White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/81893/

Version: Published Version

Proceedings Paper:
Atwell, ES (1989) A lexical database for English learners and users: the Oxford advanced learner's dictionary. In: McCrank, L, (ed.) Databases in the Humanities and Social Sciences 4: Proceedings of the International Conference on Databases in the Humanities and Social Sciences. The International Conference on Databases in the Humanities and Social Sciences, July 1987, Auburn University at Montgomery, Alabama, USA. Learned Information, 21 - 33. ISBN: 0938734377.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

A Lexical Database for English Learners and Users: The Oxford Advanced Learner's Dictionary

Eric Steven Atwell
Artificial Intelligence Group
Department of Computer Studies
Leeds University, Leeds, Yorkshire, U.K. L52 9JT1

Presented by Margaret Cooper
Edinburgh University
Edinburgh, Scotland EH1 1HN

Introduction: Dictionaries as Databases

A common approach in database primers such as (Date 1981; Byers 1986) is to draw upon an analogy between computer databases and more familiar "paper databases". The monographic dictionary is a usual example of a paper database often used without being aware of any computational complexity. However, upon closer analysis, the structure of records or entries in an English dictionary is far more complex than that of records in other favorite illustrations such as telephone directories. While it is true that some entries contain just a small number of 'standard' fields, most entries have a much more complex implicit structure. The standard fields in a typical dictionary are generally *Headword, Pronunciation, Part-of-Speech, Definition*. For example; the entry

"abbot /'abEt/ n man (Father Superior) at the head of the monks in an abbey or monastery"
consists of:

Headword: abbot

Pronunciation: /'abEt/

Part-of-Speech: n (i.e. noun)

Definition: man (Father Superior) at the head of the monks in an abbey or monastery.

However, many entries, usually the great majority, exhibit a far more complex structure, with many additional optional fields and/or subfields. Appendix 1 shows a typical sequence of entries from one dictionary, the *Oxford Advanced Learner's Dictionary of Contemporary English (OALD)*. Readers who are native speakers of English will have little difficulty in interpreting and understanding such dictionary entries, but they are probably unaware that many entries have complex underlying structural analyses.

Language Learning and Teaching

Dictionary entries are surprisingly complex, and this is particularly true of language learner's dictionaries such as the *OALD*. Because their dictionaries are designed for use by learners of English, less can be taken for granted about a reader's intuitive knowledge of the language. Generally, it must have rather more explicit, detailed information than is found in most native speaker's dictionaries. For example, most dictionaries give broad part-of-speech categories for words, and assume that more detailed grammatical restrictions are intuitively known by the reader. However, these detailed grammatical restrictions must be made explicit for the language learner; so the *OALD* in addition gives codes stating explicitly in what sentence structures a given verb may occur.

Unfortunately, given the complex structure of the dictionary, the learner (and perhaps even the teacher) may have difficulty finding in the printed dictionary the exact sort of information

they are seeking at any given time. To simplify the search process, a computerized database version of the dictionary may be useful in English Language teaching and research. Standard database techniques allow access to many sorts of information that cannot be elicited readily from the printed dictionary; for example, lists of taboo words, verbs occurring in unusual sentence-patterns, words which can be both adjectives and adverbs, etc.. Appendix 2 illustrates this enhanced flexibility of a computerized database over the printed dictionary. This is a list of all the words marked as taboo in the *OALD* (not including words like wank, snot, which are only considered vulgar, not taboo). Interestingly, note that the dictionary allows for the occasional need to use such words in typed documents, and gives break points for end-of-line hyphenation for longer taboo words! Such a list of taboo words in the *OALD* could only have been gleaned from the paper version by hours of painstaking search, but this was extracted easily from our database version.

Natural Language Processing Systems

Natural Language front-ends would make databases (and computer systems generally) far more accessible to scholars in the Humanities and Social Sciences. More generally, Natural Language Processing systems have very wide potential applications; see Johnson (1985) for a detailed analysis of the commercial applications. Like language learners, computer systems for Natural Language Processing have no intuitive knowledge of the details of grammatical restrictions to rely on, so they would find the detailed information in the *OALD* very useful. However, the information in the paper dictionary needs to be available in a readily-accessible format, as a structured database. To date, many researchers developing English interfaces to computer systems have restricted their systems to a small vocabulary and grammar; however, as systems become more sophisticated and ambitious, larger dictionaries are required. For example, the dictionary used by the CLAWS system (Leech, *et al.*, 1983a, b; Atwell, 1983; Atwell, *et al.*, 1984) contained over 7000 entries, and this was considered to be a very large dictionary compared to those used by other NLP systems. Nevertheless, the system still encountered many unknown words in its input, and had to rely on default backup mechanisms to cope with these words, leading to increased errors in analysis. A significant amount of manual editing and correction of the output of the system could have been circumvented if fuller dictionary databases (such as an *OALD* database) had been available (Atwell, 1981, 1982).

Some attempts have been made to build NLP systems which 'learn' lexical information automatically, thus avoiding the need to use an existing dictionary database (for example: Berwick, 1985; Atwell, 1986a, b; forthcoming a, b; Atwell & Drakos, 1986). However, these experiments have only met limited success to date. Furthermore, although these experiments may be theoretically interesting in their own right, it is questionable whether in general NLP researchers should deliberately ignore the vast amount of lexicographical knowledge and expertise in conventional dictionaries. Several leading-edge researchers are actively investigating how to use computerized machine-readable dictionaries in their Natural Language Processing software. Dictionaries under consideration include *Webster's English Dictionary* (Marcus, 1986), the *Longman Dictionary of Contemporary English (LDOCE)* (Alshawi, et al., 1985), the *Collins English Dictionary* (Sharman, 1986) and the *Houghton Mifflin American Heritage Dictionary* (Kaplan, 1986), as well as the *OALD*. However, of these, only the *LDOCE* and *OALD* have the higher level of explicit detail in grammatical and other information required both by English language learners and sophisticated English processing programs.

Parsing the *OALD*

A typesetting tape containing the text of the *OALD* in machine-readable form was the starting point in building a structured database. Although a computer program could in theory process the raw tape, such a program would have to be very sophisticated to determine entry structures 'on the fly'. Instead, it seemed more sensible to transform the tapefile into a more

readily searchable format, and then use the parsed file with a much simpler and faster search mechanism. Every entry in the dictionary needed to be parsed, so that the internal structure of entries could be modeled correctly in the database.

Whitton (1985) reports an initial attempt to parse the *OALD*, using a parser 'hand-coded' in Pascal. Unfortunately, this turned out to be very cumbersome: we soon found cases where the original prototype parser was unable to cope properly with certain entry structures, and these problems could only be handled by messy modifications to the low-level Pascal code. By the end of this initial attempt, it was clear that further improvements in parsing by further modifications to the internal workings of the Pascal.

YACC

A compiler-to-compiler tool would allow the task of writing the grammar to be separated out. The tool used was YACC, a program which converts a phrase-structure-like grammar (with optional embellishments) into a C parsing program. The YACC grammar devised for the *OALD* ran to nearly 500 lines and included 37 terminal symbols (corresponding to subfields in the database), and this still left some incorrect analyses. YACC grammar rules may include arbitrary C-code enclosed in curly brackets to set flags, print messages, etc.; this facility allowed us to cope to some extent with non-context-free parts of the syntactic structure of entries. Appendices 3, 4, and 5 illustrate the YACC grammar used. Number 3 shows some of the rules in the YACC grammar, defining the syntactic structure of dictionary entries. Number 4 is a list of non-terminal symbols in the grammar, equivalent to the higher-level structural units in the parsed dictionary. Number 5 lists the terminals in the grammar, both as single character codes and their fuller explanatory forms (see below). This grammar was used to parse the whole *OALD* file; the grammar grew incrementally as new entry structures were found. The file went through several processing stages:

1. Preprocessed:

   A number of comparatively minor alterations were made to the original file prior to input to the parser. For example, non-printing characters were converted to numeric codes, and occasional parts of entries which the grammar could not cope with were put in comment brackets which YACC knew to ignore; part of the preprocessed file is shown in Example 6.

2. Parsed:

   This version is the output from the parser: each line starts with a single character code denoting the grammar terminal symbol (i.e., record field or subfield). The file at this stage is illustrated in Appendix 7. The first character on each line is the field code, and as this is a key for consumption by software rather than humans, these letters are not intended to have any mnemonic value. (See possible single-letter codes, and their meanings in Appendix 5).

3. Parsed-Expanded:

   The single-character code is expanded for human readability (according to the table in Appendix 5); this version of the file is illustrated in Appendix 8. The expanded field codes take up much more disk space, but they have the same information content as single-letter codes as far as software is concerned, so we only created expanded versions temporarily for checking purposes. The final, parsed version of the dictionary is kept in unexpanded form, since we have a program to recreate easily the full field codes.

4. Problems:

This is an error file output by YACC separately from the parsed text; it notes comments ignored, and other potential problems identified during the parse for the human editor's intervention. Appendix 9 shows some the problems caught during parseing.

Availability

The parsed file exceeds six megabytes. All work to date has been done on VAXes, but many potential users would prefer a version on a personal computer. The full OALD database would require a hard disk on an IBM-PC (or compatible). We are considering alternative storage media; for example, the relatively new IBM PS/2 is rapidly taking over the IBM-PC as an international standard in personal computing, so the 200Mb optical discs available with the PS/2 could constitute a suitable widely accessible medium for distribution. The dictionary may have a wider market, including word processor users, if it were packaged with other word processing tools, such as a spelling checker, for example (Borland, 1985), or a grammar-checker (Atwell, 1983; 1986c; 1986d); but this will increase storage requirements and may detract from portability. It may be necessary to wait (hopefully not too long) for storage technology to evolve to our requirements before the OALD dictionary database becomes viable for widespread distribution and use.2

Notes

[1](EARN/BITNET: eric%uk.ac.leads.ai@rl.earn)

[2]I wish to acknowledge the contribution of Reginald Whitton to this research: his original undergraduate project to parse the OALD with a Pascal program (Whitton, 1985) was not very successful, but pointed the way forward. Whitton continued work on the project as a programmer for a few weeks after graduating; I thank my Head of Department, Denis Hutchinson, for allocating funds to this research!

References

Alshawi, H., Boguraev, B. & Briscoe, T. (1985). Towards a lexicon support environment for real time parsing. *Proceedings of the Second Conference of the European Chapter of the Association for Computational Linguistics.* Geneva.

Atwell, E. S. (1981). *LOB Corpus Tagging Project: Manual Preedit Handbook*. Lancaster: University of Lancaster, Departments of Computer Studies and Linguistics.

_____(1982). *LOB Corpus Tagging Project: Manual Postedit Handbook (A mini-grammar of LOB Corpus English, examining the types of error commonly made during automatic [computational] analysis of ordinary written English)*. Lancaster: University of Lancaster, Departmentsof Computer Studies and Linguistics.

_____(1983). Constituent-Likelihood Grammar. *Newsletter of the International Computer Archive of Modern English (ICAME NEWS), 7,* 34-67. Bergen: University, Norwegian Computing Centre for the Humanities.

_____(1986a). *Extracting a natural language grammar from raw text* (Report No. 208). Leeds: University of Leeds, Department of Computer Studies Research.

_____(1986b). A parsing expert system which learns from corpus analysis. In W. Meijs (Ed.).*Corpus Linguistics and Beyond: Proceedings of the Seventh International Conference*

on English Language Research on Computerized Corpora. Amsterdam, The Netherlands: Rodopi.

_____(1986c). How to detect grammatical errors in a text without parsing it (Research Rep. No. 212). University of Leeds, Department of Computer Studies, In *Proceedings of the Association for Computational Linguistics Third European Chapter Conference, Copenhagen, Denmark*. (forthcoming).

_____(1986d). Beyond the micro: Advanced software for research and teaching from computer science and artificial intelligence. In G. Leech, & C. Candlin (Eds.), *Computers in English language teaching and research: selected papers from the British Council Symposium on Computers in English Language Education and Research*, (pp.167-183). Lancaster, England: White Plains, NY: Longman.

_____(forthcoming a). Transforming a Parsed Corpus into a Corpus Parser, to appear in *Proceedings of the 1987 ICAME 8th International Conference on English Language Research on Computerized Corpora*. Helsinki, Finland.

_____(forthcoming b). An expert system for the automatic discovery of particles. In *Proceedings of the 1987 International Conference on the Study of Particles, Berlin, East Germany.*

Atwell, E. S., Leech, G. & Garside, R. (1984). Analysis of the LOB corpus: Progress and prospects. In J. Aarts & W. Meijs (Eds.), *Corpus Linguistics; Proceedings of the ICAME Conference on the use of Computer Corpora in English Language Research*, Nijmegen, The Netherlands: Rodopi.

Atwell, E. S. & Drakos, N. F. (1987) Pattern Recognition Applied to the Acquisition of a Grammatical Classification System from Unrestricted English Text to appear in *Proceedings of the Association for Computational Linguistics Third European Chapter Conference, Copenhagen, Denmark.*

Berwick, R (1985). *The acquisition of syntactic knowledge*. Cambridge, MA and London: MIT Press.

Borland International Inc. (1985). *Turbo Lightning: Owner's Handbook*. Scotts Valley, CA: Borland International.

Byers, R. (1986). *Everyman's Database Primer*. Torrance, CA: Ashton-Tate.

Date, C. J. (1981). *An introduction to database systems* (3rd ed.). Reading, MA: Addison Wesley.

Leech, G., Garside, R. & Atwell, E. S. (1983a). Recent developments in the use of computer corpora in English language research. *Transactions of the Philological Society*, 23-40.

Leech, G., Garside, R. & Atwell, E. S. (1983b). The automaticgrammatical tagging of the LOB corpus. *Newsletter of the International Computer Archive of Modern English (ICAME NEWS), 7*, 13-33. Bergen University: Norwegian Computing Centre for the Humanities.

Hornby, A. S., Cowie, A. P. & Gimson, A. C. (Eds.) (1974).*Oxford Advanced Learner's Dictionary of Current English* (3rd ed.). Oxford: Oxford University Press.

Johnson, T. (1985). *Natural Language Computing: the commercial applications*. Ovum, London.

Kaplan, R. (1986). Contributions to *Discussion session on the Lexicon* in Whitelock, *et al.* (1986), 146-168.

Marcus, M. P. (1986). Contributions to *Discussion session on the Lexicon* in Whitelock, *et al.* (1986), 146-168.

Procter, P. (editor-in-chief) (1978). *Longman Dictionary of Contemporary English*. White Plains, NY: Longman.

Sharman, R. (1986). Contributions to *Discussion session on the Lexicon* in Whitelock, *et al.* (1986), 146-168.

Whitelock, P., Somers, H., Bennett, P., Johnson, R., & Wood, M. M. (Eds.) (1986). *Alvey/ICL workshop on linguistic theory and computer applications: transcripts of presentations and discussions* (CCL/UMIST Report No. 86/2). University of Manchester Institute of Science and Technology, Centre for Computational Linguistics.

Whitton, R. T. J. (1985). *The production of a database of words used in contemporary English with information on the use of each*. Undergraduate Project Report,, Leeds University, Department of Computer Studies.

Appendices

Appendix 1: A typical sequence of entries from a dictionary (the *OALD*)

aback /E'bak/ adv backwards. be *taken a'back, be startled, disconcerted.<

abacus /'abEkEs/ (pl -cuses /-kEsIz/ or -ci /'abEsaI/) n frame with beads or balls sliding on rods, for teaching numbers to children, or (still in the East) for calculating; early form of digital computer.<

abaft /E'bqft US: E'baft/ adv, prep (naut) at, in, toward, the stern half of a ship; nearer the stern than; behind.<

abandon 1 /E'bandEn/ vt 1. UVP6Ae go away from, not intending to return to; forsake: The order was given to @ ship, for all on board to leave the (sinking) ship. The cruel man @ed his wife and child. 2. UVP6Ae give up: They @ed the attempt, stopped trying. They had @ed all hope, no longer had any hope. The new engine design had to be @ed for lack of financial support. 3. UVP14e @ oneself to, give oneself up completely to, eg passions, impulses: He @ed himself to despair. @ed part adj 1. given up to bad ways; depraved; profligate: You @ed wretch] 2. deserted; forsaken. @ment n UUe.<

abandon 2 /E'bandEn/ n UUe careless freedom, as when one gives way to impulses: waving their arms with @.<

abase /E'beIs/ vt UVP6Be @ oneself, humiliate or degrade oneself: @ oneself so far as to do sth, lower oneself in dignity to the extent of doing sth. @ment n UUe.<

abash /E'ba2/ vt UVP6Ae (passive only) cause to feel self-conscious or embarrassed: The poor man stood/felt @ed at this display of wealth, was confused, not knowing what to do or say.<

abate /E'beIt/ vt,vi 1. UVP6A,2Ae (liter) (of winds, storms, floods, pain, etc) make or become less: The ship sailed when the storm @d. 2. UVP6Ae (legal) bring to an end; abolish: We must @ the smoke nuisance in our big cities. @ment n UUe abating; decrease.<

abattoir /'abEtwq(r) US: *abE'twqr/ n slaughter-house (for cattle, sheep, etc).<

Appendix 2: Example search of the *OALD* database: taboo words

Abo arse ball balls bal.locks bas.tard bol.locks bug.ger bull.shit cock coolie coon crap cunt dago Darkey eff fag.got fairy fart fuck god.dam Jim Crow kaf.fir mammy nig.ger pansy piss pouf prick queen queer screw shit sod sod.ding stuff swine tit whore

Appendix 3: Sample rules from the YACC grammar used to parse the *OALD*

```
list_of_structural_elements
   : structural_element list_of_structural_elements
   | structural_element
   | /* empty_string */ ;

structural_element
   : square_box pieces
   | rectangular_box
   | cross_reference_section
   | idiom_section
   | bold_number pieces
   | derivative_section_section
   | bold_bracketed_letter_or_derivative
   | noun_type_section pieces
   | verb_pattern_section pieces
   | taboo
   | US_also_or_abbreviation_section
   | meanings_and_examples ;

headword_section
   : headword_start headword headword_end headword_superscript
     alternative_spelling_section pronounciation_section
     alternative_to_headword_section ;

alternative_spelling_section
   : comma space alternative_spelling alternative_spelling_section
   | space US_also_or_abbreviation_section space
   | space
   | /* empty_string */ ;

pronounciation_section
   : phonetic_font slash pronounciation other_pronounciations slash
   | /* empty_string */ ;

other_pronounciations
   : US_pronounciation_section
   | strong_form_pronounciation_section
   | comma list_of_pronounciations
     /* for where input has had to be modified */
   | /* empty string */ ;

list_of_pronounciations
   : pronounciation separater list_of_pronounciations | pronounciation ;

alternative_to_headword_section
   : comma space start_of_alternative_headword
     alternative_headword pronounciation_section
   | /* empty_string */ ;
```

