with reference to recent work by the authors. The ProPOSEC dataset is intended for distribution with an updated version of the Aix-MARSEC corpus project (Hirst *et al.*, 2009).

2. Dataset fields

The prototype ProPOSEC dataset merges selected information from Aix-MARSEC (*i.e.* file number; word token; SAMPA phonetic transcription; and tonic stress marks assigned to each segment) with syntactic annotations from SEC, plus corresponding syntactic annotations and canonical pronunciations in the ProPOSEL lexicon. In addition, pauses denoting the original 'gold-standard' phrase break annotations in SEC are aligned with punctuation where appropriate.

Currently, the order and content of fields in the text file is as follows:

(1) Aix-MARSEC file number; (2) word; (3) LOB PoS-tag; (4) C5 PoS-tag; (5) Aix SAM-PA phonetic transcription; (6) SAM-PA phonetic transcription from ProPOSEL; (7) syllable count; (8) lexical stress pattern; (9) default content or function word tag; (10) DISC stressed and syllabified phonetic transcription; (11) alternative DISC representation, incorporating lexical stress pattern; (12) nested arrays of phonemes and tonic stress marks from Aix.

The Lancaster-Oslo-Bergen (LOB) tag set (Johansson *et al.*, 1986) was used to syntactically annotate the original version of SEC and is more fine-grained than C5, the PoS tag set used in the British National Corpus (Leech and Smith, 2000). DISC phonetic transcriptions are unique in providing a one-to-one mapping between character and sound for both long vowels and affricates (*i.e.* the consonants in *chin* and *gin*). Lexical stress patterns are abstract representations of rhythmic structure, as in the sequence 201 for *disappear*, where each syllable is assigned a stress weighting: 1 for primary stress, 2 for secondary stress and 0 for unstressed elements. Readers are referred to recent publications from the authors for further discussion of DISC and lexical stress patterns (Brierley and Atwell, 2008a; 2008b).

2.1 Prosody fields in the dataset

Like ProPOSEL, the ProPOSEC dataset serves as a repository of key prosodic information to do with prominence and intonation. The former is a property of syllables which are perceived as being louder and longer than others and which may enact changes in pitch. The latter is generally a property of the phrase or sentence, and refers to utterance tunes: recognisable patterns in a series of pitch movements which have semantic and functional significance. In English, prominent syllables can be stressed, in which case they are perceived as strong rhythmic beats endowed with a full vowel, not a reduced one. Stressed syllables may also be accented, in which case they initiate a change in the direction of pitch and sometimes a sharp jump across the speaker's pitch range. British English has six distinct pitch accent types: level; rising; falling; rising-falling; falling-rising; and rising-falling-rising (Grabe, 2001).

Listing 1 shows linguistic annotations in ProPOSEC for a prosodic-syntactic chunk initiated by a major clause boundary, the snippet *soon after it took off from Athens airport* from Section A08 of the corpus, with items in **bold** selected for further comment (*cf.* 2.1 and 2.2).

```
A0801|soon|RB|AV0|su:n|sun|1|1|C|'sun|'sun:1|[['s',
'u:', 'n'], ['\\', '\\', '\\']]
A0801|after|CS|CJS|A:ft@|'Aft@R|2|10|F|'#f-t@R|'#f
:1 t@R:0|[['A:', 'f', 't', '@'], ['0', '0', '0', '0']]
A0801|it|PP3|PNP|rIt|It|1|1|F|'It|'It:1|[['r', 'I', 't'],
['0', '0', '0']]
A0801|took|VBD|VVD|tUk|tUk|1|1|C|'tUk|'tUk:1|[['t
', 'U', 'k'], ['`', '`', '`']]
A0801|off|RP|AVP|Qf|0f|1|1|C|'Qf|'Qf:1|[['Q', 'f'],
['0', '0']]
A0801|from|IN|PRP|fr@m|fr0m|1|1|F|'frQm|'frQm
:1|[['f', 'r', '@', 'm'], ['0', '0', '0', '0']]
A0801|athens|NP|NP0|{TInz|'&TInz|2|10|C|No
value|No value|[['{', 'T', 'I', 'n', 'z'], ['*', '0', '0', '0', '0']]
A0801|airport|NN|NN1|e@pO:t|'e@pOt|2|10|C|'8-p
$t|'8:1 p$t:0|[['e@', 'p', 'O:', 't'], ['`/', '0', '0', '0']]
A0801|PAUSE|,|,
```

Listing 1: Parallel linguistic annotations for each word token include a prototype mapping between phones and tonic stress marks

2.2 Elisions

Differences in ProPOSEC's SAM-PA transcriptions from Aix-MARSEC (field 5) and the lexicon (field 6) arise in part due to the former implementing elision rules for optimizing raw phonemic transcriptions (Auran *et al.*, 2004). Hence, in Listing 1, the Aix transcription for *it* shows a linking 'r'. Link-ups effected by w-glides and y-glides (Mortimer, 1985:46) are not included and constitute a potential enhancement for Aix-MARSEC and ProPOSEC. For example, greater verisimilitude to spoken English could be achieved quite simply by an extra rule governing use of the definite article (*cf.* 2.2).

2.3 Reduced forms

Another difference in ProPOSEC's SAM-PA transcriptions in fields (5) and (6) is more extensive representation of reduced vowels in function words in Aix-MARSEC. Hence we have an optimized versus canonical transcription for *from* in Listing 1. Definite articles in Aix-MARSEC are transcribed one of two ways: /D@/ - incorporating a schwa and identical to their SAM-PA transcriptions in the lexicon; and /DI/ - modelling coarticulation before vowels as in: /DI/ and /A:mI/ for *the army* (Aix-MARSEC A0402). As suggested in Section 2.1, elision prediction could include a linking 'y' in such instances: / DIjA:mI/ for *the ↺ army*.

3. Dataset build

In this section, we discuss the algorithm used to merge data from two different versions of the corpus (SEC and Aix-MARSEC) with canonical dictionary forms from ProPOSEL. A visual representation of the algorithm summarises preceding explanation and justification at each step in this segmented process – see Appendix.

One incentive for creating the ProPOSEC dataset was to enrich annotations in Aix-MARSEC, which at time of writing comprises multi-level prosodic annotation tiers but lacks syntactic information. NLP resources at the University of Leeds include a version of SEC tagged with the LOB tag set; but aligning word-LOB pairings in SEC with information from the concatenated version (2006:02:27) of Aix-MARSEC used was non-trivial. An initial problem is that some orthographic forms in SEC (*i.e.* hyphenated compounds and abbreviations) are decomposed into multiple phonetic and prosodic units in Aix-MARSEC: for example, the TextGrid file for A0802 in Aix shows decomposition of the word *x-ray* into two separate **n**arrow **r**hythm **u**nits (NRU), equivalent to two stressed feet.

SYLLABLES TIER: A0802B	JASSEM TIER: A0802B
8.3460000000000001	8.3460000000000001
"""" e k s"	**"NRU"**
8.3460000000000001	8.3460000000000001
8.6959999999999997	8.6959999999999997
"""" r eI"	**"NRU"**
8.6959999999999997	8.6959999999999997

Table 1: Data from 2 prosodic annotation tiers (syllables and rhythmic units) in an Aix-MARSEC TextGrid file

The first step was therefore to reconcile, manually, orthography in SEC Section A with that of Aix: for example, *TWA* (airlines) in A08 becomes *tee double u ay* and so on.

After automatically reconstituting enclitics in SEC (*e.g.* will_MD not_XNOT in LOB becomes won't_MD+XNOT) in Step 2, the most intractable problem was mapping PoS tags from SEC with data from Aix (Step 3); in this merger, files are of different lengths, due to asynchronous distribution of punctuation (in SEC) and pauses/phrase break annotations (in Aix).

