Reading in a Foreign Language April 2022, Volume 34, No. 1
ISSN 1539-0578 pp. 165–171

Counting Affixes with Morpholex: A Response to McLean and Stoeckel (2021)

Tom Cobb
Université du Québec à Montréal
Canada

Some words in English are derived words, i.e., words with an affix that changes their part of speech, like *know (v)→knowable* (adj), or meaning, like *pleasant (adj) → unpleasant (adj)*. As a reading teacher, I had always believed that derived words were not particularly problematic for my English as a second language (ESL) and foreign language (EFL) learners, owing to the recurrence of a relatively small number of different derivational affixes in what they were reading. Their problem seemed rather not knowing enough words in any form. The opposing belief, held by researchers like McLean (2017), is that the roughly 120 derivational affixes of English are inscrutable to their learners, even if they know the base words they are attached to, and thus present a significant obstacle to reading comprehension. Such a disparity of beliefs, held by people with similar experience of learners, texts, and affixes and a professional interest in the matter, shows the need for an objective way of counting affixes such as text analysis computer software might contribute to.

Laufer and Cobb (2020) conducted an empirical investigation into this question, for which I developed Morpholex v.3.1, a computer program that assesses the morphological load of an entered text. We could not find another program that performs this task. The v.2 version of Morpholex was developed to break down system-supplied word family lists and colour-code their affixations; the v.3 series extended this operation to user-supplied texts and counts up the affixes; and v.3.5, the current version, incorporates six months of feedback from users running the program on the Web following publication of the paper. The output has two main parts, the proportion of base, inflected, and derived words in a text, and the cumulative token coverage each provides. The tool is described in the paper and can be accessed at https://lextutor.ca/cgi-bin/morpho/lex.

After we had analyzed a variety of texts (academic and narrative, authentic and simplified) and calculated the number and percentage of base words and affixes in each, we found that most of the derived words in texts were formed with a limited number of affixes. Across a range of text types, just 10 affixes accounted for 86% of all derivational affixations. This is hardly surprising, being a standard instance of Zipf's Law, whereby a small number of words, affixes, or other language features are used a lot and a large number used little. In works of fiction intended for native speakers, just two derivations, typically *~er* and *~ly*, plus base words, inflections, and proper nouns, are typically sufficient to reach 95% lexical coverage, the coverage point several studies have shown corresponds to basic comprehension (e.g., Laufer, 2020).

Morpholex v.3.2 was recently reviewed in *Reading in a Foreign Language* by McLean and

Stoeckel (2021) as part of a response to Laufer (2021) in the same volume. The review raised some useful points, chief among them that Morpholex underestimates the number of derivational morphemes in a given text by counting only once words that in fact bear two affixes, whether an inflectional and a derivational (like *teachers → teach + er + s*, counted only as an inflection while it is also a derivation) or two derivational (like *remorselessly → remorse + less + ly*, counted only as one derivation rather than two). The criticism is legitimate; it was based on our decision to put program development on hold while the paper was in review and in press. Then, following publication it seemed right that, for a period, readers should be able to run the same texts through the program and 'get the same answers.' In any event, we felt that our research had arrived at a point where it was clear further refinements would make little difference to the basic finding or would strengthen it. The proportion of derived forms in texts is quite reliable, in a range from 5% in literary texts (whether graded or not) to 8% in some academic texts, the majority in either case falling in the category of 'regular affixations' (Level 3 of the Bauer and Nation, 1993, scheme that we used as our affix list) and are about 10 in number (including *~ly, ~able, un~,* and *~ness)* and involve little change to the base words they are attached to.

It is now more than a year since publication, during which a follow-up piece has been produced (Cobb & Laufer, 2021), Morpholex has collected megabytes of user data, and program development has resumed—partly thanks to nudges from McLean and Stoeckel and other critics.

