Reading in a Foreign Language April 2010, Volume 22, No. 1
ISSN 1539-0578 pp. 181–200

Learning about language and learners from computer programs

Tom Cobb
Université du Québec à Montréal and McGill University
Canada

Abstract

Making Nation's text analysis software accessible via the World Wide Web has opened up an exploration of how his learning principles can best be realized in practice. This paper discusses 3 representative episodes in the ongoing exploration. The first concerns an examination of the assumptions behind modeling what texts look like to learners with different levels of lexical knowledge; the second concerns approaches to handling proper nouns in text profiling within an international context; and the third involves the future of the Academic Word List as new frequency information appears to undermine its utility. Underlying these explorations is an argument that writing computer programs is a useful way to investigate language and language learning.

Keywords: computer text analysis, lexical frequency profiling (LFP), Range, Vocabprofile (VP), Academic Word List (AWL), Vocabulary Levels Test, text coverage, frequency list, learner modeling

Computational text analysis underlies much of Paul Nation's research agenda as well as its pedagogical bearing. The impact of this analysis within applied linguistics research is well known (through a number of high profile papers from roughly Laufer & Nation, 1995, through to Webb & Rodgers, 2009, at date of submission). Less well known is its impact on teachers and learners or how these players have used and adapted the agenda and its technologies. In essence, Nation's writings have given users a practical means of responding to the now widely accepted but still basically unoperational idea that "language learning is largely lexical learning" (Gass & Selinker, 2008, p. 173). The Range computer program (Nation & Heatley, 1994; Heatley, Nation, & Coxhead, 2002) makes it possible for teachers to devise plausible sequences of lexical acquisition, target specific lexical needs, or assess and modify the learning burdens and opportunities of texts and tests. It allows learners to develop their lexical knowledge at a particular level rather than randomly.

Many of the uses that teachers and learners have made of Nation's ideas have been through my website Lextutor (www.lextutor.ca), which is an attempt to reverse engineer some of Range's key functions in an accessible and user friendly format on the World Wide Web. Broadening and increasing this access has had the unexpected effect of creating a decade long, two-way conversation with users, who have contributed many responses and suggestions, creating in effect a whole sub-agenda of research and development.

Lextutor's original goal was to imitate Range routines as closely as possible, or to develop the tutorial dimensions of Range-based ideas such as the Levels Tests (Nation, 2001; Schmitt, Schmitt, & Clapham, 2001) or related frequency lists, and to deliver the whole package over the Web from a single location to three overlapping constituencies—learners, teachers, and researchers. Right from the start, however, the Web format posed development challenges that required most of the routines to be different from their Range equivalents in some way. For one thing, the size of files that can be handled online is smaller than on a single user PC.

The Web format also meant that many, many people could use Nation's routines, from many language backgrounds, and that their ways of using them could be tracked. The tutorial dimension meant that the programs and the ideas behind them could come into closer contact with teachers and learners than might have been the case if these had remained only research instruments. As a result, many teachers and learners have offered insights and suggestions leading to Lextutor becoming a kind of ongoing group-development project.

This close contact with the end-user has sometimes led to simple adjustments to make ideas clearer or more salient. My own use of Lextutor's version of Range's lexical frequency profiling (LFP) and Vocabprofile (VP) with classes of graduate students led me to realize that these students were not getting the true sense of the profile of a text from a bare list of word token percentages across different levels, but that they did get this from an integral version of the text with levels indicated by different colors. Closer contact has also led to the addition of features that help users in the tasks they are actually using the different tools for. Many teachers and course developers use Vocabprofile to modify the lexical profiles of instructional texts for their learners, but to do this they had to travel between the entry page and the output page repeatedly, at risk of losing work in the event of a network collapse, so a same-page system was devised with input and output appearing simultaneously.

These and many similar interface adjustments have seemed helpful to users but do not greatly advance the research agenda or pose interesting new questions. But this has not always been the case. This paper will look at three recent and representative cases where Lextutor's close contact with its user base has impacted the research agenda, creating or potentially creating interesting new information or questions. The sub-text of the paper, and I think a sub-text to Nation's research, is that language and learning can be usefully explored with computer programs, and in the conclusion I will attempt an appropriate characterization of this type of research.

Increasing the Grain Size of Learner Portraits

If Paul Nation had retired right after the publication of his book *Teaching and Learning Vocabulary* (1990), that one volume alone would have provided enough material to keep both Lextutor developer and vocabulary researchers generally busy for a lifetime. One of the book's most memorable features is an appendix with a figure summarizing one of its most interesting ideas, what a text looks like if various levels of its vocabulary (as identified by the Range profiling program) are replaced with blanks. It shows how the text is experienced, or what it looks like, to a learner who knows the most frequent 1,000 words, or 2,000, or those plus the

University Word List (Xue & Nation, 1984, a precursor of the Academic Word List [AWL], Coxhead, 2000). The topic of the text chosen for this demonstration was New Zealand forestry, a topic most readers would easily understand but not have detailed prior knowledge of. An obvious extension of the Forestry text idea was to adapt Vocabprofile to perform a similar analysis of other texts, and following that to develop a cloze passage builder that would turn such texts into exercises for learners. Using the builder program, a teacher whose students scores on the Levels Test (Nation, 1990) indicated a sound knowledge of only the first 1,000 words but little beyond that could find a text and make an exercise with second 1,000-level words blanked for replacement. Further, within the computer environment, such exercises could be linked to available online resources like online learners' dictionaries. On users' advice, the pedagogical adaptation also involved targeting some levels that had not been provisioned in the original Forestry scheme, such as the words beyond any of the other levels (the off-list words). The routine can be visited (http://lextutor.ca/cloze/vp_cloze/).

The pedagogical users of the VP cloze builder did not remain passive consumers of the program, however, but quickly began to feed their experiences of using the program back to its basic principles. Many argued that the Forestry version of what a text looks like (or, how it feels to try to read it) does not accommodate the fact that learners will know some of a word's affixes even if they do not know the word itself. In the Forestry text the blanks are of equal sizes, but in a real text learners can easily see whether the word is short (probably a function word) or long (definitely a content word). Most interesting, many found the knowledge levels too general. A learner rarely knows all or nothing at a particular level, as the Forestry scheme appears to suggest, although of course this is only a programming convenience.

