

1 ProPOSEL Lexicon (LREC'08) & ProPOSEC (LREC'10)

- At LREC'08, the authors reported on ProPOSEL, a purpose-built Prosody and PoS English Lexicon of 104049 word forms compatible with the Python Natural Language Toolkit (Bird *et al.*, 2009).
- ProPOSEC is a new corpus research resource built using this lexicon, and intended for distribution with an updated version of the Aix-MARSEC dataset (Hirst *et al.*, 2009).
- ProPOSEC comprises multi-level parallel annotations, juxtaposing prosodic and syntactic information from different versions of the Spoken English Corpus (Taylor & Knowles, 1988), with canonical dictionary forms, in a query format optimized for Perl, Python, and text processing programs.
- As an experimental dataset, ProPOSEC can be used to study correlations between these annotation tiers, one application being to integrate significant findings as additional features in phrasing models for Text-to-Speech Synthesis and Automatic Speech Recognition.
- As a training set, ProPOSEC can be used for machine learning tasks in Information Retrieval and Speech Understanding systems.

2 Fields in ProPOSEC

1	Aix-MARSEC File ID	7	Syllable Count
2	Word	8	Lexical Stress Pattern
3	LOB PoS Tag	9	Default Content/Function Word Tag
4	C5 PoS Tag	10	DISC Stressed & Syllabified Phonetic Transcription
5	Aix SAMPA Phonetic Transcription	11	Alternative DISC Transcription, with Syllables Mapped to Stress Weightings
6	Canonical SAMPA Phonetic Transcription from ProPOSEL Lexicon	12	Prototype Phoneme-to-TSM Mapping (TSM: Tonic Stress Marks)

3 Example: Linguistic Annotations in ProPOSEC Text Files

Linguistic annotations in ProPOSEC for a prosodic-syntactic chunk initiated by a major clause boundary in the string:

“...soon after it took off from Athens airport...”

```
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', '@'],
['o', 'o', 'o', 'o']]
A0801|it|PP3|PNP|rt|lt|1|1|F|'lt|'lt:1|[['r', 'l', 't'], ['o', 'o', 'o']]
A0801|took|VBD|VVD|tUk|tUk|1|1|C|'tUk|'tUk:1|[['t', 'U', 'k'], [' ', ' ', ' ']]
A0801|off|RP|AVP|Qf|Of|1|1|C|'Qf|'Qf:1|[['Q', 'f'], ['o', 'o']]
A0801|from|IN|PRP|fr@m|frOm|1|1|F|'frQm|'frQm:1|[['f', 'r', '@', 'm'], ['o', 'o', 'o', 'o']]
A0801|athens|NP|NP0|{TInz|'&TInz|2|10|C|No value|No value|[['{', 'T', 'l', 'n', 'z'],
['*', 'o', 'o', 'o', 'o']]
A0801|airport|NN|NN1|e@pO:t|'e@pOt|2|10|C|'8-p$t|'8:1 p$t:0|[['e@', 'p', 'O:'],
['t'], [' ', 'o', 'o', 'o']]
A0801|PAUSE|,|,
```

This sentence snippet shows differences (in bold) between phonetic transcriptions in speech corpus and lexicon.

Speech corpus has **link-up** in the form of **'r-elisions'**, plus **vowel reduction** in **function words**.

But speech corpus only represents *one* phrasing variant...

4 Building Dataset Non-trivial: Several Stages Involved

- MANUAL:** Reconcile orthography in SEC file with Aix → *amended version of SEC file*
- AUTOMATIC:** Reconstitute enclitics in SEC & lower case all words
- AUTOMATIC:** Merge PoS from SEC with data from Aix, coping with asynchronous distribution of punctuation & pauses → *file with LOB PoS tags subsumed into Aix data*
- AUTOMATIC:** Map set of C5 PoS tags in ProPOSEL lexicon to arrays of corresponding LOB tags, where one-to-many mappings predominate
- AUTOMATIC & MANUAL:** Iterate through output from (3), seeking match between LOB tags in data file and live mapping in (4) to trigger event: insertion of C5 tag at designated index position in data file array. Implement patch for instances of one-to-many mappings LOB<C5. Conduct manual inspection.
- AUTOMATIC:** Create instance of ProPOSEL transformed into Python dictionary with compound (word + C5) keys mapped to prosodic-syntactic value arrays. Match between dictionary keys and (word + C5) pairings in output from (5) triggers event: append designated prosodic-syntactic information from lexicon to dataset arrays. Re-run lookup seeking match between word tokens only for any untagged items.

5 Experimentation: Significance Tests & Machine Learning

- We have empirical evidence of a significant correlation in English between 'gold-standard' phrase break annotations in the ProPOSEC dataset and words containing complex vowels (diphthongs and triphthongs) in their canonical dictionary pronunciations via the DISC phonetic transcription set in ProPOSEL and the chi-squared statistic (Brierley and Atwell, 2009; 2010).
- Multi-level parallel annotations in the ProPOSEC dataset facilitate statistical analyses of this kind. However, simply translating each field in ProPOSEC into a WEKA attribute as in the mock .arff file below would be insufficient as a training set for phrase break prediction.

@relation phraseBreak	1. One problem is the potential number of values for each attribute e.g. PoS trigram sequences; lexical stress patterns.
@attribute word {took, off, from, athens, airport}	
@attribute pos { AVP, NPO, NN1, PRP, VVD }	
@attribute lexicalStress { 1, 10, 01 }	2. Another problem is incorporating sufficient context into the language model: e.g. the researcher may be interested in a window of <i>N</i> words either side of a given index position.
@attribute break { yes, no }	
@data	3. It would require instead a series of complex transformations on the dataset to capture context and to summarise attribute-value pairs: e.g. applying a series of conditions which dictate whether or not a word carries a beat.
took, VVD, 1, no	
off, AVP, 1, no	
from, PRP, 1, no	
athens, NPO, 10, no	
airport, NN1, 10, yes	

6 References used in this poster

Auran, Bouzon & Hirst (2004) The Aix-MARSEC project: an evolutive database of spoken British English. In *Proc. SP-2004*. 561-564

Bird, Klein & Loper (2009) *Natural Language Processing with Python*. Sebastopol: O'Reilly Media Inc. 2009.

Brierley & Atwell (2008) ProPOSEL: A Prosody and POS English lexicon for Language Engineering. In *Proceedings of LREC'08: Language Resources and Evaluation Conference*, Marrakech, Morocco.

Brierley & Atwell (2009) Exploring Complex Vowels as Phrase Break Correlates in a Corpus of English Speech with ProPOSEL, a Prosody and PoS English Lexicon. In *Proceedings of INTERSPEECH'09*, Brighton.

Brierley & Atwell (2010) Complex Vowels as Boundary Correlates in a Multi-Speaker Corpus of Spontaneous English Speech. In *Proceedings of Speech Prosody 2010*, Chicago.

Hall, Frank, Holmes, Pfahringer, Reutemann & Witten (2009) The WEKA Data Mining Software: An Update. In *SIGKDD Explorations*. 11: 1.

Taylor & Knowles (1988) Manual of Information to Accompany the SEC Corpus: The machine readable corpus of spoken English. Accessed: January 2010. <http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM>