ProPOSEL: A Prosody and POS English Lexicon for Language Engineering

Claire Brierley, Eric Atwell

School of Computing, University of Leeds, LEEDS LS2 9JT, U.K.
E-mail: claireb@comp.leeds.ac.uk, eric@comp.leeds.ac.uk

Abstract

ProPOSEL is a prototype prosody and PoS (part-of-speech) English lexicon for Language Engineering, derived from the following language resources: the computer-usable dictionary CUVPlus, the CELEX-2 database, the Carnegie-Mellon Pronouncing Dictionary, and the BNC, LOB and Penn Treebank PoS-tagged corpora. The lexicon is designed for the target application of prosodic phrase break prediction but is also relevant to other machine learning and language engineering tasks. It supplements the existing record structure for wordform entries in CUVPlus with syntactic annotations from rival PoS-tagging schemes, mapped to fields for default closed and open-class word categories and for lexical stress patterns representing the rhythmic structure of wordforms and interpreted as potential new text-based features for automatic phrase break classifiers. The current version of the lexicon comes as a textfile of 104052 separate entries and is intended for distribution with the Natural Language ToolKit; it is therefore accompanied by supporting Python software for manipulating the data so that it can be used for Natural Language Processing (NLP) and corpus-based research in speech synthesis and speech recognition.

1. ProPOSEL: Derivation and Rationale

A pronunciation lexicon is an integral part of the front-end NLP module in a generic Text-to-Speech (TTS) synthesis system and constitutes a natural way of giving such a system both prosodic and syntactic insights into input text. For English, three such resources - originally developed for Automatic Speech Recognition (ASR) and listing words and their phonetic transcriptions - are widely used: CELEX-2 (Baayen et al, 1996); PRONLEX (Kingsbury et al, 1997); and CMU, the Carnegie-Mellon Pronouncing Dictionary (Carnegie-Mellon University, 1998). The latter is used in Edinburgh's Festival speech synthesis system (Black *et al*, 1999; Williams, 2008) and is included as one of the datasets in NLTK - the Natural Language ToolKit (Bird et al, 2007a). Similarly, lexicons or machine-readable dictionaries have been developed for TTS engines in other languages: for the German TTS system MARY (Schroder and Trouvain, 2003); for French (cf. Auberge, 1993; Thomas, 2003); for Norwegian (cf. Stensby et al, 1993; Heggtveit and Natvig, 2001). Recently, Nokia have used an extensive lexicon of 92,901 words and 68 PoS for Mandarin TTS (Tian et al, 2005); and large lexica with phonetic, prosodic and morpho-syntactic content have been generated for 13 languages, including US-English, as part of the LC-Star project (Hartikainen *et al*, 2003).

The starting point for our new prosody and PoS lexicon is CUVPlus (Pedler, 2002). This is a computer-usable dictionary of wordforms, derived from CUV2 (Mitton, 1992) and the Oxford Advanced Learner's Dictionary of Current English (Hornby, 1974), which identifies wordclass for each entry via C5 PoS tags, the syntactic annotation scheme used in the BNC or British National Corpus (Burnard, 2000). LC-Star and associated publications (cf. Hartinkainen *et al*, 2003; Vriend *et al*, 2003; Conejero *et al*, 2003) highlight a shortage of language resources that meet the needs of ASR, TTS and speech-to-speech translation applications which depend on wide coverage lexica with detailed morpho-syntactic information. The incorporation of C5 PoS-tags in CUVPlus provides this kind of detail and distinguishes this lexicon from other paper-based and electronic English dictionaries, including CELEX, PRONLEX and CMU; it also facilitates linkage with machine-readable corpora like the BNC. However, CUVPlus entries compact PoS variants for a given wordform into one field; ProPOSEL introduces one-to-one mappings of wordform to wordclass to facilitate their use as compound keys when the lexicon is transformed into a Python dictionary or associative array.

Phonological data in ProPOSEL has been generated from CELEX-2 and CMU. An analysis of prosodic and syntactic information in all three sources - CUVPlus, CELEX-2 and CMU - plus a full account of lexicon build is planned for a subsequent paper. Our lexicon was originally created to assemble information relevant to prosody in one language resource customised for language engineering tasks which involve the prosodic-syntactic chunking of text; and we want to make this resource freely available to other researchers.

2. Fields in the Prosody-PoS English Lexicon

The prototype prosody lexicon comes as a textfile of 104052 separate entries, each comprising 14 pipe-separated fields arranged as follows:
(1) wordform; (2) C5 tag; (3) capitalisation flag; (4) SAM-PA phonetic transcription; (5) CUV2 tag and frequency rating; (6) C5 tag and BNC frequency rating; (7) syllable count; (8) lexical stress pattern; (9) Penn Treebank tag; (10) default content or function word tag; (11) LOB tag; (12) C7 tag; (13) IPA syllabified phonetic transcription; (14) stressed and unstressed values mapped to syllable transcriptions.

```
sunniest|AJS|0|'sVnIIst|Os%|AJS:0|3|100|JJS|C
|JJT|JJT|'sV-nI-Ist|'sV:1 nI:0 Ist:0
```
Table 1: Example entry from ProPOSEL textfile

One field of particular interest to our research into automatic phrase break prediction is lexical stress pattern, where the rhythmic structure of wordforms is represented symbolically as a string of numbers: thus the pattern for the wordform `,objec'tivity` - with secondary stress on the first syllable and primary stress on the third syllable - is `20100`. For some homographs, this lexical stress pattern can fluctuate depending on part-of-speech category and meaning. The wordform `present` is a case in point, as demonstrated by fields 1, 2, 4, 7, 8 and 10 for all its entries in ProPOSEL:

```
present | AJ0 | 'preznt | 2 | 10 | C |
present | NN1 | 'preznt | 2 | 10 | C |
present | VVI | prI'zent | 2 | 01 | C |
present | VVB | prI'zent | 2 | 01 | C |
```
Table 2: Rhythmic structure for the homograph `present` is inverted when it functions as a verb

3. Prosodic Phrase Break Prediction

As previously stated, the purpose of this work is to integrate information from different dictionaries into one lexicon, customised for language engineering tasks which involve the prosodic-syntactic chunking of text. One such task is automated phrase break prediction: the classification of junctures (whitespaces) between words in the input text as either breaks (the minority class) or non-breaks (Brierley and Atwell, 2007a,b,c). The machine learner is trained on the annotated speech corpus, processed as a list of tokenised PoS tags including punctuation and boundary tags. The latter are represented by pipe symbols: /|/ for minor tone unit boundary; /||/ for pause (Roach, 2000).

Phrase break classifiers have been trained on additional text-based features besides PoS tags. The CFP status of a token - is it a *content* word (e.g. nouns or adjectives) or *function* word (e.g. prepositions or articles) or *punctuation* mark? - has proved to be a very effective attribute in both deterministic and probabilistic models (Liberman and Church, 1992; Busser *et al*, 2001) and therefore, a default content-word/function-word tag is assigned to each entry in the prosody-PoS lexicon in field (10). It is anticipated that further research will suggest modifications to this default status when the CFP attribute interacts with other text-based features.

Syllable counts - field (7) in our lexicon - have already been used in phrase break models for English (Atterer and Klein, 2002). This rather assumes uniformity in terms of duration of syllables whereas we know that in connected speech, an indefinite number of unstressed syllables are packed into the gap between one *stress pulse* (Mortimer, 1985) and another, English being a *stress-timed* language. A lexical stress pattern for each entry has therefore been included in ProPOSEL - fields (8) and (14) - because of its potential as a classificatory feature in the machine learning task of phrase break prediction. This intimation is further supported by the presence of rhythmic annotation tiers in the Aix-MARSEC corpus project (Auran *et al*, 2004), with its focus on speech synthesis applications and the theoretical modelling (acoustic, phonetic and phonological) of intonation and speech prosody.

4. Manipulating Data in the Lexicon: Python Dictionaries

The Python programming language has a dictionary mapping object with entries in the form of (key, value) pairs. Each key must be unique and immutable (e.g. a string or tuple), while the values can be any type (e.g. a list). This syntax can be exploited by transforming the prosody lexicon into a Python dictionary, where the lookup keys are (wordform, C5 tag) tuples and the corresponding values are lists of tokens representing selected information from the remaining fields for a given entry. Thus, using a sample of 4 entries to represent our lexicon and version 0.8 of NLTK, we can use the code in Listing 1 below to transform the lexicon into Python dictionary format.

```
from nltk.book import * # import statement for NLTK version 0.9 would be: import nltk, re, pprint
lexicon = """
cascaded|VVD|0|k&'skeIdId|Ic%,Id%|VVD:1|3|010|VBD|C|VVD|VBD
cascaded|VVN|0|k&'skeIdId|Ic%,Id%|VVN:0|3|010|VBN|C|VVN,VVNK|VBN
cascading|VVG|0|k&'skeIdIN|Ib%|VVG:1|3|010|VBG|C|VVG,VVGK|VBG
cascading|AJ0|0|k&'skeIdIN|Ib%|AJ0:0|3|010|JJ|C|JJ,JK|JJ,JJB,JNP
"""
lexicon = [line.split('|') for line in list(tokenize.line(lexicon))]
lexKeys = [(index[0], index[1]) for index in lexicon]
lexValues = [[index[6], index[7], index[9]] for index in lexicon]
proPOSEL = dict(zip(lexKeys, lexValues))
```

Listing 1: Code snippet using Python list comprehensions and built-ins to transform the prosody-PoS English Lexicon into an associative array

The Python dictionary method for displaying a list of (key, value) pairs returns an as yet unsorted dictionary; nevertheless, listing 2 below demonstrates how multiple values representing a series of linguistic observations on syllable count, lexical stress pattern and content/function word status have now been mapped to compound keys (cf. Bird et al, 2007b, chapter 6; Martelli et al, 2005 pp. 173-5).

```
proPOSEL.items()
# calls built-in method which returns a list of
key-value pairs
[(('cascaded', 'VVN'), ['3', '010', 'C']),
(('cascading', 'VVG'), ['3', '010', 'C']),
(('cascaded', 'VVD'), ['3', '010', 'C']),
(('cascading', 'AJ0'), ['3', '010', 'C'])]
```

Listing 2: Each individual entry tuple is a collection of objects with different linguistic interpretations

Incoming corpus text - also in the form of (token, tag) tuples - can now be matched against dictionary keys; and thus intersection enables corpus text to accumulate additional values which have the potential to become features for machine learning tasks. There is one caveat, however. Listing 2 identifies an instance where the *-ing* form of a verb (the present participle) is sometimes tagged as an adjective, whereas the *-ed* form (the past participle) is not. This is just one example of a general problem: a corpus may include syntactic variants as yet unrecorded in ProPOSEL; and if so, these will not be matched because the lookup keys and the syntactic values generated from them - the Penn, LOB and C7 fields - can only represent the variance in CUVPlus and its parent corpus, the BNC.

5. Manipulating Data in the Lexicon: Managing Different Tagsets

The aforementioned lookup mechanism is relatively straightforward for corpora tagged with C5, the basic tagset used in the BNC. For corpora tagged with alternative schemes, incoming tokens and tags will first need to be matched against wordforms and the corresponding tagset fields in the lexicon. Different tagsets (Penn, C7 and LOB) were mapped to C5 as part of the lexicon build; we are still experimenting with these mappings and it is anticipated that user feedback will also be important in fine-tuning them. However, the lexicon is supported by a range of Python software compatible with NLTK to facilitate the cross-referencing of linguistic data from the lexicon's record structure to corpus text (cf. Bird *et al*, 2007b, chapter 13).

It is possible, nominally, to map between C5 and different PoS-tagging schemes in ProPOSEL via a one-step process. In the following line of code, C5 tags are mapped to LOB:

```
mapTags = list(set([(line[1], line[10]) for line
in lexicon]))
```

Listing 3: Code snippet maps the set of all C5 PoS tags in the prosody-PoS English Lexicon to equivalent symbolic values in LOB

However, the resulting *mapTags* object uncovers a new set of problems. The C5 tagset comprises 62 part-of-speech tags, including 4 tags for punctuation; but the set of C5 tags in the lexicon includes combinations for enclitics and possessive forms like *"I'll"* <PNP+VMO> and *"Lloyd's"* <NPO+POS> and has 95 items. The *mapTags* object also reveals 39 instances (around 41%) of one-to-many mappings - mostly, but not entirely, in the direction C5 > LOB. The challenges of converting between different tagsets have been extensively documented (Atwell *et al*, 1994; Atwell *et al*, 2000; Atwell, 2007). One-to-many mappings uncover 'indelicate' areas of each tagset and syntactic information is lost both ways even when the tagsets favour fine-grained linguistic distinctions.

ProPOSEL is supported by a toolkit of software solutions and an explanatory tutorial to help surmount such problems, with sections on: preparing the textfile for NLP; mapping variant syntactic information (with subsidiary sections on enclitics, Saxon genitives and one-to-many mappings); using the lexicon as a prosodic annotation tool; and implementing ProPOSEL as a Python dictionary. In the following code snippet, the Python itertools() module is used to loop through two parallel iterables: *match* - a list of token, C5 tuples; and *corpusText* - a list of lists comprising the original token, LOB tag pairings from the corpus plus an equivalent C5 tag generated from the lexicon. For each item, a successful lookup via the dictionary object *proPOSEL* in turn generates a deeply nested sequence object holding orthographic form mapped to both LOB and C5, plus further annotations from whichever additional fields have been selected.

```
for x, y in itertools.izip(match, corpusText):
    if x in proPOSEL.keys():
    # if tuple format matches dictionary keys
        y.append(proPOSEL[x])
        # append corresponding values for selected
        fields
    else:
        y.append('No match')
```