To the likelihood of mixed knowledge levels one could add that it is also not necessarily the case that learners know more words at a higher frequency level than they do at a lower frequency level. Words seem to be learned roughly in order of frequency in first language (L1) development (Biemiller & Slonim, 2001), but this cannot be assumed in a second language (L2). My own research using a more recent version of the Levels Test (Nation & Beglar's, 2007, Vocabulary Size Test) with several groups of both school and adult learners in Quebec has often showed these learners knowing as many words at a medium-frequency level (3k, 5k) as at a higher frequency level (1k, 2k). Obvious reasons that L2 learners may not follow a sequential growth pattern could include that they have cognate L1s where the shared items are from medium- and low-frequency levels, or that they have reached a high level of proficiency within a technical domain in the L2 but no experience of everyday situations and interactions.

Gradually these three responses were encoded as options on Cloze_VP. Gaps can now be constructed with one underscore for each letter (*cat* = ___, *interesting* = ___________). Inflections and affixes from Bauer and Nation's (1993) levels 0 (common inflections) and 1 (high-frequency affixes that do not change the base word) are left attached to the gap, provided the gap represents a standalone word (thus *replaced* becomes *re____d*, but *replacing* does not become *re____ing* since *plac** does not stand alone). Mixed knowledge levels can be chosen for each level in the classic 1k, 2k, and AWL scheme. For example, 60% across levels can be chosen via a menu, and the program chooses these proportions randomly.

Figure 1 shows first the third paragraph from the original Forestry text in the first column, then the same paragraph with all post-1,000 word families replaced by an equal sized gap in the middle column as in Nation (1990), and finally the same paragraph with 40% of 1k families, 30% of 2k, and 20% of AWL replaced by actual size gaps and bearing original affixes in the right column. The final text is arguably more accurately "what the text looks like" to a learner who knows 60% of 1k items, 70% of 2k, and 80% of AWL. These modifications lead to the interesting question of what sorts of information learners actually use to fill these gaps, and whether it interacts with level, and this software would make the creation of a set of experimental materials straightforward.

Forestry A: Intact version	Forestry B: 1990 version, 1k words known	Forestry C: Mixed profile of words known + affixes + size information
Even if used in an unprocessed form, the increasing wood supplies will require a larger labour force, an improved roading network, and expanded transport and processing facilities. If the trees are to be exported, then certain investments must be made. They will include investments in: logging machinery and equipment; logging trucks, and other vehicles required for the transport of processed products; upgrading and maintaining roads (or rail or coastal shipping facilities where appropriate); and port facilities. The list could be extended to include overseas shipping, and accommodation and township facilities for forestry workers.	Even if used in an unprocessed form, the increasing _____ _____ will require a larger labour force, an improved roading network, and _____ _____ and processing facilities. If the _____ are to be _____, then certain _____ must be _____. They will _____ investments in: logging machinery and _____; logging trucks, and other vehicles _____ for the transport of _____ products; upgrading and maintaining _____ (or _____ or coastal ______ _____ where appropriate); and port _____. The list could be extended to include _____ ______, and _____ and township facilities for forestry _____.	Even if used in an unprocessed form, the increasing ____ ________ will require a larger labour force, an improved roading network, and ______ed _________ and processing facilities. If the ____s are to be ______ed, then certain __________s must be ____. They will ______ investments in: logging machinery and _____ment; logging trucks, and other vehicles _______d for the transport of _______ed products; upgrading and maintaining ____s (or ____ or coastal ________ __________ where appropriate); and port __________. The list could be extended to include ________ ________, and _____________ and township facilities for forestry ______s.

Figure 1. Different versions of what texts look like to learners.

Another calculation these modifications make possible is vocabulary size and coverage for mixed profile learners. As proposed above, these may be quite common in L2 acquisition. What difference would it make if a learner knew 2,000 words in sequence versus 2,000 words total but at various levels? The default answer to this question is that knowing the most frequent 2,000 word families gives a learner knowledge of 80% of the terms in average texts, so the learner who knows these should have the advantage. But this might not be true for all types of texts.

Take our learner again who knows 60% of first 1,000 items, and 80% of both second 1,000 and AWL items (570 word families). This learner thus knows 600 + 800 + 432 = 1,856 word families. What percentage of the Forestry's lexis does this learner know, in comparison to a hypothetical learner with a sequential accumulation of the most common 2,000 words? Table 1 shows VP_Cloze's coverage percentages for these two knowledge profiles for four types of texts: academic (Forestry and a section of an applied linguistics article), quality journalism (two pieces from Canadian commentator Rex Murphy, Globe & Mail, from www.theglobeandmail.com /news/opinions/columnists/rex-murphy/), fiction writing (a chapter from Jack London's, 1903, *Call of the Wild*), and simplified fiction (chapters from the simplified versions of the Oxford Bookworm Series' *Call of the Wild* and *Elephant Man*).

Table 1. *Two ways of calculating percentage of words known*

		Coverage	
	Number of words	First 2,000 words known	1,832 words known at mixed levels
Forestry	374	76%	83%
Applied linguistics	1,012	80%	84%
Rex Murphy 1	882	85%	87%
Rex Murphy 2	937	86%	85%
Call of the Wild, Ch 1, Original	3,719	85%	87%
Call of the Wild, Ch 1, Simplified	877	96%	75%
Elephant Man, Ch 1, Simplified	1,131	97%	75%

Note. The program assumes that all function words and proper nouns are known or interpretable, and no off-list words are known. In the Forestry text, 38% of items are function words, and 7% are off-list words.

Table 2. *More AWL and technical = less 1k*

			Percent of word tokens		
Genre	Percent 1k	2k	AWL	Off-List	AWL + Off-list
Fiction	87	4	1	9	10
Fiction	82	7	0	12	12
Scientific	65	6	18	11	29
Scientific	85	5	13	16	29
News	87	6	4	3	7
News	85	3	5	6	11
Mean (*SD*)	82 (8)	5 (1.5)	7 (7)	10 (5)	16 (10)

Note. Percentages are rounded.

Is there anything new in these coverage figures? I think so. They suggest that reading a natural text is pretty hard going for either of these learners, but if we accept that texts begin to come into focus at 90% known lexis (Schmitt, Jiang, & Grabe, 2010) and are fully in focus only at 98% (Nation, 2006), the mixed profiler nevertheless comes off marginally better for the more difficult texts (applied linguistics and Rex Murphy) but much worse for the simplified readers—and by inference also for English as a second language (ESL) course materials, or tests of elementary reading comprehension.

Perhaps it is not so surprising that a strong AWL + off-list can compensate to some extent for a weak 1k in academic or specialist texts. This would make sense if proportions of 1k and AWL/specialist items were inversely related, as appears to be the case. A possible reason for this would be that many English words have higher and lower frequency versions (*sweat* vs. *perspiration* and others). Table 2 shows classic profiles from VP's six demonstration texts across a range of types (two unsimplified fiction texts, two academic texts, and two newspaper articles). The table shows that variation at the 1k level is rather high (*M* = 82, *SD* = 8); at the 2k level much less (*M* = 5, *SD* = 1.5); and at the AWL + off-list levels again high (combined *M* = 16, *SD*

= 10). The high variations in the first and last columns are moderately strong and negatively related, *r* = -.65, *p* < .001. It appears that a large AWL component can predict about 5% reduction in the 1k component. Such a difference may seem small but as Nation reminds us a difference of 5% is one word in 20. In other words, mixed-profile readers could have up to 5% fewer 1k items to deal with if they stuck to academic or specialist texts, and thus any weakness they had in the high-frequency zone would affect them less.

This investigation is clearly preliminary and is intended mainly to give a taste of the interesting questions that can be raised by modifying Range according to practitioners' needs and responses. Nevertheless, if confirmed with a larger number and length of texts, this finer-grained portrait of the learner would raise at least two practical questions. First, if we used the Levels Test for diagnostic purposes, should the 60–70–80 profiler be placed in the beginner's class or in the advanced class? What injustice would be committed if we tested this learner's reading ability by his comprehension of a simplified text? Does this information tell us anything about the problems that teachers sometimes experience getting adult learners to read simplified stories? It seems conceivable that for such learners the simplified content might seem silly but the language difficult.

The second question concerns how typical the nonlinear profile is, and more broadly whether there is an L1-L2 split on the question of learning sequence by frequency. Milton (2009) has found some evidence for nonlinear profiles, particularly in early stages of learning, in a 7-year study of French as a foreign language students in a British school. My research suggests that many adult ESL learners in Montreal are mixed profilers who, for reasons suggested above, perform better with technical texts than with easy texts or conversations. A test of this would be to give large number of such learners Nation and Beglar's (2007) 14k Vocabulary Size Test, count the sequential and non-sequential profiles, and determine empirically what is in fact hard and easy for each group to read. Third there is the question of how we should use frequency based vocabulary tests as an aid to needs analysis and instructional design in such cases. My hunch is that if the goal of a learner with a 60–80–80 profile is to live in English, then he or she should do something about the weakness with very common words, but if the goal is to read in a professional domain, then technical lexis is probably the shortest route to higher coverage.

On the Proper Treatment of Proper Nouns

In some cases, like the one above, user modifications have led to interesting theoretical speculations and potential new research. In other cases they have led to programming challenges. An example of the latter involves the question of how to handle proper nouns in Vocabprofile. In recent versions of Range, Nation has included as a separate category (called 15k since it follows the currently final 14th thousand list) an itemized and growing list of as many proper nouns as possible, derived from the British National Corpus (BNC) and elsewhere and treated in the output as non-lexical items. A 50-word stretch of an early (12,424 item) version of this itemization is shown in Appendix 1. Lextutor users have not found this particular categorization satisfactory, coming as they do from a wide variety of language backgrounds, each with its own massive array of names and other proper nouns. It is unlikely that any particular listing can ever pick up a meaningful proportion of all the proper nouns from all the texts that are run through

Range or Lextutor on a daily basis from every corner of the world. A programmatic rather than itemizing solution to this problem would therefore be interesting, if it could be achieved and was effective.

It is difficult to explain to novice Vocabprofilers that proper nouns are not lexical items. The text “Pierre lives in Beaurepaire” is comprehensible enough without knowing more than that *Pierre* is somebody’s name and *Beaurepaire* is the place this person lives in. More information is added if we know that these words are French and that Pierre is a name for males, but the sentence can be processed well enough to get the reader to the next sentence without knowing this. If the text went on to develop a rich portrait of life for this person in this place, the reader would gain further encyclopedic knowledge centered on these proper nouns, but *Pierre* and *Beaurepaire* would still not amount to generative lexical items.

This point grasped, the next hurdle is to show the novice profiler that a proper noun is nonetheless a factor in a text’s lexical density and is hence factored into the calculation of text coverage. There are two ways of calculating the profile of the example sentence above from the perspective of a beginner who knows 1,000 words of English. By one method, *lives* and *in* are both common first 1,000 (1k) words, while *Pierre* and *Beaurepaire* are off-list words, so for this reader the sentence comprises 50% known items. By another method, if we can assume the learner understands the concept of a proper noun and the main kinds of these (persons, places) then the text is 100% known or at least comprehensible. The second method is clearly more realistic, but it is not obvious how it is best realized.

Nation (e.g., 2006) and his student Stuart Webb (e.g., Webb & Rodgers, 2009) follow the second method, calculating proper nouns as a separate category and adding them to the level or levels they are investigating to get a coverage figure. Webb and Rodgers, for example, report results showing “that knowledge of the most frequent 3,000 word families plus proper nouns . . . provided 95.45% coverage” (of most television shows, p. 335). Two problems with this approach are that the proper noun calculation is an extra step taken by the researchers that is not actually shown in the program’s output, and that practitioners using Lextutor for various materials design and action research projects rarely adopted it. As a result, especially in the case of fiction (still the main reading diet in language classes), without the many names of people and places factored into the coverage calculation, texts are made to seem more lexically challenging than they really are. A way of incorporating proper nouns into a coverage estimate that has seemed clearer to Lextutor users is to give them the option of reclassifying proper nouns as first 1,000 items. The various Web versions of Vocabprofile make it easy to do this by simply double-clicking on the words to reclassify them in the input text.

How proper nouns are handled makes a big difference to an output profile. Nation (2006) provides a table showing the difference in overall coverage between two methods of handling proper nouns in creating profiles for Lawrence’s (1929) *Lady Chatterley’s Lover*. One method classifies proper nouns as off-list items, and the other classifies these, as found in Nation’s 15k collection, as known items. The difference is a reliable 2% (*SD* = 0.02), as shown in Table 3 below, reproduced from Nation (2006), but with a differences column and mean differences row added for the purposes of the present investigation. Basically, a 2% difference is established by adding all the story’s proper nouns to the first 1,000, and this difference is maintained through the remaining levels, accounting at the 14th-thousand level to over 99% of the story's lexical items. As already noted, small percentages can have big effects on text coverage.

However, *Lady Chatterley's Lover* is a thoroughly English story, and it is quite likely that Range's itemized proper nouns list handled these particular proper nouns rather well—better than might be the case for an article in The Teheran Times, The South China Morning Post, or indeed The Montreal Gazette.

Table 3. *Cumulative percentage coverage figures for Lady Chatterley's Lover by the fourteen 1,000 word families from the BNC, with and without proper nouns, achieved by itemized lists*

1,000-level	Coverage without proper nouns (%)	Coverage including proper nouns (%)	Difference
1	80.88	82.93	2.05
2	88.09	90.14	2.05
3	91.23	93.28	2.05
4	93.01	95.06	2.05
5	94.08	96.13	2.05
6	94.77	96.88	2.11
7	95.38	97.43	2.05
8	95.85	97.9	2.05
9	96.17	98.22	2.05
10	96.41	98.46	2.05
11	96.62	98.67	2.05
12	96.82	98.87	2.05
13	96.93	98.98	2.05
14	96.96	99.01	2.05
	Mean difference		2.05
	SD		0.02

Note. From Nation (2006), with Difference column added.

Following years of comments from Lextutor users and then a discussion with Batia Laufer at a conference in Mexico in October 2008 (personal communication), I began looking for a more global method of identifying the proper nouns in an English text. Laufer's suggestion was to develop an algorithm to find all the mid-sentence capitals of a text. The algorithm has now been built from regular expressions in the PERL scripting language, and has been deployed as an option on all of Lextutor's various versions of VP, with Laufer's collaboration on points of interpretation. This is the algorithm:

```
@capwords = ($no_lines =~ /[^\.!?:]\s+(?=(\b[A-Z][A-Za-z]+\b))/g)
```

Here is what the algorithm does: A version of the input text with no line endings is created ($no_lines), and from this list is generated an itemized array (@capwords) consisting of all the words (strings between spaces \b and \b) that begin with a capital letter ([A-Z]), are followed by any number of other letters whether capitalized or not ([A-Za-z]+), but *not* (^) preceded by a terminal punctuation mark ([^\.!?:]) plus any number of spaces (\s+)—and this throughout the text, or globally (/g). Once created, @capwords is added to the 1k list and the normal profiling procedure is begun.

The @capwords approach successfully creates a list of candidate proper nouns but still raises some issues about how it is to be deployed. First, a name or other proper noun can often occur at the beginning of a sentence (e.g., *Simon* in "*Simon* thought he was alone") and thus will not join the list. But this is only a problem in very short texts (say, under 250 words); in texts of any length, it is doubtful that a name will appear only once, or always as the first word in a sentence (evidence for this is offered below). If the algorithm finds the word at mid-sentence even once, it is added to @capwords and handled as a proper noun throughout the text including at a sentence boundary. Second, there are names, particularly of places, that while proper nouns, are also lexically meaningful to a greater or lesser degree. In the sentence, "We went to the top of the Statue of Liberty," clearly the learner who knows *statue* and *liberty* gets more from the sentence than the learner who knows only that it is *the name of something you can go to the top of* (example from Batia Laufer). The solution to this is to run the proper-finding algorithm only on off-list items; that way, any lexical element or connotation the name may have will get its due. In Statue of Liberty, *statue* will appear as a 6k word and *liberty* as 4k (by the BNC scheme), and the learner who knows words at these levels can be predicted to enjoy a comprehension advantage over a learner who does not. And finally, the converse problem to the preceding is that some names, usually of people, have no lexical dimension when used as names (such as George *Bush*) and should therefore not be counted as lexical items (*bush* is 3k on the BNC scheme). This problem arises fairly rarely, and at present there is no programmatic way to deal with it. VP gives users a way to block such words manually from making the text appear richer than it is (i.e., to designate *Bush* as a proper not lexical item and enter it into 1k).

Table 4. *Cumulative percentage coverage figures for Lady Chatterley's Lover by the fourteen 1,000 word families from the BNC, with and without proper nouns, achieved by algorithm*

1,000-level	Coverage without proper nouns (%)	Coverage including proper nouns (%)	Difference
1	83.08	85.26	2.18
2	88.61	90.79	2.18
3	91.69	93.87	2.18
4	93.24	95.42	2.18
5	94.17	96.35	2.18
6	94.84	97.02	2.18
7	95.35	97.53	2.18
8	95.65	97.83	2.18
9	96.01	98.19	2.18
10	96.24	98.42	2.18
11	96.46	98.64	2.18
12	96.63	98.81	2.18
13	96.8	98.98	2.18
14	96.88	99.06	2.18
		Mean difference	2.18
		SD	0.00

Note. From Nation (2006), with Difference column added

Does this approach sort the proper nouns properly? The first test is to repeat Nation's (2006) exercise with *Lady Chatterley*. Table 4, which can be compared to Table 3 above, shows that the algorithm is slightly more successful at pulling out proper nouns than the itemized list was (2.18% mean increase over calculation without proper nouns, compared to Nation's 2.05%). The slightly higher coverage across the levels is due to Lextutor's separation of contractions into component words.

The second test is to check that the increased coverage has not been achieved by extracting items that are not actually proper nouns, such as capitalized words at the beginnings of quoted dialogue. Appendix 2 shows the 290 Chatterley items that were extracted and added to the first 1,000 category in order of appearance. Apart from some possible typos in the electronic version of the novel (*Ev*, *Wor*), and some foreign words that are probably meant to have lexical meaning to those who know the language (*La Terre, Auto Da Fe*), it seems only three English items, *Charlestoned*, *Bolshevistic and Londonized*, carry potential lexical content.

The final test is to apply the algorithm to an English text from a non-English speaking zone, such as an English-language newspaper in a country where English is not the primary language, the type of problem that inspired our interest in an algorithm in the first place. The English translation of de Maupassant's Boule de Suif (1880/1990; 14,436 words) can serve to represent this type of text. It is an extended English text but with all its persons and places in French or German. Its proper noun output, shown in Appendix 3, does not appear to contain any content words, with the possible exception of *Bonapartist* and the unexpected *Godforsaken*.

And finally the off-list component of the Boule de Suif profile (0.84% of tokens, or 114 words) is shown in Appendix 4, the usual assortment of misspellings, Briticisms or Americanisms, foreign words and unclassified nonce words and compounds, showing that only two proper nouns have somehow failed to be identified by the algorithm, Catherine and Judith. Closer inspection reveals that Catherine is a remnant of "Ste-Catherine's Hill," a place name, leaving one error in 14,000 words.

The conclusion appears to be that an algorithmic approach is substantially correct—pulls at least as many proper nouns out of an average text as a dedicated list does, is able to find proper nouns in texts from anywhere that English is used, and additionally is fully adaptable to Vocabprofiling in languages other than English (as explored in Ovtcharov, Cobb, & Halter, 2006).

3. Is There an AWL in English?

At the EUROSLA vocabulary conference organized by Batia Laufer and Paul Bogaards at the University of Leyden in March, 2002, Paul Nation and I discussed the findings of a study by Hazenberg and Hulstijn (1996) which had appeared to suggest that a Dutch reader would need to know 90% of the vocabulary of a Dutch academic text to achieve basic comprehension of its content, which in their analysis would correspond to knowing 10,000 word families. We agreed that this rather high number was probably a result of the Dutch language's not possessing a zone of lexis corresponding to the AWL in English, or at least of no one having found one yet, as Averil Coxhead (2000) under Nation's supervision had found for English.

Capitalizing on some accidents in the development of English (the Norman conquest and bifurcation of the language), Coxhead showed in a corpus study that a smallish set of 570 mainly Greco-Latin word families, of medium (post-2,000 level) frequency in English as a whole but much higher frequency in the discourse of scientific texts, when added to the 2,000 families of the General Service List (GSL; West, 1953) will normally give academic learners about 90% coverage in the texts they are studying (or a little more since they will also know some technical items in their subjects). All of this made a rather convenient fit with existing research showing that knowledge of 95% of an English text's lexis was sufficient for basic comprehension (Hirsch & Nation, 1992; Laufer, 1989). Accordingly, the ESL and EFL (English as a foreign language) course writing industry set about developing course books and Websites devoted to teaching and learning the AWL (e.g., Schmitt & Schmitt, 2005, or of course the VP_Cloze web routine mentioned earlier).

Later when Nation began experimenting with BNC versions of Range, based on frequency lists from a 100-million word corpus, he made some discoveries that appeared to unsettle the happy GSL + AWL picture. Following a familization and carve-up of the massive BNC frequency list (Leech, Rayson, & Wilson, 2001) into 1,000 family divisions, Nation built an updated Range and corresponding Vocabulary Size Test (Nation & Beglar, 2007), and from about 2005 began using these to re-pose some of the basic questions of his agenda. One of these was the question about percentage of text lexis needed for unassisted reading and how many word families this corresponds to. Hu and Nation (2000) had determined that 98% is the percentage needed, and Nation (2006) finds that this is typically achieved when one knows 8,000–9,000 word families for written texts or 6,000–7,000 for spoken. The other question was about the validity of the AWL. Running the AWL headwords through the BNC version of Range showed that about half of them are first-2,000 level items under the new scheme, or in other words "the AWL is an artifact of the GSL" (of using the GSL as a basis for defining the AWL, personal communication, 2006). The conclusion appeared to be that English is like Dutch after all.

Raising the learning task from 2,580 to 8,000 word families gives a rather different picture of what is needed to read in a second language. I have recently used corpus data (Cobb, 2007) to show that the natural distribution of words in texts makes it very difficult for L2 learners to get much beyond 2,000 word families on their own through reading, for the demonstrable reason that post-2,000 words simply do not appear often enough for reliable learning. This finding appears to accord quite well with Laufer's (2000) compilation of seven sets of Levels Test results from eight countries showing an average vocabulary size for academic learners of about 2,100 word families (*SD* = 977). How can the aspiring academic ESL or EFL learner ever acquire the 8,000 words that Nation's data suggests they need in order to read effectively in their studies?

In the field, it is not clear that this new information has sunk in. Practitioners seem to be ignoring the Nation (2006) findings, at least the Lextutor users worldwide who are sticking with the classic version of VP (i.e., the GSL + AWL) at a ratio of five to one despite the obvious advantages of the BNC version (such as the vast reduction of uncategorized or off-list items). However, reactions to the Nation bombshell are slowly coming in. One reaction has been to look in the data for continua rather than the cut-offs or thresholds that Nation has often seemed more interested in. Schmitt et al. (2010) tagged comprehension success to percentage of text-lexis known, and found instead of a cut-off a continuum of comprehension from 50% (*SD* = 18) when 90% of lexis is known ,through 75% (*SD* = 15) when 100% is known, with the remainder presumably accounted for by topic knowledge and other factors. It is indeed true that learners, even when studying academic subjects via English, manage to survive with less than perfect reading comprehension, and that unassisted reading may be an unnecessarily lofty goal. On behalf of Lextutor users, I propose another type of reaction, the search for a modified AWL within the BNC framework, beginning with the following feasibility study.

The weak link in the GSL + AWL scheme was indeed the pre-corpus era GSL. The exact problem can be seen if we run the GSL's 2,000 headwords through the BNC version of Vocabprofile, the result of which is shown on the left column in Table 5. It seems that somehow the GSL contains about 500 fairly infrequent items. From a drill-down into the data, here are the 23 GSL items that the BNC scheme classifies as 6k: *accustom*, *applaud*, *applause*, *barber*, *beak*, *cape*, *coarse*, *conquer*, *inquire*, *noun*, *oar*, *paste*, *procession*, *quarrel*, *quart*, *rejoice*, *roar*, *saucer*, *scent*, *tame*, *tribe*, *vain*, and *veil*. Some of these apparently miscategorized words are probably useful classroom words, as West originally argued (the item at K13 in Table 5 is *scold*) but these are arguably better housed in a dedicated specialist list. It seems quite likely that the useful part of the GSL was really a list of about 1,500 high-frequency word families, with another roughly 500 along for the ride but rarely appearing in Range or Vocabprofile outputs.

Table 5. *The GSL's 2,000 families and the AWL's 570 families as seen by the BNC*

BNC frequency level (1,000's)	Number of GSL families	Number of AWL families
K1	849	82
K2	534	198
K3	325	87
K4	134	98
K5	55	60
K6	23	19
K7	10	13
K8	2	7
K9	2	2
K10	2	2
K11		
K12		
K13	1	

The fate of the 570 AWL families is shown in the right column of Table 5. A total of 280 AWL items falls within the first two 1,000 levels of the BNC. Looking at the BNC's coverage of the GSL and AWL together, we might draw the conclusion that the first two BNC 1,000 lists have both trimmed a lot of unessential items from the GSL and are doing a lot of the work that used to be done by the AWL.

The power of the first two BNC lists is further illustrated by looking at typical text coverage. In many types of texts, the first two BNC lists provide as much coverage as was previously achieved by the GSL and AWL together. As Lextutor runs both old and new versions of the software, this proposition can be tested by simply running the same text through the two versions. For example, classic VP analysis for one of Rex Murphy's lexically rich diatribes (demoed on the entry page) shows just under 88% of words claimed by the GSL and AWL (2,570 families), while BNC-VP analysis shows just under 88% of words claimed by the first two 1,000 levels alone (2,000 families). With this encouragement I ran four random 1,000-paragraph sections of the BNC Written samplers corpus (Oxford Computing Services, 2005) through VP and again found about 90% coverage for the first two BNC 1,000 lists plus proper nouns (see Table 6).

Table 6. *Coverage of the first two BNC 1,000 lists across randomly chosen written texts*

BNC-written sampler section	Number of words	1k + 2k coverage (%)
1	5,529	85.60
2	7,308	87.91
3	6,390	90.35
4	7,815	95.73
Mean	6,761	90,00
SD	1,010.9	4.34

The power of the first two BNC lists could be explained as an artifact of the written language bias of the BNC corpus from which they are derived. In fact, Nation took pains to include spoken language components of the BNC in the first two lists to be sure that words like *please* and *thanks* did not end up as third and fourth 1,000 items (personal communication, 2008). A test of the lists' spoken coverage would be to run a speech corpus through the two versions of VP. The recently created 105,000 word ALERT corpus of ESL classroom teacher talk (used in Collins, Trofimovich, White, Cardoso, & Horst, 2009) was thus run through the two versions, with equal handling of proper nouns, with the result that the first two 1,000 lists of the classic VP (GSL+AWL) claimed 95.82% of tokens, the BNC version 95.56%, virtually identical. The provisional conclusion appears to be that the new lists provide a large increase in written text coverage at the high-frequency levels, compared to the GSL, at no cost in spoken text coverage.

This higher text coverage for the BNC-2,000 affects the prospects or indeed the need for an AWL. If 85%–90% coverage by the first two lists is consistent across text types, then there might be "no room for an AWL" in English as Cobb and Horst (2004) argued was the case for French. After all, there has to be some space for domain-specific and true low-frequency items. But is 85–90% coverage by the BNC-2k plus proper nouns indeed consistent? Unfortunately, for texts that were AWL-heavy in the classic framework (>10% of tokens), the coverage provided by the first two BNC 1,000 lists remains about the same. Figure 2 shows a pair of texts that have lived as user demos on Lextutor for years, the first with a classic profile of GSL = 70% and AWL = 18%, the second with a classic profile of GSL = 72% and AWL = 13%. For these texts, the GSL and BNC first two 1,000 list coverages are very similar at 74% and 72%, respectively, and 90% coverage is achieved only after 6,000 and 5,000 words, respectively. In this type of text, in other words, there is still room for an AWL.

Where might a new AWL come from? Some of it could possibly be found in the remainder of the old one. To test this possibility, Lextutor's BNC version of VP was configured to add an extra list to those already deployed. This list was the "rump AWL," the remainder of 288 AWL families that had not been claimed by the BNC first two 1,000 lists (from Table 5). What coverage does the rump AWL provide in these texts? BNC-Vocabprofile gives it 7.95% coverage for the first text of Figure 6, and 3.97% for the second. These are rather large figures considering the list is fewer than 300 families. With this encouragement, I ran four random 1,000-paragraph stretches of the BNC medical sub-corpus through the BNC version of VP, configured to add the rump AWL as an extra category, and the results are shown in Table 7. The exercise would need to be repeated with more kinds of academic texts, but the pattern here seems to be that for at least some academic texts the first two BNC 1,000 lists provide a low enough coverage to allow some room for an AWL to function (*M* = 83%, *SD* = 7.7%), in which the rump of Coxhead's AWL claims just under 2% (*M* = 1.86%, *SD* = 0.47%).

Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families.

Abstract Annotation of the biological function of different protein sequences is a time-consuming process currently performed by human experts. Genome analysis tools encounter great difficulty in performing this task. Database curators, developers of genome analysis tools and biologists in general could benefit from the access to tools able to suggest functional annotations and facilitate access to functional information.

In the paper, we present a prototype system for the automatic annotation of protein function. The system is triggered by collections of abstracts related to a given protein, and it is able to extract biological information directly from scientific literature, i.e., MEDLINE abstracts. Relevant keywords are selected by their relative accumulation in comparison with a domain-specific background distribution. Simultaneously, the most representative sentences and MEDLINE abstracts are selected and presented to the end-user. Evolutionary information is considered as a predominant characteristic in the domain of protein function. Our system consequently extracts domain-specific information from the analysis of a set of protein families.

The system has been tested with different protein families, of which three examples are discussed in detail in the paper: 'ataxia-telangiectasia associated protein', 'ran GTPase' and 'carbonic anhydrase'. We found generally good correlation between the amount of information provided to the system and the quality of the annotations. The current limitations and future developments of the system are discussed.

Relativistic heavy ion physics is of international and interdisciplinary interest to nuclear physics, particle physics, astrophysics, condensed matter physics and cosmology. The primary goal of this field of research is to re-create in the laboratory a novel state of matter, the quark-gluon plasma (QGP), which is predicted by the standard model of particle physics (Quantum Chromodynamics) to have existed ten millionths of a second after the Big Bang (origin of the Universe) and may exist in the cores of very dense stars.

STAR searches for signatures of quark-gluon plasma formation and investigates the behavior of strongly interacting matter at high energy density by focusing on measurements of hadron production over a large solid angle. It utilizes a large volume Time Projection Chambers (TPC) for tracking and particle identification in a high track density environment. STAR will measure many observables simultaneously on an event-by-event basis to study signatures of a possible QGP phase transition and the space-time evolution of the collision process at their respective energy. The goal is to obtain a fundamental understanding of the microscopic structure of hadronic interactions, at the level of quarks and gluons, at high energy densities.

STAR is one of two large-scale experiments under construction at the Relativistic Heavy Ion Collider (RHIC) at the Brookhaven National Laboratory (BNL) on Long Island (New York) for operation in 1999. It has been designed to focus primarily on hadronic observables and features a large acceptance for high precision tracking and momentum analysis at center of mass (c.m.) rapidity. Specific to RHIC will be: significantly increased particle production (thousands of particles produced); hard parton-parton scattering in heavy ion collisions.

Figure 2. Seriously academic texts.