Appendix 4: Non terminal symbols in the YACC grammar used to parse the *OALD*

A accent apostrophe arrow bold_italic_font box circumflex clear_font colon comma dollar_sign end_of_file eoe equals exclamation_mark font_6 full_stop grave headword_end headword_start hyphen italic_font label_end label_start large_bold_font left_bracket left_square_bracket one other_digits other_letters phonetic phonetic_2 phonetic_font plus question_mark quote rec_box right_bracket right_square_bracket roman_font semicolon slash small_bold_font space stress taboo_mark three_dots tilde times word_break

Appendix 5: List of single-character field codes, and their meanings

a alternative headword section
b change in part of speech
c change in part of speech for derivative
d definition
e doubling of consonants
f doubling of consonants (sometimes)
g doubling of consonants (but not in US)
h end of entry
i pronunciation unchanged
j spelling of conjugation or plural unchanged from headword
k start of pieces
l subentry
m subentry definition
n taboo symbol

A abbreviation
B abbreviation pronunciation
C also pronunciation
D also spelling
E alternative headword
F alternative spelling of headword
G comparative
H conjugation or plural spelling
I conjugation or plural label
J cross reference
K derivative
L headword
M headword superscript
N idiom
O nountype
P pronunciation
Q strong form pronunciation
R superlative
S text
T US pronunciation
U US Spelling
V verb pattern
W word class label

Appendix 6: Sample of the PREPROCESSED *OALD* file

330B327, b 316/bi/ 315(pl 313B's, b's 316/biz/315) 314the seco
nd letter of the English alphabet.304
330baa327 316/bq/ 315n 314cry of a sheep or lamb. 307 315vi (3
13baaing, baaed 314or 313baa'd 316/bqd/315) 314make this cry; blea
t. 310246@-lamb 315n 314child's word for a sheep or lamb.304
330baas327 316/bqs/ 315n 314(S Africa) boss.304
330babble327 316/246babl/ 315vi,vt 3101 312[VP2A,B,C] 314talk
in a way that is difficult to understand; make sounds like a baby; (of
streams, etc) murmur. 3102 312[VP6A,15B] 311@ (out), 314repeat fool
ishly; tell (a secret): 315@ (out) nonsense/secrets. 307 315n 312[U
] 3101 314childish or foolish talk; confused talk not clearly to be u
nderstood (as when many people are talking at once). 3102 314gentle s
ound of water flowing over stones, etc. 310bab322bler 316/246bablE(
r)/ 315n 314person who @s, esp one who tells secrets.304
330babe327 316/beIb/ 315n 3101 314(liter) baby. 3102 314inexpe
rienced and easily deceived person. 3103 314(US sl) girl or young wom
an.304
330babel327 316/246beIbl/ 315n 3101 the Tower of B@, 314tower b
uilt to reach heaven. 313(Gen 11). 3102 315(sing 314with 315indef
art) 314scene of noisy and confused talking: 315What a @! A @ of voic
es could be heard from the schoolroom.304
330ba322boo327, babu 316/246bqbu/ 315n 314(as Hindu title) Mr;
Hindu gentleman; Hindu clerk; (old use, pej) Hindu affecting English sp
eech and manners.304
330ba322boon327 316/bE246bun 315US: 316ba-/ 315n 314large mon
key (of Africa and southern Asia) with a dog-like face. 313333 314th
e illus at 313ape.304
330baby327 316/246beIbI/ 315n (pl 313-bies315) 3101 314very y
oung child: 315She has a 311*315@-311246315boy/311246315girl.
Which of you is the @ 314(= the youngest member) 315of the family? 3
11(be left) carrying/holding/to carry/to hold the @, 314(colloq) be le
ft responsible for sth one does not wish to be responsible for (because
of its difficulty or distastefulness. 310246@ carriage, 314(US) pra
m. 310246@-faced, 314looking much younger than one's age. 310246@-
farmer, 314(often pej) woman who contracts to keep (esp unwanted) babi
es. 310246@-minder, 314woman paid to look after a @ for long periods
(e254g while the mother is out working). 310246@-sit322ter, 314pe
rson paid to look after a @ for a short time (e254g while its parents
are at the cinema). Hence, 310246@-sit 315vi, 310246@-sit322ting
315n 310246@-talk 315n 314kind of speech used by or to babies with
distorted vocabulary and syntax. 3102 314(used attrib) very small of
its kind: 315a 311315@ car, 314a small motor-car. 310@ 246grand,
314small grand piano. 3103 314(sl) girl; sweetheart. 307 315vt 3
12[VP6A] 314(colloq) treat like a @: 315Don't @ the boy! 310246@32
2hood 315n 314state of being a @; time when one is a @. 310@322ish
315adj 314of or like a @: 315@ish behaviour.304
330bac322ca322laur322eate327 316/*bakE246lcrIEt/ 315n 312[C]
3101 314last secondary school examination in France. 3102 314univer
sity degree of Bachelor.304