Listing 4: Code snippet illustrating one solution for automatic dictionary lookup

Outputs from this lookup process after formatting functions have been applied are shown in Listing 5 below. These illustrate problems with function words mostly, which can only be addressed through further research:

- What are the best default CFP settings for phrase break prediction in field (10)?
- What is the CFP status of an enclitic?
- Under what circumstances do function words carry a beat? In LC-Star lexica, primary stress is marked on all items including function words.
- The lexical stress pattern for `necessarily` would not be everyone's choice here; nevertheless, it *is* the principal pronunciation form in CELEX2: P\'nE-s@-s@-r@-lI\

```
Wordform:           to
PoS tag:            TO
syllable count:     1
stress pattern:     1
CFP tag:            F
stress distribution: 'tu:1

wordform:           attribute
PoS tag :           VB
syllable count :    3
stress pattern:     010
CFP tag:            C
stress distribution: '{:0 trI:1 bjut:0

wordform:           to
PoS tag:            IN
syllable count:     1
stress pattern:     1
CFP tag:            F
stress distribution: 'tu:1

wordform:           them
PoS tag :           PP3OS
syllable count :    1
stress pattern:     1
CFP tag:            F
stress distribution: 'Dem:1

wordform:           roles
PoS tag:            NNS
syllable count:     1
stress pattern:     1
CFP tag:            C
stress distribution: 'r5lz:1

wordform:           which
PoS tag:            WP
syllable count:     1
stress pattern:     1
CFP tag:            F
stress distribution: 'Wij:1

wordform:           aren't
PoS tag:            BER+XNOT
syllable count:     1
stress pattern:     1
CFP tag:            CF
stress distribution: No value

wordform:           necessarily
PoS tag:            RB
syllable count:     5
stress pattern:     10000
CFP tag:            C
stress distribution: 'Ne:1 s@:0 s@:0 r@:0 Li:0

wordform:           theirs
PoS tag :           PP$$
syllable count :    1
stress pattern:     1
CFP tag:            F
stress distribution: 'D8z:1
```

Listing 5: LOB-tagged corpus text has accumulated new prosodic values via automated lookup from ProPOSEL

6. Conclusions

This paper describes a new combined prosody and PoS-tag lexicon for corpus-based research and language engineering in English. The lexicon builds on established language resources and maps wordform entries to a range of attributes which have proved, or may prove, significant for machine learning and linguistic analysis of prosody: the open or closed-class status of words, for example, and their symbolic rhythmic structure. The paper argues the case for word class identification via PoS tags in computer-usable dictionaries. Syntax is an important intermediary in TTS systems between input text and synthesized speech output (Loquendo, 2004). The prototype prosody lexicon already holds four variant PoS-tagging schemes widely used in English speech corpora. Finally, it is planned to make this lexicon freely available to other speech and language researchers - under the auspices of the open source Natural Language Toolkit - and therefore it is supported by Python software and tutorial documentation.

7. References

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

Atwell, E., Hughes, J., and Souter, C. (1994). AMALGAM: Automatic Mapping Among Lexico-Grammatical Annotation Models. In Judith Klavens and Philip Resnik (eds.) *The Balancing Act - Combining Symbolic and Statistical Approaches to Language. Proceedings of the Workshop in conjunction with the 32nd Annual Meeting of the Association of Computational Linguistics.* New Mexico State University, Las Cruces, New Mexico, USA.

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

Atwell, E. (2007) Development of tag sets for part-of-speech tagging. In Anke Ludeling & Merja Kyto (eds.) *Corpus Linguistics: An International Handbook* Mouton de Gruyter. 2007.

Auberge, V. (1993). Prosody Modelling with a Dynamic Lexicon of Intonative Forms: Application for Text-to-Speech Synthesis *ESCA Workshop on Prosody* Lund, Sweden Sept 27-29, 1993.

Auran, C., C. Bouzon and D. Hirst (2004) The aix-MARSEC project: an evolutive database of spoken british English. In SP-2004, 561-564.

Baayen, R. H., R. Piepenbrock and L. Gulikers (1996). CELEX2 Linguistic Data Consortium, Philadelphia.

Bird, S., E. Loper and E. Klein (2007a) NLTK-lite 0.8 beta [June 2007] Available online from: http://nltk.sourceforge.net/index.php/Main_Page (accessed: 21/06/07).

Bird, S., E. Klein and E. Loper (2007b). "Natural Language Processing" Available online from: http://nltk.sourceforge.net/index.php/Book (accessed: 21/09/07).