No full-language/any-text routine like Morpholex will ever be 100% accurate, owing mainly to the unruly nature of English morphology. Derivational affixes consist of words, former words, foreign words, pieces of words, and single letters attached to base words which may or may not have shape-shifted, whether orthographically, phonologically, or both, to accommodate them. These affixes are unlike those of the inflectional system (*~ing, ~ed,* etc.) which are fewer in number, as types, are attached to their roots in predictable ways, and have typically been the focus of much drill and practice in the ESL grammar class. In comparison, the derivational system is a Wild West show, particularly in English. As Prasada and Pinker (1993, p. 55) put it, "No language uses rule-like productivity nearly as much as it could" and English uses it considerably less. A particular problem in coding and counting derivational affixes is that few apparent affixes actually are so in most of the words they appear in. Of 10,000 words ending in *~er* in a corpus of graded stories, just 974 really are affixes; the rest are coincidental copies of the same form (*dinner, rather, bother,* etc) and must be blocked from appearing as derivations. Finding and creating these blocks is a significant task, and while frequency lists can be used for some of it, it will mainly come from encounters with individual texts. The program's output was thus made fully explicit, with a view to sharing out the work. McLean and Stoeckel (2021) indicate some needed exclusions thrown up by the program, like day (*da + y*), and question (*quest + ion*), which, despite comprising properly formed suffixes attached to plausible base words, are in fact not derived forms, though only a human English speaker would know it. These join the hundreds of other misclassified derivations reported by users over 100,000+ runs of the program in 2020-2021, which are now incorporated in an ever-growing exceptions file in Morpholex v.3.5. There will always be more exceptions to trap for, many of them in conflict with others, which is why the paper stated an error rate (p. 976; 1% in v.3.2, down to .05% in v.3.5). But, as mentioned, the trend is towards fewer derived forms, not more.

Another line of program development spurred by McLean and Stoeckel and others' responses involves the ongoing expansion of the database of plausible root words to which affixes may legally be attached. The database began life as an unstructured list of all the individual words in Nation's (2012) BNC/COCA 25 thousand-family word lists, comprising 105,630 items, though it is larger now, used as follows: following an affix-stripping model, Morpholex subtracts potential affixes from words and then checks whether what remains is present in the database. If so, the word is classed as an affixed form; if not, it either happens to include an affix-like letter string (as in *ceiling* and *rather*) or is a bound morpheme (without identifiable base word, as in *imperative* and *dissent*) and either way is classed as a base word despite the apparent affixation. For example, the program comes to *comprehension*, strips *~ion*, checks whether *comprehens* or even *prehens* is present in the database, and finding they are not, counts *comprehension* as a base word that cannot be further reduced. McLean and Stoeckel dispute *comprehension* being classed a base word and want it classed instead as a derived word. The idea raises interesting questions about program development and analogies with language acquisition in humans.

The affix-stripping model of v.3.1 was rulebound and somewhat rigid. A degree of rigidity seemed useful for our particular investigation, which involved hypothetical learners who knew base words but were nonplussed when certain affixes were added to them. The base words such learners would know are presumably the canonical versions of words, as opposed to the large number of variants that affixation might impose, so a fixed collection of standard base words seemed to model this situation. Standard base words are analogous to the '*wugs'* devised by Berko (1958) to study morphological development in children. It could be argued, however, that non-child L2 learners familiar with *comprehend* might well be able to see *comprehens* in it, where a child or binary-minded computer program would not. So should *comprehens* be added to the database and *comprehension* classed a derived word? It depends whether *comprehens* can be judged recognizable to a learner who would know the canonical version, which is an empirical question, or more likely a professional judgment, that cannot be made for all learners or for all *~ion* suffixes. (It is interesting that McLean and Stoeckel assume that learners who cannot cope with regular affixations like *know → know + able* will easily see the relationship between *comprehend* and *comprehens*, surely a more strenuous feat.)

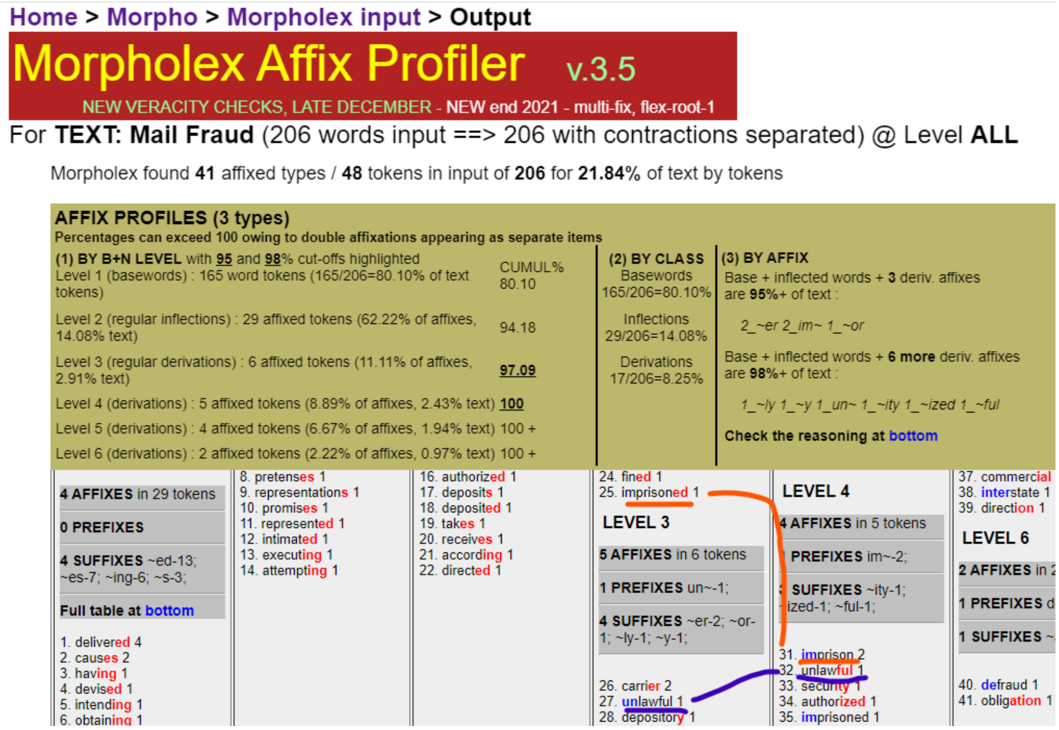
A consistent approach to modifying the base words database was clearly needed and here is what we have come up with so far. With each run of the program, Morpholex culls a list of all the words it has classed as base words that nonetheless bear a legitimate derivational affix at either end. The list is sorted by frequency after each run of the program with recurring items at the top. *Comprehension* has appeared near the top of this list throughout the gap year, possibly as a result of its recurrence in one of the test texts used in the paper and linked to the interface (Laufer & Ravenhorst-Kalovski, 2010). The accumulated list is periodically inspected, and decisions are made by native speakers about its recurring items. The rule of thumb that has evolved is that high frequency variants with just one letter changed from the canonical are provisionally added to the list of base words. Thus Morpholex v.3.5 now includes *comprehens* as a base word or stem to which affixes may legally be appended, which upgrades *comprehension, comprehensive* and *comprehensible* to the status of derived forms, while *retention, sanction* and other *~ion* items remain base words. And so on, one at a time, for the many candidates thrown up over many runs of the program.

System-level refinements have also been made to the program, chief among them a procedure for dealing with more than one layer of affixation, particularly inflections upon derivations such as occurs in words like *teachers,* as noted by McLean and Stoeckel (2021). Thus, after every act of affix-stripping, Morpholex now adds the remainder item to the end of the text where it will be processed anew like any other word in the text. For example, the word *teachers* is classed an inflection, the ~*s* stripped, and *teacher* appended to the end of the list; in the next loop *teacher* is classed a derivation and *teach* is appended for another round; *teach* itself, however, does not end up in the base words column, unless it appears in the text in that form. This procedure will clearly increase the percentage of derived forms, though not much, and differently in different types of texts. In 10,000 running words of *The Great Gatsby* (used in our study and linked from the interface), derived forms increase from 4.49% of the text's lexis to 5.33%, or 94 additional words, about one per 100 words of text. In graded stories it is closer to one per 1,000. So although this change to the algorithm was probably worth making, it is mainly for better face validity and does not affect our basic argument: the number of different derived forms in the texts that learners are likely to read remains low, recurrent, and pedagogically manageable.

What is 'pedagogically manageable' in this context? What proportion of derived words can learners be expected to cope with as they read? If as mentioned basic comprehension is associated with knowledge of 95% of a text's words, then the number of derived forms needed to reach 95% is the relevant target and Morpholex duly provides a computation of this number. The number of derivational affixes needed to reach 95% text coverage ranges from zero to 2-3 in the texts we have run and re-run so far. For example, the demonstration text in the introduction to Laufer and Cobb (2020, p. 978) was a short (205-word) but for illustration purposes morphology-heavy Wikipedia entry on Mail Fraud. Despite its morphology, however, just three derivational affixes were needed to account for 95% of its words. Figure 1 shows program output for this text under the updated rules in Morpholex v.3.5, with two items appearing in two categories which had previously appeared in just one: *unlawful* and *imprisoned*, the latter appearing as an inflection (~*ed*) and subsequently as a derivation (*im*~). Despite these additions and some minor reshuffling of the elements, however, the morphology burden of Mail Fraud is unchanged: 95% of its words are comprehensible to a learner who knows its base words, their inflections, and three derivational affixes.

Figure 1

Multiple Affixes and New Word Counts in v.3.5



Home > Morpho > Morpholex input > Output

Morpholex Affix Profiler v.3.5

NEW VERACITY CHECKS, LATE DECEMBER - NEW end 2021 - multi-fix, flex-root-1

For **TEXT: Mail Fraud** (206 words input ==> 206 with contractions separated) @ Level **ALL**

Morpholex found **41** affixed types / **48** tokens in input of **206** for **21.84%** of text by tokens

AFFIX PROFILES (3 types)

Percentages can exceed 100 owing to double affixations appearing as separate items

(1) BY B+N LEVEL with **95** and **98**% cut-offs highlighted

	CUMUL%
Level 1 (basewords) : 165 word tokens (165/206=80.10% of text tokens)	80.10
Level 2 (regular inflections) : 29 affixed tokens (62.22% of affixes, 14.08% text)	94.18
Level 3 (regular derivations) : 6 affixed tokens (11.11% of affixes, 2.91% text)	**97.09**
Level 4 (derivations) : 5 affixed tokens (8.89% of affixes, 2.43% text)	**100**
Level 5 (derivations) : 4 affixed tokens (6.67% of affixes, 1.94% text)	100 +
Level 6 (derivations) : 2 affixed tokens (2.22% of affixes, 0.97% text)	100 +

(2) BY CLASS
Basewords 165/206=80.10%
Inflections 29/206=14.08%
Derivations 17/206=8.25%

(3) BY AFFIX
Base + inflected words + **3** deriv. affixes are **95%**+ of text :

2_~er 2_im~ 1_~or

Base + inflected words + **6 more** deriv. affixes are **98%**+ of text :

1_~ly 1_~y 1_un~ 1_~ity 1_~ized 1_~ful

Check the reasoning at bottom

4 AFFIXES in 29 tokens

0 PREFIXES

4 SUFFIXES ~ed-13; ~es-7; ~ing-6; ~s-3;

Full table at bottom

1. delivered 4
2. causes 2
3. having 1
4. devised 1
5. intending 1
6. obtaining 1

8. pretenses 1
9. representations 1
10. promises 1
11. represented 1
12. intimated 1
13. executing 1
14. attempting 1

16. authorized 1
17. deposits 1
18. deposited 1
19. takes 1
20. receives 1
21. according 1
22. directed 1

24. fined 1
25. imprisoned 1

LEVEL 3

5 AFFIXES in 6 tokens

1 PREFIXES un~-1;

4 SUFFIXES ~er-2; ~or-1; ~ly-1; ~y-1;

26. carrier 2
27. unlawful 1
28. depository 1

LEVEL 4

4 AFFIXES in 5 tokens

PREFIXES im~-2;

SUFFIXES ~ity-1; ized-1; ~ful-1;

31. imprison 2
32. unlawful 1
33. security 1
34. authorized 1
35. imprisoned 1

37. commercial
38. interstate 1
39. direction 1

LEVEL 6

2 AFFIXES in 2

1 PREFIXES d

1 SUFFIXES ~

40. defraud 1
41. obligation 1

But Mail Fraud is a short and possibly atypical text, chosen to fit a small page in a research journal and possibly less than ideal to generalize from. It is therefore convenient that as part of their review McLean and Stoeckel (2021) undertook a hand reanalysis of a longer paper that was also used in our study (Laufer & Ravenhorst-Kavloski, 2010, with 7,454 words including appendices) which could be more seriously impacted by whether doubly affixed words are counted once or twice. This particular text contains a large token count of such words (though a small type count, mainly *teachers* → *teach* + *er* + *s* and *learners* → *learn* + *er* + *s*), so adding these to the profile as both inflections and derivations should substantially increase the number of the latter, which is possibly why it was chosen. But increased to what effect?

The critics' re-analysis does not include a coverage calculation, but since Morpholex 3.5 incorporates most of their suggestions, perhaps they will accept ours. The v.3.1 analysis of Laufer and Ravenhorst-Kavloski's paper found that just one derivational suffix was needed to reach 95% text coverage for learners who knew the text's base words and inflections (namely the suffix *~age*, repeated 121 times in a paper mainly about coverage and comprehension). In v.3.5, instead of one derivation, three derivations are now needed to reach the same coverage (122 instances of *~age*, 101 of *~er*, and 139 of *~ion*, with *comprehension* now added to the line-up). The point, however, is the same. Only a small number of different derivational suffixes need be known (or taught or reviewed) to enable basic comprehension of the text. And this is despite the fact that the total number of words in the text has not been increased to reflect double entries. This particular text is academic in nature, of course, and beyond the scope of many classroom

ESL/EFL learners, who are more likely to be cutting their teeth on graded readers and course books. In *Lord Jim*, from Oxford Bookworms collection (also provided on the interface), just one derivational suffix (~*ly*) is needed for 95% coverage; in v.3.1, none were needed. In an ungraded work of fiction, *The Great Gatsby,* just one derivational affix is needed to reach 95% coverage (again ~*ly*), unchanged from v.3.1.

McLean and Stoeckel (2021) could hardly be expected to be aware of these issues in program development and are right to say what they see, thereby playing a possibly unwitting role in moving the process forward. Some of their criticisms are less helpful, however, and may show a less than careful reading of our paper, Laufer and Cobb (2020). These include:

- assembling program errors out of context to suggest they are more prominent than they really are (despite our explicit provision of an error rate);

- suggesting but not showing that such errors would undermine our claim about the points at which 95% and 98% coverage are reached (admittedly doing so would be difficult without some sort of code-and-count software);

- criticizing our use of smallish demonstration texts (while our stated purpose was precisely to use such texts to counteract 'the corpus effect' by which infrequent affixations appear more prominent than they really are in particular texts or genres);

- claiming not to understand how a word like *comprehension* could be classed as a base word (when the inclusion criteria for base words had been clearly explained);

- stating that we 'did not report the individual words in which each affix appeared' (p. 255, whereas Morpholex does precisely that, with colour coding (see Figure 1);

- re-tweeting our words with added connotations, such as our describing derivational affixes as 'rare' (p. 257; while we had merely suggested they are few enough to be manageable);

- claiming rather inscrutably that Laufer and Cobb (2020) had 'assumed [that] knowledge of only the most frequently occurring derivational affixes in a given text is sufficient as an estimate of the total number of affixes a learner needs to know to understand the text' (p. 256). Come again? We did not assume anything, we rather worked from the finding of numerous studies (e.g., Laufer, 2020) showing that knowledge of 95% a text's base words along with the most frequent affixes *are* 'the total number of affixes a learner needs' in the sense that it provides a sufficient basis to make inferences about the rest of the words whether affixed or not.

But we take our feedback where we find it. Neither Google, nor corpus linguistics, nor the commercial software industry will give us the tools we need to answer the questions we have, so we create them ourselves as we are able. The interesting questions that MacLean, Stoeckel, and others have raised can only be answered by a combination of empirical and computational research. Few studies bring these together – Laufer and Cobb (2020) was an attempt to do it in the context of morphology and reading.

Note

All texts referred to are linked from the Morpholex input page.

References

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2–3), 150–177. doi:10.1080/00437956.1958.11659661.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*, 253–279. https://doi.org/10.1093/ijl/6.4.253.

Cobb, T., & Laufer, B. (2021). The Nuclear Word Family List: A list of the most frequent family members, including base and affixed words. *Language Learning*, *71*(3), 834–871.

Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly, 54*, 1076–1085. https://doi.org/10.1002/tesq.3004

Laufer, B. (2021). Lexical thresholds and alleged threats to validity: A storm in a teacup? *Reading in a Foreign Language, 33*, 238–246.

Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text-coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*, 15–30.

Laufer, B., & Cobb, T. (2020). How much knowledge of derived words is needed for reading? *Applied Linguistics*, *41*(6), 971–998. https://doi.org/10.1093/applin/amz051

McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*, 823–45.

McLean, S., & Stoeckel, T. (2021). Lexical mastery thresholds and lexical units: A reply to Laufer. *Reading in a Foreign Language*, *33*(2), 247-259.

Nation, P. (2012). The BNC/COCA word family lists. Retrieved from https://www.wgtn.ac.nz/__data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf.

Prasada, S. & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*(1), 1–56.

About the Author

Tom Cobb has been a reading teacher, program coordinator, teacher of reading teachers, and developer of reading software in several countries over many years. His software is assembled at the Lextutor website (www.lextutor.ca) and offers a range of takes on data-driven language learning and course/curriculum design. His principal interest at present is solving puzzles in the reading research with text analysis. In breaks from research and development, he helps UNESCO share competency-based learning principles with educators in developing countries.