The ProPOSEC dataset includes PoS tags from two schemes which differ in 'delicacy' (Atwell, 2008). C5 is a much sparser tagset than LOB. It is also integral to dictionary lookup via ProPOSEL. The algorithm addresses this mismatch in delicacy between the tagsets in Steps 4 and 5. The former instantiates a live one-to-many mapping of C5<LOB PoS tags from the imported ProPOSEL lexicon. Examples in Table 2 show rafts of

LOB tags mapped to C5 in the single category of adverbs, plus category combinations involving proper nouns, along with potential problems which lurk the other way: prepositions and subordinating conjunctions in LOB with more than one equivalent in C5.

Syntactic Category	**C5**	**LOB**
Adverbs	AV0	['QL', 'QLP', 'RB', 'RI', 'RBR', 'RBT', 'RN']
Enclitic: proper noun with *has*	NP0+POS	['NP$', 'NPL$', 'NPLS$', 'NPS$', 'NPT$', 'NPTS$']
Preposition: *of*	PRF	IN
Prepositions	PRP	IN
Subordinating conjunction: *that*	CJT	CS
Subordinating conjunctions	CJS	CS

Table 2: One-to-many mappings for C5 and LOB occur both ways

A match between LOB tokens in the merged dataset and the live mapping in ProPOSEL appends the corresponding C5 tag to dataset arrays (Step 5) and a patch is implemented to remove redundant C5 tags in cases of LOB<C5. Very few items remain untagged at this stage and can therefore be repaired manually: for example there were only 15 untagged items remaining out of 629 word tokens in Section A08.

Finally, we transform ProPOSEL into a Python dictionary using ProPOSEL's bespoke software tools (Brierley and Atwell, 2008a), with compound (word + C5) keys mapped to prosodic-syntactic value arrays from selected fields in the lexicon. Intersection between dictionary keys and (word + C5) pairings in the dataset appends dictionary values to the parallel position in that sequence object (Step 6).

4. Experimentation with the dataset

The ProPOSEL lexicon was purpose-built to integrate and leverage domain knowledge from several well-established lexical resources for corpus-based research and language engineering tasks in English. One such task is supervised learning of phrase break prediction, which requires the binary classification of word tokens in the training set into breaks (*i.e.* words followed by a major or minor intonation unit boundary) or non-breaks. Listing 2a shows as an example a sentence snippet in A11 in the ProPOSEC dataset, '…*palace which houses the central committee of the communist party*…' annotated with both LOB and C5 part-of-speech tags; this illustrates sparse prosodic phrasing. Listing 2b shows break classifications for C5 PoS trigrams derived from this snippet. The snippet initially consists of word tokens carrying LOB and C5 tags but only the latter are used.

```
...palace_IN_NN1 which_WP_PNQ houses_VBZ_VVZ the_ATI_AT0 central_NP_NP0 committee_NP_NP0 of_IN_PRF the_ATI_AT0 communist_JNP_AJ0 party_NP_NP0 in_IN_PRP...
```

Listing 2a: The string '…*palace which houses the central committee of the communist party*…' annotated with both LOB and C5 part-of-speech tags

```
n = 3
snippet = [snippet[i:i+n] for i in range(len(snippet)-n+1)]
for index in snippet: print index[0][0], index[1][0], index[2][0], index[2][1]
```

```
NN1 PNQ VVZ non_break
PNQ VVZ AT0 non_break
VVZ AT0 NP0 non_break
AT0 NP0 NP0 non_break
NP0 NP0 PRF non_break
NP0 PRF AT0 non_break
PRF AT0 AJ0 non_break
AT0 AJ0 NP0 break
AJ0 NP0 PRP non_break
```

Listing 2b: Python code to extract sliding windows of size 3 capturing C5 PoS trigrams, plus break classification for each trigram

In recent work, the authors have found empirical evidence of a significant correlation in English between 'gold-standard' phrase break annotations in the ProPOSEC dataset and words containing complex vowels in their canonical dictionary pronunciations via the DISC phonetic transcription set in ProPOSEL (Brierley and Atwell, 2009; 2010). This finding suggests English speakers may favour diphthong/triphthong-bearing words as *tonics* (*i.e.* nuclear prominences in tone groups).

Multi-level parallel annotations in the ProPOSEC dataset facilitate statistical analyses of this kind. For example, interesting patterns may emerge in the co-occurrence of tonic stress marks and pauses (perceived phrasing) with punctuation (conceptual phrasing) in particular syntactic contexts. Listing 3 shows one such instance (in **bold**) in A04 where a high fall plus pause (minor boundary) co-occurs with a comma and major clause boundary.

```
A0407|while|CS|CJS|[['w', 'aI', 'l'], ['_', '_', '_']]
A0407|they|PP3AS|PNP|[[ 'D', 'e'], ['0', '0']]
A0407|may|MD|VM0|[['m', 'eI'], ['0', '0']]
A0407|ache|VB|VVI|[['eI', 'k'], ['`', '`']]
A0407|for|IN|PRP|[['f', '@'], ['0', '0']]
A0407|peace|NN|NN1|[['p', 'i:', 's'], ['`', '`', '`']]
A0407|PAUSE|,|,
```

Listing 3: The high fall on *peace* coincides with a minor intonation unit boundary, a comma, and a major clause boundary, and is suggestive of contrastive stress

Thus one application for ProPOSEC would be as part of a training set for the task of automatic punctuation annotation for structuring the output of speech recognizers within Information Retrieval or Speech

Understanding systems (*cf.* Christensen *et al.*, 2001).

4.1 ProPOSEC and machine learning

The ProPOSEC dataset constructs a syntactic, rhythmic, and phonetic profile for each word in the corpus. However, converting this raw data into feature vectors for phrase break prediction using a machine learning toolkit such as WEKA (Hall *et al.*, 2009) is challenging for a number of reasons, especially if the researcher is interested in how interrelationships between syntax, rhythm and pronunciation influence break placement. One problem is the potential number of values for each attribute: number of PoS in the tag set; number of trigram sequences (*cf.* Listing 2b); number of lexical stress patterns. Added to this is the problem of incorporating sufficient context into the language model: for example, the researcher may be interested in a window of *N* words either side of a given index position. Listing 4 shows as an example a basic WEKA arff input file for the snippet *soon after it took off from Athens airport,* derived from the ProPOSEC sample from Listing 1, but only including word, C5 PoS-tag, lexical stess pattern, and a final extra attribute showing whether or not the following ProPOSEL field marks a PAUSE or prosodic break. Even for these restricted fields, simply translating each field in ProPOSEC into a WEKA attribute as in Listing 4 would be cumbersome in terms of values and superficial in terms of modelling.

```
@relation phraseBreak

@attribute word { after, it, took, off, from, athens, airport}
@attribute pos { AVP, CJS, NP0, NN1, PNP, PRP, VVD }
@atttribute lexicalStress { 1, 10, 01 }
@attribute break { yes, no }

@data

soon, AV0, 1, no
after, CJS, 10, no
it, PNP, 1, no
took, VVD, 1, no
off, AVP, 1, no
from, PRP, 1, no
athens, NP0, 10, no
airport, NN1, 10, yes
```

Listing 4: Example .arff file for machine learning in WEKA, but still unsuitable as a training set for phrase break prediction as it does not capture context.

It would require instead a series of complex transformations on the dataset to summarise attribute-value pairs (*e.g.* applying a series of conditions which dictate whether or not a word carries a beat) and to take context into account.

5. Conclusions

ProPOSEC is a new corpus research resource merging ProPOSEL with SEC and Aix-MARSEC, with multi-level parallel annotations juxtaposing linguistic information from different versions of the corpus with canonical dictionary forms, in a query format optimized for text processing programs. The motivation for compiling ProPOSEC was to study correlations between these multi-level annotation tiers and to formulate significant findings as additional features for phrasing models integral to text-to-speech, speech recognition, and related systems.

6. References

Atwell, E. 2008. 'Development of tag sets for part-of-speech tagging.' In Anke Ludeling & Merja Kyto (eds.) *Corpus Linguistics: An International Handbook*. Mouton de Gruyter.