Table 7. *Coverage of the rump AWL (288 word families) in medical texts*

BNC-med sections	Number of words	1k + 2k coverage (%)	Rump AWL coverage (%)
1	25,289	72.39	2.44
2	21,300	84.53	2.05
3	19,247	90.80	1.45
4	18,777	84.62	1.51
Mean	21,153	83.08	1.86
SD	2.966	7.70	0.47

I interpret these coverage figures to mean that Coxhead's original AWL research can be usefully replicated in the new BNC framework, adopting the methodology of the earlier project (possibly incorporating suggestions from the subsequent discussions of Eldridge, 2008; Hyland & Tse, 2007; and Granger & Pacquot, 2010). What difference would it make to establish an AWL in the new framework? If there is group of perhaps 500 high coverage mid-frequency sub-technical academic word families lurking somewhere between the second and eighth 1,000 zones, then discovering these words might well reduce the learning investment needed for academic reading in English as a second language well below 8,000 word families, a wish shared by many Lextutor users. We may be within sight of the real 2k list and the real AWL.

Conclusion

The examples could be multiplied, but perhaps the point is made. Nation has set running a highly productive set of ideas and idea-generating tools that, via the social possibilities of the Web, many teachers, researchers and even learners can participate in developing and adapting.

More broadly, the combination of computer programming and empirical research presents itself as a potent medium for exploring language and language learning. Many of the questions raised by Lextutor users did not have obvious answers, and playing with the code has generated both answers and new questions. It was not obvious what texts look like to learners with mixed profiles, or how many proper nouns can be accounted for by an algorithm. New questions like whether a new AWL can once again reduce the learning burdens of academic reading will again be solved through a blend of empirical and computational research.

More broadly still, the enterprise of exploring language processing with computer programs is hardly new in language study. Outstanding cases of its application in language development include a learning-based account of vocabulary bursts (Elman et al., 1997); a simulation of past-tense learning without recourse to abstract rules (Rumelhart & McClelland, 1986); and a proof of associative learning between non-adjacent items (Ellis & Schmidt, 1997). Without these and many other concrete demonstrations of what is computationally and hence logically possible, useless debates about what language is and how it can be learned could have gone on forever.

But there exist more elemental versions of this agenda employing vastly simpler computer programs that can still play transformational roles in how we represent and understand language at a practical level. One example is the discovery of the sheer degree of repetition that exists in natural language, by, for example, Sinclair (1990), which has changed our whole view of “what the language looks like” to a learner and also has put to rest some other less useful accounts. Similarly, Nation has used computation to search for and locate manageable zones of recurrence amid the oceans of lexis that confront language learners. The prescription that “language learning is largely lexical learning” was basically a banishment of language pedagogy to a hopeless regime of item learning, unless someone could find a way to impose system and learnability on a vast learning task. No one has done more to achieve this than Paul Nation.

References

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*, 273–279.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*, 498–520.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*, 38–63.

Cobb, T., & Horst, M. (2004). Is there room for an AWL in French? In P. Bogaards & B. Laufer, (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15–38). Amsterdam: John Benjamins.

Collins, L., Trofimovich, P., White, J., Cardoso, W., & Horst, M. (2009). Some input on the easy/difficult grammar question. *The Modern Language Journal*, *93*, 336–353.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*, 213–238.

de Maupassant, G. (1880). *Boule de suif* [A. McMaster, Trans.]. Retrieved from http://www.gutenberg.org/files/3077/3077.txt

Eldridge, J. (2008). "No, there isn't an 'academic vocabulary,' but . . . ." A reader responds to K. Hyland and P. Tse's "Is there an 'academic vocabulary'?" *TESOL Quarterly*, *42*, 109–113.

Ellis, N., & Schmidt, R. (1997). Morphology and longer distance dependencies: Laboratory research illuminating the A in SLA. *Studies in Second Language Acquisition*, *19*, 145–171.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1997). *Rethinking innateness.* Cambridge, MA: MIT.

Gass, S., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3rd ed.). New York: Taylor & Francis.

Granger, S., & Pacquot, M. (2010). *In search of a general academic vocabulary: A corpus-driven study.* Manuscript in preparation. Retrieved from http://cecl.fltr.ucl.ac.be /.../In_search_of_a_general _academic_english.pdf

Hazenberg, S., & Hulstijn, J. (1996). Defining a minimal receptive second language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics*, *17*, 145–163.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range [Computer software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*, 689–696.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430.

Hyland, K. & Tze, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*, 235–253.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Clevedon, UK: Multilingual Matters.

Laufer, B. (2000). Task effect on instructed vocabulary learning: The hypothesis of "involvement." *Selected papers from AILA '99* (pp. 47–62). Tokyo: Waseda University Press.

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.

Lawrence, D. H. (1929). Lady Chatterley's Lover (online version). Retrieved from http://gutenberg.net.au/ebooks01/0100181h.html

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus.* London: Longman.

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Bristol, England: Multilingual Matters.

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* London: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63,* 59–81.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nation, I. S. P., & Heatley, A. (1994). Range: A program for the analysis of vocabulary in texts [software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Ovtcharov, V., Cobb, T., & Halter, R. (2006). La richesse lexicale des productions orales: Mesure fiable du niveau de compétence langagière. *The Canadian Modern Language Review, 63,* 107–125.

Oxford Computing Services (2005). *British National Corpus Samplers.* Retrieved from http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=products

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2. Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press / Bradford Books.

Schmitt, N., Jiang, X., & Grabe, W. (2010). *The percentage of words known in a text and reading comprehension*. Manuscript submitted for publication.

Schmitt, N., & Schmitt, D. (2005). *Focus on vocabulary: Mastering the Academic Word List.* London: Longman.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18,* 55–88.

Sinclair, J. M. (1991). *Corpus, concordance, collocation.* London: Oxford University Press.

Webb, S., & Rodgers, M. (2009). Vocabulary demands of television programs. *Language Learning, 59,* 335–366.

West, M. (1953). *A general service list of English words.* London: Longman.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning & Communication, 32,* 215–219.

Appendix A

From Range's (Heatley, Nation, & Coxhead, 2002) *Itemization of Proper Nouns*

Wallingford
Wallington
Wallis
Wallsend
Wally
Walpole
Walsall
Walsh
Walsingham
Walt
Walter
Walters
Waltham
Walthamstow
Walton
Walworth
Wanda
Wandsworth
Wang
Wantage
Wapping
Waqar
Warburg
Warburton
Wardle
Warenne
Warhol
Wark
Warley
Warne

Warner
Warnie
Warnock
Warrington
Warsaw
Warton
Warwick
Warwickshire
Washburn
Washington
Wasim
Watanabe
Waterford
Watergate
Waterhouse
Waterloo
Waterman
Waterstone
Watford
Watkin
Watkins

Appendix B

Proper Nouns in Lady Chatterley Identified by VP Algorithm

Chatterley
Lawrence
Constance
Clifford
Flanders
Wragby
Chatterleys
Malcolm
Reid
Fabians
Raphaelite
Hilda
Florence
Hague
Berlin
Wandervogel
Connie
Kensington
Cambridge
Westminster
Bonn
Kitchener
Geoffrey
Emma
Herbert
Tommies
Ab
Ovo
Lloyd
George
Horatio
Bottomley
Tevershall
Sussex
Trent
Mester
Ashby
Michaelis
Dublin
Mayfair
Sheffield
Renoir
Cezanne
Streety
Une
Esperance
La
Terre
Mick
Tommy
Charles
Hammond
Julia
Charlie
Arnold
Hors
De
Socrates
Plato
Alcibiades
Cathedra
Bolshevists
Bolshevist
Hades
Hildebrand
Mansfield
Mellors
Betty
Martin
Charlestoned
Methuselah
Th
Henry
James
Jerusalem
Autre
Betts
Eva
Bennerley
Winterslow
Strangeways
Carrara
Du
Coeur
Cannes
Biarritz
Sicily
Bolton
Shardlow
Leiver
Ted
Edith
Uthwaite
Ev
Morn
Persephone
Absalom
John
Juno
Derbyshire
Wor
Gaskell
Eliot
Mitford
Allsopp
Pye
Bestwood
Kinbrook
Willcock
Harrison
Mary
Edgar
Thompson
Pally
Charlestons
Doncaster
Nottingham
Mikado
Whiteover
Colwick
Madrid
Frankfurt
Linley
Leslie
Shipley
Edward
Marehay
Josephine
Luke
Balaam
Bacchante
Bacchae
Francis
Racine
Bolshevistic
Oliver
Bertha
Coutts
Geoffery
Edwin
Landseers
William
Shortlands
Ess
Weedon
Fillingwood
Alexander
Cooper
Esmeralda
Venice
Gare
Nord
Calais
Sam
Nelson
Matlock
Warsop
Coningsby
Chadwick
Anne
Tom
Jones
Weatherleys
Eastwood
Victoria
Sandringham
Bentley
Cavell
Cappadocia
Lipton
Tres
Flossie
Brutus
Thysen
Maupassant
Thout
Royce
Jupiter
Neptune
Wragbys
Shipleys
Nero
Jenny
Papp
Dieppe
Schieber
Proust
Guerre
Comme
Ollerton
Butterley
Persepolis
Timbuctoo
Birmingham
Duckfoot

Sautes
Thomas
Jane
Gloire
Dijon
Retford
Colemans
Grantham
Adam
Da
Fe
Te
Deum
Laudamus
Joe
Jonah
Moses
Jinny
Aristotle
Minerva
Athena
Crosshill
Bolsover
Yorkshire
Abelard
Heloise
Columbia
Wae
Londonized
Flaneurs
Bois
Luxembourg
Brenner
Bernard
Lucerne
Tyrol
Mestre
Giovanni
Costanza
Chioggia
Scotchman
Contessa
Guthries
Edinburgh
Lind
Lucchese
Florian
Goldoni
Daniele
Apollo
Duncan
Forbes
Dan
Beggarlee
Fred
Kirk
Phillips
Guthrie
Coty
Couttses
Benvenuto
Cellini
Joan
Rabelais
Crippen
Dee
Sade
Rodrigo
Finley
Burroughs
Judith
Stewart
Coburg
Hartland
Cliffords
Berthas
Charing
Dahomey
York
Hauteur
Vulcan
Carrington
Lecky
Tennyson
Magna
Mater
Boue
Heanor
Richards
Smitham
Herefords
Notts
Pentecost
Juan

Appendix C

Proper Nouns in English Translation Boule De Suif, Identified by Algorithm

Rouen
Pont
Audemer
Bourg
Achard
Darnetal
Boisguillaume
De
Ville
Norman
Croisset
Dieppedalle
Biessart
Havre
Dieppe
Normandie
Loiseau
Carre
Lamadon
Comte
Comtesse
Hubert
Breville
Normandy
Henry
Orleanist
Nantes
Louis
Philippe
Brevilles
Cornudet
Boule
Suif
Tantalus
Dieu
Rubicon
Crassane
Leveque
Napoleon
Bonapartist
Du
Follenvie
Elisabeth
Rousset
Loiseaus
Orleans
Guesclin
Joan
Godforsaken
Yvetot
Holofernes
Lucrece
Sextus
Cleopatra
Hannibal
Capua
Abraham
Crimea
Gad
Lamadons
Gruyere
Marseillaise
Conduis
Liberte

Appendix D

The Off-List Items From English Translation of Boule De Suif, Proper Nouns Having Been Previously Identified by Algorithm

abashed
adepts
amour
areally
aregence
artillerymen
avec
aves
banditti
baser
benumbed
bezique
braggart
breeched
capered
catherine
chandlers
cherie
complaisance
consumptive
cudgeling
cupful
cur
dainties
daybreak

debauchees
defenseurs
demeanor
despotic
devastations
easygoing
ecarte
effectually
etrelles
execrating
faultlessly
foie
forethought
foundered
freemasonry
gnawings
gras
hearers
hoarfrost
imaginings
indignation
induced
ins
intrenchments
intrusted
invalids
irregulars
iv
judith
kissable
la
larded
lunching
maneuvers
mesdames
mustache
mustachios
needful
nicephore
nogg
noiseless
outcry
outdistance
overstrained
partridges
paternosters
patrie
paunched
pawing
pleasanter
pocketknife
potbellied
prefecture
repast
repeopling
reproached
repugnance
sacre
savior
seamed
sha
sheepskin
shipwrecked
skillfully
snowdrift
soldiery
solicitations
somber
soutiens
stertorous
supposition
tenfold
tes
timeworn
trente
uhlans
uncorking
undersized
unedifying
uninterruptedly
unreasoning
untiring
unutterably
vapor
vengeurs
visionaries
warmers
whereat
witticisms

About the Author

Tom Cobb teaches courses in computer development for TESL trainees at the Université du Québec à Montréal and at McGill University in Montreal. He has previously worked in and directed reading and writing programs in British Columbia, Hong Kong, and The Arabian Gulf. He recently gave a graduate course in CALL development with Paul Nation at Victoria University of Wellington, New Zealand. E-mail: cobb.tom@uqam.ca