Black A.W., P. Taylor and R. Caley. *The Festival Speech Synthesis System: System Documentation Festival*

version 1.4 (1999) Available online from: http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html (Accessed: 07/03/08)

Brierley, Claire; Atwell, Eric. (2007a) Prosodic phrase break prediction: problems in the evaluation of models against a gold standard. *Traitement Automatique des Langues,* vol. 48.1.

Brierley, Claire; Atwell, Eric. (2007b) Corpus-based evaluation of prosodic phrase break prediction. In: *Proceedings of Corpus Linguistics 2007*.

Brierley, Claire; Atwell, Eric. (2007c) An approach for detecting prosodic phrase boundaries in spoken English. *ACM Crossroads journal*, vol. 14.1.

Burnard, L. (ed.) (2000). Reference Guide for the British National Corpus (World Edition) Available online from: http://www.natcorp.ox.ac.uk/docs/userManual/ (accessed: 20/05/07).

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

Carnegie-Mellon University (1998). The CMU Pronouncing Dictionary (v. 0.6) Available online from: http://www.speech.cs.cmu.edu/cgi-bin/cmudict (accessed: 21/06/07).

Conjero, D., J. Giménez, V. Arranz and A. Bonafonte (2003) Lexica and Corpora for Speech-to-Speech Translation: A Trilingual Approach. In *Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH-2003*. September, 1-4, 2003. Geneva, Switzerland.

Hartinkainen, E., G. Maltese, A. Moreno, S. Shammass and U. Ziegenhain (2003) Large Lexica for Speech-to-Speech Translation: From Specification to Creation. In *EUROSPEECH-2003*, 1529-1532.

Heggtveit, P.O., and J.E. Natvig (2001). Intonation modelling with a lexicon of natural F0 contours In *EUROSPEECH-2001*, 1163-1166.

Hornby, A.S. (1974) *Oxford Advanced Learner's Dictionary of Current English* (third edition) Oxford: Oxford University Press

Kingsbury, P., S. Strassel, C. McLemore, and R. MacIntyre (1997). CALLHOME American English Lexicon (PRONLEX) Linguistic Data Consortium, Philadelphia

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

Loquendo (2004). SSML 1.0: an XML-based language aimed to improve TTS rendering Available online from: http://www.loquendo.com/en/news/whitepaper_SSML.htm (accessed: 12/10/07)

Martelli, A., A. Martelli Ravenscroft and D. Ascher (2005) *Python Cookbook* (second edition) Sebastopol: O'Reilly Media, Inc.

Mitton, R. (1992) A description of a computer-usable dictionary file based on the Oxford Advanced Learner's Dictionary of Current English. Available online (and accessed 07/03/08) from: http://coral.lili.uni-bielefeld.de/Classes/Winter98/ExPhon/SynthDBNotes/synthdbnotes/node18.html

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

Pedler, J. (2002) CUVPlus [Electronic Resource] Oxford Text Archive Available online from: http://ota.ahds.ac.uk/textinfo/2469.html (accessed: 21/06/07)

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

Schröder, M. and J. Trouvain (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6 pp. 365-377. Available online from: http://mary.dfki.de/ (accessed: 12/10/07)

Stensby, S., B. Horvei and G.E. Ottesen (1993) Lexicon and prosodic structure in a text-to-speech system *ESCA Workshop on Prosody* Lund, Sweden Sept 27-29, 1993

Thomas, C. (2003) Automatic Generation of French Speech in *Crossroads ACM Student Magazine* 9.3 Spring 2003 Available online from: http://oldwww.acm.org/crossroads/xrds9-3/french.html (accessed: 12/10/07)

Tian, J., J. Nurminen, I. Kiss (2005) Duration Modeling and Memory Optimization in a Mandarin TTS System, submitted to *Interspeech 2005*, Lisbon, Portugal, Sept 4-8, 2005

UCREL (2007) University of Lancaster Centre for Computer Corpus Research on Language: "UCREL CLAWS5 Tagset" Available online from: http://www.comp.lancs.ac.uk/ucrel/claws5tags.html Accessed: 21/09/07

Vriend, F. de., N. Castell, J. Giménez and G. Maltese (2004) LC-STAR: XML-coded Phonetic Lexica and Bilingual Corpora for Speech-to-Speech Translation Available online from: http://www.lc-star.com/LC-STAR_papillon_2004.PDF (Accessed: 07/03/08)

Williams, B. (2008) Re: [Corpora-List] transcribe English text. Email post to *Corpora* discussion list. Available online from: http://www.uib.no/mailman/public/corpora/2008-March/006135.html (Accessed: 07/03/08)