Appendix 7: Sample of the parsed *OALD* file with single-character field keys

```
LB
Fb
Pbi
k
Ipl
HB's
Hb's
Pbiz
d
Sthe second letter of the English alphabet.
h
Lbaa
Pbq
k
Wn
d
Scry of a sheep or lamb.
b
Wvi
S(baaing, baaed or baa'd /bqd/) make this cry; bleat.
l
K%@-lamb
Wn
m
Schild's word for a sheep or lamb.
h
Lbaas
Pbqs
k
Wn
d
S(S Africa) boss.
h
Lbabble
P%babl
k
Wvi
Wvt
d
V2A
V2B
V2C
Stalk in a way that is difficult to understand; make sounds like a baby;
(of streams, etc) murmur.
d
V6A
V15B
l
N@ (out)
S, repeat foolishly; tell (a secret): @ (out) nonsense/secrets.
```

Appendix 8: Sample of parsed *OALD* file with full text field keys

```
headword                          B
alternative spelling of headword  b
pronunciation                     bi
+++++++start of pieces+++++++
conjugation or plural label       pl
conjugation or plural spelling    B's
conjugation or plural spelling    b's
pronunciation                     biz
----------definition----------
text                              the second letter of the English alphabet.
**********end of entry**********
headword                          baa
pronunciation                     bq
+++++++start of pieces+++++++
word class label                  n
----------definition----------
text                              cry of a sheep or lamb.
***change in part of speech***
word class label                  vi
text                              (baaing, baaed or baa'd /bqd/) make this cry; bleat.
====subentry====
derivative                        %@-lamb
word class label                  n
---subentry definition---
text                              child's word for a sheep or lamb.
**********end of entry**********
headword                          baas
pronunciation                     bqs
+++++++start of pieces+++++++
word class label                  n
----------definition----------
text                              (S Africa) boss.
**********end of entry**********
headword                          babble
pronunciation                     %babl
+++++++start of pieces+++++++
word class label                  vi
word class label                  vt
----------definition----------
verb pattern                      2A
verb pattern                      2B
verb pattern                      2C
text                              talk in a way that is difficult to understand;
                                  make sounds like a baby; (of streams, etc) murmur.
----------definition----------
```