Auran, C., Bouzon, C. and Hirst, D. 'The Aix-MARSEC Project: an Evolutive Database of Spoken British English' in *Proc. Speech Prosody (SP-2004)*: 561-564.

Brierley, C. and Atwell, E. 2008a. 'ProPOSEL: A Prosody and PoS English Lexicon for Language Engineering' in *Proc. 6th Language Resources and Evaluation Conference (LREC 2008).* Marrakech.

Brierley, C., and Atwell, E. 2008b. 'A Human-oriented Prosody and PoS English Lexicon for Machine Learning and NLP' in *Proc. 22nd International Conference on Computational Linguistics (Coling 2008), Workshop on Cognitive Aspects of the Lexicon*. Manchester.

Brierley, C. and Atwell, E. 2009. 'Exploring Phrase Break Correlates in a Corpus of English Speech with ProPOSEL, a Prosody and PoS English Lexicon' in *Proc. Interspeech 2009*. Brighton.

Brierley, C., and Atwell, E. 2010. 'Complex Vowels as Boundary Correlates in a Multi-Speaker Corpus of Spontaneous English Speech'. To appear in *Proc. Speech Prosody 2010*. Chicago.

Christensen, H., Gotoh, Y. and Renals, S. 2001. 'Punctuation Annotation using Statistical Prosody Models' in Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding.

Grabe, E. 2001. 'Prosodic Annotation.' PowerPoint. *9th ELSNET European Summer School on Language and Speech Communication*, *Prague*. Last Accessed: 2006.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. 2009. The WEKA Data Mining Software: An Update. In *SIGKDD Explorations*. 11: 1.

Hirst, D., Auran, C. and Bouzon, C. 2009. 'The Aix-MARSEC Database: version 2009.' Online. Accessed: February 2010. http://crdo.up.univ-aix.fr/voir_depot.php?lang=en&id=33

Johansson, S., Atwell, E., Garside, R. and Leech, G. 1986. 'The Tagged LOB Corpus: Users' Manual.' Bergen. Norwegian Computing Centre for the Humanities.

Leech, G. and Smith, N. 2000. 'Manual to Accompany The British National Corpus (Version 2) with Improved Word-class Tagging.' Online. Accessed: January 2010. http://www.natcorp.ox.ac.uk/docs/bnc2postag_manual.htm

Mortimer, C. 1985. *Elements of Pronunciation*. Cambridge. Cambridge University Press.

Taylor, L.J. and Knowles, G. 1988. 'Manual of Information to Accompany the SEC Corpus: The machine-readable corpus of spoken English.' Online. Accessed: January 2010. http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM

Step 1: Manual	
Reconcile orthography in SEC file with Aix	Amended version of SEC file
Step 2: Automatic	
Reconstitute enclitics in SEC; lower case all words	
Step3: Automatic	
Merge PoS from SEC with data from Aix, coping with asynchronous distribution of punctuation & pauses	File with LOB PoS tags subsumed in to Aix data
Step 4: Automatic	
Map set of C5 PoS tags in ProPOSEL to arrays of corresponding LOB tags, where one-to-many mappings predominate	
Step 5: Automatic & Manual	
Iterate through output file from Step 3, seeking a match between LOB tags in data file and live mapping from Step 4. A match triggers an event: insertion of C5 tag at designated index position in data file array. Implement a patch for instances of one-to-many mappings LOB<C5. Conduct manual inspection.	File with C5 as well as LOB PoS tags subsumed into Aix data, with one-to-one correspondence between taggings
Step 6: Automatic	
Create instance of ProPOSEL transformed into a Python dictionary with compound (word + C5) keys mapped to prosodic-syntactic value arrays. A match between dictionary keys and word + C5 pairings in output file from Step 5 triggers an event: designated prosodic-syntactic information from ProPOSEL is appended to dataset arrays. Re-run lookup seeking match between word tokens only for any untagged items.	Dataset subfiles for Section A of the corpus

Appendix: Stages in dataset build