Appendix 9: Sample of problems found during the parse of the *OALD*

```
baa          in text: possible unexpected pronounciation
bacchanal      in text: possible unexpected pronounciation
ballistic    in bracketed derivative: possible mistake
bank 3        in comments |315|
bath         in text: dummy conjugation or plural section
bay 3         in comments |*replaced ';' with ','*|
be 1         in comments |315after "I': 316|
be 1         in comments |314; 315otherwise: 316*inserted ','*|
be 1         in comments |314; 315strong form: 316*inserted ','*|
be 1         in comments |314; 315but 316*inserted ','*|
be 1         in comments | 315after 316p, t, k, f, T314; 315strong form: 316*inserted ','*|
be 1         in comments |*inserted 314';'*|
be 1         in comments | contracted forms,|
be 1         in comments | neg|
be 1         in comments |315Am I not 314is contracted to|
be 1         in comments |*replaced 315 with 313 ie "aren't I" become clearfont*|
beacon        in text: possible unexpected pronounciation
bed 1         in comments |315with 316|
bed 1         in comments | 315as in "dry', not separated as in "head-room'316|
beet         in comments | 315with 316tr 315as in "try'316|
beget        in comments | 314old use|
begin        in comments |*inserted 313*|
begin        in comments |313|
begin        in comments |,
begin        in comments |*inserted ') '*|
begin        in comments |*inserted '('*|
begin        in text: possible unexpected pronounciation
begin        in text: possible unexpected pronounciation
begin        in comments | 314the 315<v> 314being understood312|
behest       in comments |*replaced 316 with 315*|
benefaction    yylex sees unknown character 35 # replaced with a >
benzene        in bracketed derivative: possible mistake
bereave      in comments |*inserted ')'*|
bet          in comments |*replaced ',' with ')'*|
bet          in comments |*inserted '('*|
betake        in comments |, 314reflex312|
bevel        in comments | =|
beware        in comments | 314in the imperative and infinitive only312|
bicarbonate    in bracketed derivative: possible mistake
bicycle      in text: possible unexpected pronounciation
bid 1        in text: possible unexpected pronounciation
bid 1        in text: possible unexpected pronounciation
bilabial     in text: possible unexpected pronounciation
billetdoux    in comments |314, pronunciation unchanged|
billy        in headword: brackets
bind         in comments | 314in 315progressive tenses 314only312|
binomial       yylex sees  unknown character 323 replaced with a >
binomial       yylex sees  unknown character 323 replaced with a >
```

Databases in the Humanities and Social Sciences–4

Proceedings of The International Conference on Databases in the Humanities and Social Sciences held at Auburn University at Montgomery, July, 1987

Edited by Lawrence J. McCrank
Auburn University at Montgomery

Learned Information, Inc.
Medford, New Jersey 08055
U.S.A.

Published by Learned Information, Inc.
143 Old Marlton Pike
Medford, New Jersey 08055
U.S.A.

Copyright© 1989 by Learned Information, Inc.
All rights reserved.

Printed and bound in the United States of America.
ISBN 0-938734-97-7

Contents

Preface: ICDBHSS '87 and Beyond
Lawrence J. McCrank ix

PROFILE: A Humanities Computing Workbench
Peter Adman 1

Optical Disk and the Developing Countries
S. Nazim Ali 9

Natural Language Interface Without Artificial Intelligence
S. Pal Asija 13

A Lexical Database for English Learners and Users:
The Oxford Advanced Learner's Dictionary
Eric Steven Atwell 21

"M" to "Moonless": Lexical Databases in Development
Paul Beam and *Frank Huntley* 35

Buildings as Structures, as Art and as Dwellings:
Data Exchange Issues in an Architectural Information Network
David Bearman 41

Information Delivery in the Social Sciences and Humanities:
The Changing Role of the Library
Tony Carbo Bearman and *Linda H. Schumacher* 49

Access to Electronic Information:
Problems in Resource Allocation and Policy Formation
Joseph Behar 55

New Computer Technologies and Social Science Research Methods
Howard Besser, Robert Yamashita and *Troy Duster* 61

The Dissertation Abstracts International Database and the Humanities Researcher: Theoretical Applications and Practical Applications
David J. Billick 71

An Expert Dedision Support System for a Prosopographical Database
Caroline Bourlet and *Jean-Luc Minel* 79

Visual Databases for Biomedical Teaching
Charles E. Branch and *James W. Woods* 85

Hi-Tech Information Storage in the National Archives
Frank G. Burke 93

An Expert System for Online Retrieval in the Humanities: