



Metrics for Investigations into L2 Knowledge of Derivational Affixes

Dale Brown *Kanazawa University***Phil Bennett** *University of Niigata Prefecture***Geoffrey Pinchbeck** *Carleton University*

Abstract

Knowledge of derivational affixes makes an important contribution to second language learners' success when reading. Yet while the effects of some learner variables (L2 proficiency, L1 background) have been investigated, there has been little research addressing the effects of varying characteristics of affixes on their acquisition. The goal of this study was to develop a range of metrics concerning the characteristics of derivational affixes with respect to their frequency of occurrence, semantic salience, and orthographic and phonological form. The study presents 19 metrics (58 when including variants) for 38 derivational affixes (12 prefixes; 26 suffixes) with a range of frequencies. Each metric is calculated across progressively larger vocabulary size levels in recognition of the fact that as learners' vocabulary knowledge develops, their exposure to and knowledge of words including derivational affixes grows. Examples of a selection of metrics for one affix are provided (the full data set being available online; <https://osf.io/2vcg9/>) as well as some global observations on the data set. It is hoped that these metrics will allow future analyses that provide insights into the process of L2 derivational affix acquisition (by exploring which metrics and to what degree the metrics contribute to acquisition) as well as insights into the order in which affixes are learnt and at what stage in L2 development different affixes are acquired.

Background

Derivational affix knowledge is considered important in L2 vocabulary studies. Teaching affixes has been reported to be a well-established classroom fixture (Ur, 2022), receiving some attention in L2 textbooks from intermediate proficiency levels (Brown, 2011; Vu & Michel, 2021), and has been promoted as a key vocabulary learning strategy (Nation, 2001). Such knowledge is part of morphological awareness, “the ability to reflect upon and manipulate morphemes and employ word formation rules” (Kuo & Anderson, 2006, p. 161). This might include recognising roots or bases (whether bound or free) in morphologically complex words and the relationship between roots/bases and affixes, the impact of affixes on part of speech and word meaning, and the category(ies) of base(s) an affix may attach to.

Morphological awareness has been found to significantly predict the reading comprehension of L1 children (Liu et al., 2024) and L2 learners (Jeon, 2011; Keiffer & Lesaux, 2012; Zhang, 2016) even after other predictors (e.g., lexical/phonological knowledge) are controlled for. More specifically, morphological awareness has a direct effect on reading comprehension as well as indirect effects via its influence on other aspects of knowledge such as vocabulary (Levesque et al., 2021).

There is evidence that knowledge of affixes develops alongside overall L2 proficiency and vocabulary knowledge (Iwaizumi & Webb, 2021, 2022; Mochizuki & Aizawa, 2000; Raudszus et al., 2021). The L1 background of learners also seems to influence learning. Zhang and Koda (2014) report that learners whose L1 has different word formation patterns from their L2 have greater difficulty developing L2 morphological awareness due to reduced positive transfer. Moreover, Stewart et al. (2023) found that the difficulty of specific affixes varied notably among learners from three L1 backgrounds and via L1–L2 contrastive analysis described features that may influence this.

The characteristics of individual affixes also likely influence their acquisition. Sasao and Webb (2017) report that affixes grouped into three levels of difficulty based on test data differ in terms of type frequency (i.e., how many different words an affix appears in) and token frequency (i.e., how many times an affix appears in a corpus). However, many other item variables concerning affix frequency, semantics, and form (both orthographic and phonological) could have an impact on when and how well particular affixes are acquired.

The metrics developed in this study were intended to allow greater insights into these issues. It is our hope that this work may in turn feed into the development of new descriptions of affix learning/acquisition, such as Bauer and Nation’s (1993) well-known classification of affixes. The metrics were developed in the context of a broader ongoing study in which data are being collected on L1-Japanese learners’ knowledge of 1,001 target words. Accordingly, some of the metrics reflect the Japanese context (e.g., JAPANESE LOANWORDS, Metric 9), but the majority should have some general applicability.

Methods

Fifty affixes (22 prefixes; 28 suffixes) were selected through a sampling procedure which resulted in all highly frequent affixes being included along with a number of less frequent affixes (see Appendix A for further details).

The selection of metrics was informed by literature on morphology (concerning factors affecting the order of acquisition of inflectional affixes, the productivity of derivational affixes [i.e., whether an affix tends to be used in the formation of new words], and the degree to which derivational forms may be parsed as combinations of morphemes) and on statistical language learning mechanisms (concerning how different aspects of frequency may have effects on affix acquisition); see Appendix C for specific rationales. The metrics are categorised according to their focus: frequency, semantics, and orthographic or phonological form.

Frequency measures were drawn from a lemma list compiled by the third author based on a corpus of subtitles for “family-viewing” TV and movies. The raw English OPUS 2018 corpus (Lison & Tiedemann, 2016; Tiedemann, 2018) was cleaned and filtered to identify files that were (i) originally produced in English and (ii) classified as “Family” in the “Genre” attribute on the Internet Movie Database (IMDB, n.d.) The focus on this genre follows previous research (Pinchbeck et al., 2022) which found strong correlations between word frequency rank in a corpus of this genre and the word knowledge of Japanese learners of English. The corpus files were tagged with TreeTagger via TagAnt (Anthony, 2014; Schmid, 1994), and lemma-based word frequency lists generated using AntConc (Anthony, 2018). The metrics utilise the first 11,000 lemmas in this list. This threshold was adopted since it was found that this figure provided 95% coverage of various genres, which tallies with figures in Vilkaitė-Lozdienė and Schmitt (2020). The 11,000 lemmas in this list were coded for the presence of the 50 affixes (see Appendix B for details) to serve as the basis for the metrics. At this stage, it was found that 12 of the affixes occurred in fewer than 10 lemmas in the list, so metrics were not developed for these affixes. That is, in this study, metrics were developed and described for 38 affixes (12 prefixes; 26 suffixes).

The metrics were calculated at successively larger vocabulary sizes. That is, we calculated each metric on the basis of the first 1,000 most frequent lemmas in the list, then on the first 2,000 most frequent lemmas (cumulatively), and so on. This is because as learner vocabulary knowledge grows, they gain more experience with affixes.

Several metrics provided here refer to *basewords* and *precursor* words. The term *baseword* is used to refer to a morphologically simplex relative of an affixed form. For example, *fame* is the baseword of *famous*. We use the term *precursor* to refer to a relative of an affixed form which is not morphologically simplex itself but does not contain the particular affix that affixed form includes. For example, when the affix of interest is *-ly*, *famous* is a precursor of *famously*, where the baseword is *fame*. There are two reasons for making this distinction. Firstly, many words do not have a baseword that is a stand-alone form (i.e., a free morpheme) in English. For example, we code the word *predict* as containing the affix *pre-*, leaving only the Latin form *dict*. This is problematic when considering how learners might deal with the word *predictable*. Rather than saying that there is no baseword for this word, our position is that knowledge of the word *predict* could assist learners, so we allow for precursor words in our metrics. Secondly, affixed forms can have infrequent basewords but frequent precursors. One example is the word *emotionally*, for which the precursors *emotional* and *emotion* are likely to be more recognisable than the baseword *emote*.

All metrics that involve the baseword of affixed lemmas also use precursor words. For example, SEMANTIC TRANSPARENCY (Metric 7) concerns whether there is

a transparent semantic link between an affixed lemma and its base¹ form. Codings for this metric were based on semantic transparency between the affixed form and either a baseword or a precursor (see Appendix C for details on the method). Thus, for the affix *-able*, the baseword/affixed form pair *accept–acceptable* was coded as exhibiting a transparent semantic link, as was the precursor–affixed form pair *predict–predictable*, while the baseword–affixed form pair *consider–considerable* was not.

There are several versions of some metrics:

- **Types/tokens:** One version is based on the number of different lemmas with a given feature (i.e., a types count) and another on the number of tokens with the feature (i.e., the summed frequency of all types with the feature).
- **Counts/proportions:** One version is based on the number of lemmas with a given feature and another on the proportion of lemmas with that feature. For example, for SEMANTIC TRANSPARENCY (Metric 7), at the ~1000 vocabulary size, there are four lemmas featuring the affix *-able*, only one of which exhibits semantic transparency, producing a count of 1 and a proportion of 0.25. Count- and proportion-based versions are often crossed with type and token versions so that there are four versions of a metric. Differing effects for type/token frequencies and the importance of proportions are discussed in Hay and Baayen (2002).
- **Alternative approaches:** Sometimes, more than one approach to computing a metric was identified in the literature. For example, base–affix boundary phonotactics were calculated via two methods: LOW PROBABILITY PHONOTACTICS (Metrics 16a–d) and FORWARDS/BACKWARDS TRANSITIONAL PROBABILITY OF BASE–AFFIX BOUNDARY (Metrics 16e–h). Another example of alternative approaches is ORTHOGRAPHIC TRANSPARENCY (Metric 11). Here, there are versions with a strict interpretation of whether a baseword’s orthographic form is preserved in the lemma containing the target affix and more lenient versions that allow for common, regular changes in orthography, such as the truncation of final *-e* (e.g., *adore–adorable*).

Table 1 provides a description of each metric. Details on the rationale and method of operationalisation for each metric are provided in Appendix C, with worked examples of metric calculations in Appendix D.

Results

The metrics are publicly available at the following Open Science Framework site: <https://osf.io/2vcg9/>

With 19 metrics (or 58 when including different versions) for 38 affixes, each of which is calculated across 11 vocabulary sizes, the affix metrics constitute an extremely rich data set. Here, we present a small selection of metrics for one affix by way of example (Table 2) and provide some global observations on the data set.

Table 2 shows three metrics for the affix *-able*. The first metric, AFFIX TYPES – NUMBER (Metric 1), shows how many lemmas were coded as including *-able* at each vocabulary size. That is, within the first 1,000 lemmas, 4 included *-able*, within the

Table 1 *Metric Descriptions*

<i>Metrics focused on frequency</i>	
Metric name	Description
AFFIX TYPES – NUMBER (Metric 1)	Number of different lemmas (i.e., lemma types) featuring the affix within a given vocabulary size.
AFFIX TYPE–TOKEN RATIO (Metric 2)	Number of affix types divided by number of affix tokens within a given vocabulary size.
RELATIVE ENTROPY AMONG AFFIX TYPES (Metric 3)	Evenness of the frequency distribution of types featuring the affix within a given vocabulary size.
FORM–FUNCTION CONTINGENCY (Metric 4)	Number of affix types divided by number of apparent affix types within a given vocabulary size.

<i>Metrics focused on semantics</i>	
Metric name	Description
KNOWN BASEWORDS/ PRECURSORS – PROPORTION (Metric 5)	Proportion of affix types that have a baseword or precursor within a given vocabulary size.
SEMANTIC VARIABILITY – NUMBER (Metric 6)	Number of senses an affix has within the affix types within a given vocabulary size.
SEMANTIC TRANSPARENCY – TYPES – NUMBER (Metric 7a)	Number of affix types for which there is a transparent semantic link between the affix type and its baseword or precursor within a given vocabulary size.
SEMANTIC TRANSPARENCY – TYPES – PROPORTION (Metric 7b)	Proportion of affix types for which there is a transparent semantic link between the affix type and its baseword or precursor within a given vocabulary size.
SEMANTIC TRANSPARENCY – TOKENS – NUMBER (Metric 7c)	Number of affix tokens for which there is a transparent semantic link between the affix type and its baseword or precursor within a given vocabulary size.
SEMANTIC TRANSPARENCY – TOKENS – PROPORTION (Metric 7d)	Proportion of affix tokens for which there is a transparent semantic link between the affix type and its baseword or precursor within a given vocabulary size.
MORPHOLOGICAL FAMILY SIZE – TYPES – MEAN (Metric 8a)	Average morphological word family size of the affix types within a given vocabulary size, calculated across the affix types.

(Continued)

<i>Metrics focused on semantics</i>	
Metric name	Description
MORPHOLOGICAL FAMILY SIZE – TOKENS – MEAN (Metric 8b)	Average of the summed frequencies of each affix type's word family members within a given vocabulary size.
JAPANESE LOANWORDS – TYPES – NUMBER (Metric 9a)	Number of affix types within a given vocabulary size that are loanwords.
JAPANESE LOANWORDS – TYPES – PROPORTION (Metric 9b)	Proportion of affix types within a given vocabulary size that are loanwords.
JAPANESE LOANWORDS – TOKENS – NUMBER (Metric 9c)	Number of affix tokens within a given vocabulary size that are loanwords.
JAPANESE LOANWORDS – TOKENS – PROPORTION (Metric 9d)	Proportion of affix tokens within a given vocabulary size that are loanwords.
DECOMPOSED FORMS – TYPES – NUMBER (Metric 10a)	Number of affix types within a given vocabulary size that are less frequent than their baseword or precursor.
DECOMPOSED FORMS – TYPES – PROPORTION (Metric 10b)	Proportion of affix types within a given vocabulary size that are less frequent than their baseword or precursor.
DECOMPOSED FORMS – TOKENS – NUMBER (Metric 10c)	Number of affix tokens within a given vocabulary size that are less frequent than their baseword or precursor.
DECOMPOSED FORMS – TOKENS – PROPORTION (Metric 10d)	Proportion of affix tokens within a given vocabulary size that are less frequent than their baseword or precursor.

<i>Metrics focused on form (orthographic or phonological)</i>	
Metric name	Description
ORTHOGRAPHIC TRANSPARENCY, STRICT, OF BASEWORD OR PRECURSOR – TYPES – NUMBER (Metric 11a)	Number of affix types within a given vocabulary size for which the baseword or precursor is strictly preserved in the affix type.
ORTHOGRAPHIC TRANSPARENCY, STRICT, OF BASEWORD OR PRECURSOR – TYPES – PROPORTION (Metric 11b)	Proportion of affix types within a given vocabulary size for which the baseword or precursor is strictly preserved in the affix type.

(Continued)

<i>Metrics focused on form (orthographic or phonological)</i>	
Metric name	Description
ORTHOGRAPHIC TRANSPARENCY, STRICT, OF BASEWORD OR PRECURSOR – TOKENS – NUMBER (Metric 11c)	Number of affix tokens within a given vocabulary size for which the baseword or precursor is strictly preserved in the affix type.
ORTHOGRAPHIC TRANSPARENCY, STRICT, OF BASEWORD OR PRECURSOR – TOKENS – PROPORTION (Metric 11d)	Proportion of affix tokens within a given vocabulary size for which the baseword or precursor is strictly preserved in the affix type.
ORTHOGRAPHIC TRANSPARENCY, LENIENT, OF BASEWORD OR PRECURSOR – TYPES – NUMBER (Metric 11e)	Number of affix types within a given vocabulary size for which the baseword or precursor is preserved in the affix type allowing minimal changes: (i) truncation of final <i>-e</i> ; (ii) shifts from <i>-y</i> to <i>-i</i> or <i>-ie</i> ; (iii) doubling of final consonants.
ORTHOGRAPHIC TRANSPARENCY, LENIENT, OF BASEWORD OR PRECURSOR – TYPES – PROPORTION (Metric 11f)	Proportion of affix types within a given vocabulary size for which the baseword or precursor is preserved in the affix type allowing minimal changes.
ORTHOGRAPHIC TRANSPARENCY, LENIENT, OF BASEWORD OR PRECURSOR – TOKENS – NUMBER (Metric 11g)	Number of affix tokens within a given vocabulary size for which the baseword or precursor is preserved in the affix type allowing minimal changes.
ORTHOGRAPHIC TRANSPARENCY, LENIENT, OF BASEWORD OR PRECURSOR – TOKENS – PROPORTION (Metric 11h)	Proportion of affix tokens within a given vocabulary size for which the baseword or precursor is preserved in the affix type allowing minimal changes.
ORTHOGRAPHIC TRANSPARENCY OF AFFIX – TYPES – NUMBER (Metric 12a)	Number of affix types within a given vocabulary size in which the affix occurs in canonical orthographic form (rather than as an allomorph or having undergone changes due to co-occurrence with other affixes).
ORTHOGRAPHIC TRANSPARENCY OF AFFIX – TYPES – PROPORTION (Metric 12b)	Proportion of affix types within a given vocabulary size in which the affix occurs in canonical orthographic form.
ORTHOGRAPHIC TRANSPARENCY OF AFFIX – TOKENS – NUMBER (Metric 12c)	Number of affix tokens within a given vocabulary size in which the affix occurs in canonical orthographic form.
ORTHOGRAPHIC TRANSPARENCY OF AFFIX – TOKENS – PROPORTION (Metric 12d)	Proportion of affix tokens within a given vocabulary size in which the affix occurs in canonical orthographic form.

(Continued)

<i>Metrics focused on form (orthographic or phonological)</i>	
Metric name	Description
PHONOLOGICAL TRANSPARENCY OF BASEWORD OR PRECURSOR – TYPES – NUMBER (Metric 13a)	Number of affix types within a given vocabulary size in which the baseword or precursor's phonological form is preserved in the affix type.
PHONOLOGICAL TRANSPARENCY OF BASEWORD OR PRECURSOR – TYPES – PROPORTION (Metric 13b)	Proportion of affix types within a given vocabulary size in which the baseword or precursor's phonological form is preserved in the affix type.
PHONOLOGICAL TRANSPARENCY OF BASEWORD OR PRECURSOR – TOKENS – NUMBER (Metric 13c)	Number of affix tokens within a given vocabulary size in which the baseword or precursor's phonological form is preserved in the affix type.
PHONOLOGICAL TRANSPARENCY OF BASEWORD OR PRECURSOR – TOKENS – PROPORTION (Metric 13d)	Proportion of affix tokens within a given vocabulary size in which the baseword or precursor's phonological form is preserved in the affix type.
PHONOLOGICAL TRANSPARENCY OF AFFIX – TYPES – NUMBER (Metric 14a)	Number of affix types within a given vocabulary size in which the affix is in its canonical phonological form.
PHONOLOGICAL TRANSPARENCY OF AFFIX – TYPES – PROPORTION (Metric 14b)	Proportion of affix types within a given vocabulary size in which the affix is in its canonical phonological form.
PHONOLOGICAL TRANSPARENCY OF AFFIX – TOKENS – NUMBER (Metric 14c)	Number of affix tokens within a given vocabulary size in which the affix is in its canonical phonological form.
PHONOLOGICAL TRANSPARENCY OF AFFIX – TOKENS – PROPORTION (Metric 14d)	Proportion of affix tokens within a given vocabulary size in which the affix is in its canonical phonological form.
AFFIX SONORITY – TYPES – MEAN (Metric 15a)	Average sonority of the affix across affix types within a given vocabulary size.
AFFIX SONORITY – TOKENS – MEAN (Metric 15b)	Average sonority of the affix across affix tokens within a given vocabulary size.
LOW PROBABILITY PHONOTACTICS, WITHIN-WORD – TYPES – MEAN (Metric 16a)	Average probability of the affix boundary phonotactics, based on within-word syllable boundaries within the monomorphemic types within a given vocabulary size.
LOW PROBABILITY PHONOTACTICS, BETWEEN-WORD – TYPES – MEAN (Metric 16b)	Average probability of the affix boundary phonotactics, based on between-word boundaries within the monomorphemic types within a given vocabulary size.

(Continued)

Metrics focused on form (orthographic or phonological)

Metric name	Description
LOW PROBABILITY PHONOTACTICS, WITHIN-WORD – TOKENS – MEAN (Metric 16c)	Average probability of the affix boundary phonotactics, based on within-word syllable boundaries within the monomorphemic tokens within a given vocabulary size.
LOW PROBABILITY PHONOTACTICS, BETWEEN-WORD – TOKENS – MEAN (Metric 16d)	Average probability of the affix boundary phonotactics, based on between-word boundaries within the monomorphemic tokens within a given vocabulary size.
FORWARDS TRANSITIONAL PROBABILITY OF BASE-AFFIX BOUNDARY – TYPES (Metric 16e)	Average forwards transitional probability for the phoneme combinations at the base–affix boundary across affix types, based on the transitional probabilities of phoneme combinations within the monomorphemic types within a given vocabulary size.
BACKWARDS TRANSITIONAL PROBABILITY OF BASE-AFFIX BOUNDARY – TYPES (Metric 16f)	Average backwards transitional probability for the phoneme combinations at the base–affix boundary across affix types, based on the transitional probabilities of phoneme combinations within the monomorphemic types within a given vocabulary size.
FORWARDS TRANSITIONAL PROBABILITY OF BASE-AFFIX BOUNDARY – TOKENS (Metric 16g)	Average forwards transitional probability for the phoneme combinations at the base–affix boundary across affix tokens, based on the transitional probabilities of phoneme combinations within the monomorphemic tokens within a given vocabulary size.
BACKWARDS TRANSITIONAL PROBABILITY OF BASE-AFFIX BOUNDARY – TOKENS (Metric 16h)	Average backwards transitional probability for the phoneme combinations at the base–affix boundary across affix tokens, based on the transitional probabilities of phoneme combinations within the monomorphemic tokens within a given vocabulary size.
ORTHOGRAPHIC AFFIX LENGTH – TYPES – MEAN (Metric 17a)	Average affix length in letters across affix types within a given vocabulary size.
ORTHOGRAPHIC AFFIX LENGTH – TOKENS – MEAN (Metric 17b)	Average affix length in letters across affix tokens within a given vocabulary size.
PHONOLOGICAL AFFIX LENGTH – TYPES – MEAN (Metric 18a)	Average affix length in phonemes across affix types within a given vocabulary size.

(Continued)

Metrics focused on form (orthographic or phonological)	
Metric name	Description
PHONOLOGICAL AFFIX LENGTH – TOKENS – MEAN (Metric 18b)	Average affix length in phonemes across affix tokens within a given vocabulary size.
AFFIX AS INITIAL/FINAL MORPHEME – TYPES – NUMBER (Metric 19a)	Number of affix types within a given vocabulary size in which the affix is the initial morpheme (for prefixes) or the final morpheme (for suffixes).
AFFIX AS INITIAL/FINAL MORPHEME – TYPES – PROPORTION (Metric 19b)	Proportion of affix types within a given vocabulary size in which the affix is the initial morpheme (for prefixes) or the final morpheme (for suffixes).
AFFIX AS INITIAL/FINAL MORPHEME – TOKENS – NUMBER (Metric 19c)	Number of affix tokens within a given vocabulary size in which the affix the initial morpheme (for prefixes) or the final morpheme (for suffixes).
AFFIX AS INITIAL/FINAL MORPHEME – TOKENS – PROPORTION (Metric 19d)	Proportion of affix tokens within a given vocabulary size in which the affix is the initial morpheme (for prefixes) or the final morpheme (for suffixes).

first 2,000 lemmas, 10 included *-able*, and so on. In the case of *-able*, the cumulative number of lemmas including the affix climbs steadily across the vocabulary sizes. The second metric shown, SEMANTIC TRANSPARENCY – TYPES – NUMBER (Metric 7a), concerns whether the meaning of an affixed form can be understood via knowledge of its baseword/precursor and knowledge of the affix. Thus, just one of the four lemmas within the first 1,000 lemmas that include *-able* is semantically transparent. This metric, too, increases steadily across vocabulary sizes. The final metric illustrated here, AFFIX AS INITIAL/FINAL MORPHEME – TYPES – PROPORTION (Metric 19b), shows the proportion of lemmas featuring an affix in which the affix is word-initial (for prefixes) or word-final (for suffixes). At the high-frequency vocabulary levels, *-able* (a suffix) is the final morpheme in approximately three quarters of lemmas (cf. *possible* [word-final] and *probably* [non-word-final]), while it is word-final in approximately five sixths of cases when lower frequency level lemmas are included.

The complexity and richness of the data set makes a general summary challenging. In an attempt to do so, Figure 1 shows the mean and SD for each affix calculated across all the metrics. Since different metrics operate across different scales, these figures were calculated after first converting each set of metric values at a given vocabulary size to a z-score. Thus, this figure illustrates the metric values for the affixes relative to each other.

The highest global means are for *-ly* (adverbs) and *-er* (nouns). This in part reflects the fact that these are the most frequently occurring affixes. However, frequency alone

Table 2 Three Metrics for the Affix -able

Affix	Vocabulary size	AFFIX TYPES – NUMBER (Metric 1)	SEMANTIC TRANSPARENCY – TYPES – NUMBER (Metric 7a)	AFFIX AS INITIAL/FINAL MORPHEME – TYPES – PROPORTION (Metric 19b)
-able	~1000	4	1	0.750
-able	~2000	10	3	0.700
-able	~3000	17	6	0.765
-able	~4000	22	7	0.727
-able	~5000	24	8	0.750
-able	~6000	27	10	0.778
-able	~7000	30	11	0.800
-able	~8000	35	13	0.829
-able	~9000	39	15	0.846
-able	~10000	44	19	0.864
-able	~11000	51	23	0.824

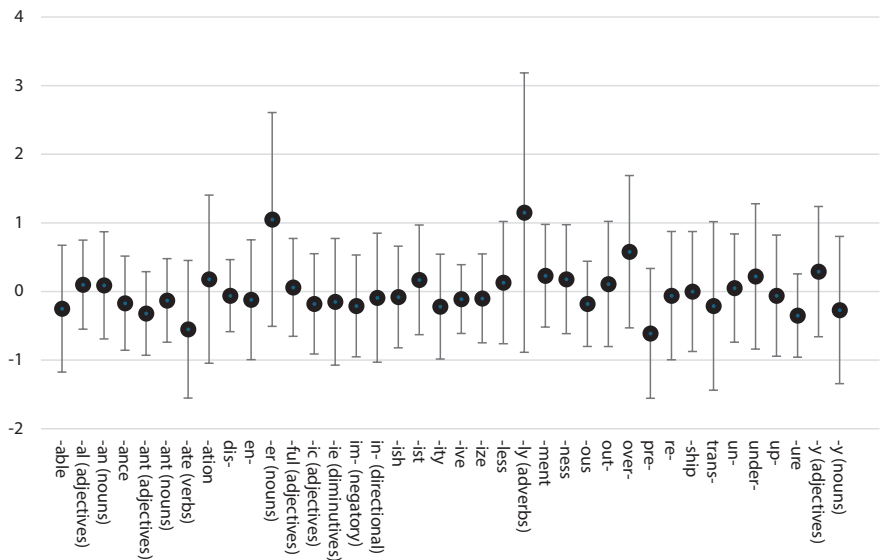


Figure 1 Global Means and SD (z-scores) for each Affix Calculated Across all Metrics and all Vocabulary Size Levels.

Note. Means are for metric values converted to z-scores. The z-scores were calculated based on the metric values for each affix at each specific vocabulary size. That is, the values across the 38 affixes for a given metric at the ~1000 vocabulary size level were converted to z-scores, the values across the affixes for that metric at the ~2000 vocabulary size level converted to z-scores, and so on. Each set of z-scores thus indicated, at a given vocabulary size, metric values for the affixes relative to each other. For each affix, the mean value of all z-scores at a given vocabulary size was then calculated (e.g., the mean of the 58 z-scores for -able at the ~1000 vocabulary size level). Finally, the mean value for each affix across the 11 vocabulary size levels was calculated.

does not determine these global means. The third highest mean is for *over-* despite it occurring in a relatively small number of lemmas (none at all within the first 4,000 lemmas). This might indicate that, once it is encountered, *over-* has many characteristics that make it somewhat easier to acquire.

The lowest global means are for the affixes *pre-* and *-ate* (verbs). The former occurs in relatively few lemmas (though in a similar number to *over-*), while the latter is quite frequent. Both, however, have low values across many metrics, including those concerning frequency, semantics, and form.

Conclusion

The affix metrics introduced here allow a variety of analyses to be conducted. These analyses may focus on the metrics themselves (i.e., investigating which metrics are of greater or lesser importance for the acquisition of affixes) or on the affixes (i.e., investigating which affixes are likely to be relatively easier or more difficult to learn). Thus, in addition to the metrics themselves, the OSF site for this project also includes tables and graphs intended to facilitate analyses following these two approaches.

Importantly, the metrics allow a dynamic view of such questions since they were calculated across successively larger vocabulary sizes. That is, when it comes to learners' acquisition of affixes, what is important is not how many words feature a given affix in the language as a whole, but rather how many such words a learner knows or has been exposed to. Thus, the metrics allow dynamic investigations of how different factors may influence acquisition as learners' experience of the language increases and their vocabulary knowledge develops. This, to our knowledge, is a unique feature of the metrics and one that we hope will help advance understanding of knowledge and acquisition of English affixes.

Note

1. Here and elsewhere in the text, the term *base* is used, after Bauer, Lieber and Plag (2013), to refer to "any morphological element to which other elements are added in the creation of words" (p. 18). That is, *base* is an umbrella term covering both *baseword* and *precursor* as defined in the previous paragraph.

References

- Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.
- Allen, D. (2020). A procedure for determining Japanese loanword status for English words. *Vocabulary Learning and Instruction*, 9(1), 33–50. <https://doi.org/10.7820/vli.v09.1.Allen.a>
- Anthony, L. (2014). *TagAnt* (Version 1.1) [Computer software]. Waseda University. <https://www.laurenceanthony.net/software>
- Anthony, L. (2018). *AntConc* (Version 3.58) [Computer Software]. Waseda University. <https://www.laurenceanthony.net/software>
- Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford reference guide to English morphology*. Oxford University Press.

- Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279. <https://doi.org/10.1093/ijl/6.4.253>
- Becker, T. (1993). Back-formation, cross-formation, and ‘bracketing paradoxes’ in paradigmatic morphology. In G. Booij & J. Van Marle (Eds.), *Yearbook of morphology 1993* (pp. 1–25). Kluwer Academic Publishers.
- Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language*, 42(3), 390–405. <https://doi.org/10.1006/jmla.1999.2681>
- Brown, D. (2011). What aspects of vocabulary knowledge do textbooks give attention to? *Language Teaching Research*, 15(1), 83–97. <https://doi.org/10.1177/1362168810383345>
- Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant’s age. *Frontiers in Psychology*, 7, 1–11. <https://doi.org/10.3389/fpsyg.2016.01116>
- Bybee, J. (2006). *Frequency of use and the organization of language*. Oxford University Press.
- Coltheart, M. (1981). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4), 497–505. <https://doi.org/10.1080/14640748108400805>
- Ford, M. A., Davis, M. H., & Marslen-Wilson, W. D. (2010). Derivational morphology and base morpheme frequency. *Journal of Memory and Language*, 63(1), 117–130. <https://doi.org/10.1016/j.jml.2009.01.003>
- Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the “natural order of L2 morpheme acquisition” in English: A meta-analysis of multiple determinants. *Language Learning*, 51(1), 1–50. <https://doi.org/10.1111/1467-9922.00147>
- Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sanchez & M. Almela (Eds.), *A mosaic of corpus linguistics: Selected approaches* (pp. 269–291). Peter Lang.
- Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, 65(S1), 228–255. <https://doi.org/10.1111/lang.12119>
- Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, 39(6), 1041–1070. <https://doi.org/10.1515/ling.2001.041>
- Hay, J. (2003). *Causes and consequences of word structure*. Routledge.
- Hay, J., & Baayen, H. (2002). Parsing and productivity. In G. E. Booij, & J. Van Marle (Eds.), *Yearbook of morphology 2001* (pp. 203–235). Kluwer Academic Publishers.
- Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences*, 9(7), 342–348. <https://doi.org/10.1016/j.tics.2005.04.002>
- IMDb. (n.d.). *IMDb—Movies, TV and celebrities*. IMDb. Retrieved June 11, 2018, from <http://www.imdb.com/>
- Iwaizumi, E., & Webb, S. (2021). To what extent does productive derivational knowledge of adult L1 speakers and L2 learners at two educational levels differ? *TESOL Journal*, 12, Article e640. <https://doi.org/10.1002/tesj.640>
- Iwaizumi, E., & Webb, S. (2022). Measuring L1 and L2 productive derivational knowledge: How many derivatives can L1 and L2 Learners with differing vocabulary

- levels produce? *TESOL Quarterly*, 56(1), 100–129. <https://doi.org/10.1002/tesq.3035>
- Jarmulowicz, L., & Taran, V. L. (2013). Lexical morphology: Structure, process, and development. *Topics in Language Disorders*, 33(1), 57–72. <https://doi.org/10.1097/TLD.0b013e318280f5c0>
- Jeon, E. H. (2011). Contribution of morphological awareness to second-language reading comprehension. *The Modern Language Journal*, 95(2), 217–235. <https://doi.org/10.1111/j.1540-4781.2011.01179.x>
- Kieffer, M. J., & Lesaux, N. K. (2012). Direct and indirect roles of morphological awareness in the English reading comprehension of native English, Spanish, Filipino, and Vietnamese speakers. *Language Learning*, 62(4), 1170–1204. <https://doi.org/10.1111/j.1467-9922.2012.00722.x>
- Kuo, L., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist*, 41(3), 161–180. https://doi.org/10.1207/s15326985ep4103_3
- Kuperman, V., & Van Dyke, J. (2011). Individual differences in visual comprehension of morphological complexity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33, 1643–1648.
- Laufer, B., & McLean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly*, 13(3), 202–217. <https://doi.org/10.1080/15434303.2016.1210611>
- Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.
- Levesque, K. C., Bredmore, H. L., & Deacon, S. H. (2021). How morphology impacts reading and spelling: Advancing the role of morphology in models of literacy development. *Journal of Research in Reading*, 44(1), 10–26. <https://doi.org/10.1111/1467-9817.12313>
- Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. <https://opus.nlpl.eu/OpenSubtitles2018.php>
- Liu, Y., Groen, M., & Cain, K. (2024). The association between morphological awareness and reading comprehension in children: A systematic review and meta-analysis. *Educational Research Review*, 42, Article 100571. <https://doi.org/10.1016/j.edurev.2023.100571>
- Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, 28(2), 291–304. [https://doi.org/10.1016/S0346-251X\(00\)00013-0](https://doi.org/10.1016/S0346-251X(00)00013-0)
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
- Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, 80(3), 674–685. <https://doi.org/10.1111/j.1467-8624.2009.01290.x>
- Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A., & Rayson, P. (2015, May 31–June 5). *Development of the Multilingual Semantic Annotation System* [Paper presentation].

- Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, Denver, CO, United States.
- Pinchbeck, G. G., Brown, D., McLean, S., & Kramer, B. (2022). Validating word lists that represent learner knowledge in EFL contexts: The impact of the definition of word and the choice of source corpora. *System*, 106, 102771. <https://doi.org/10.1016/j.system.2022.102771>
- Raudszus, H., Segers, E., & Verhoeven, L. (2021). Use of morphological and contextual cues in children's lexical inferencing in L1 and L2. *Reading and Writing*, 34(6), 1513–1538. <https://doi.org/10.1007/s11145-021-10122-z>
- Rogers, J., Webb, S., & Nakata, T. (2015). Do the cognacy characteristics of loanwords make them more easily learned than noncognates? *Language Teaching Research*, 19(1), 9–27. <https://doi.org/10.1177/1362168814541752>
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926–1928. <https://doi.org/10.1126/science.274.5294.1926>
- Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35(4), 606–621. <https://doi.org/10.1006/jmla.1996.0032>
- Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, 50, 1568–1580. <https://doi.org/10.3758/s13428-017-0981-8>
- Sasao, Y., & Webb, S. (2017). The Word Part Levels Test. *Language Teaching Research*, 21(1), 12–30. <https://doi.org/10.1177/1362168815586083>
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* (pp. 44–49).
- Stewart, J., Brown, D., Bennett, P., Robles-García, P., Sánchez-Gutiérrez, C. H., Miguel, N. M., Vitta, J. P., Nicklin, C., Stoeckel, T., & McLean, S. (2023). The contribution of affixes to productive English vocabulary knowledge for Chinese, German and Spanish learners: A comparison. *System*, 115, 103035.
- Tiedemann, J. (2018). *OpenSubtitles 2018 search interface* [Corpus resource webpage]. OPUS. <https://opus.nlpl.eu/OpenSubtitles-v2018.php>
- Ur, P. (2022). How useful is it to teach affixes in intermediate classes? *ELT Journal*, 76(3), 330–337. <https://doi.org/10.1093/elt/ccac018>
- Vilkaitė-Lozdienė, L., & Schmitt, N. (2020). Frequency as a guide for vocabulary usefulness. In S. Webb (Ed.). *The Routledge handbook of vocabulary studies* (pp. 81–96). Routledge.
- Vu, D. V., & Michel, M. (2021). An exploratory study on the aspects of vocabulary knowledge addressed in EAP textbooks. *Dutch Journal of Applied Linguistics*, 10. <https://doi.org/10.51751/dujal9345>
- Willis, M., & Ohashi, Y. (2012). A model of L2 vocabulary learning and retention. *The Language Learning Journal*, 40(1), 125–137. <https://doi.org/10.1080/09571736.2012.658232>

- Zhang, D. (2016). Derivational morphology in reading comprehension of Chinese-speaking learners of English: A longitudinal structural equation modeling study. *Applied Linguistics*, 38(6), Article 871895. <https://doi.org/10.1093/applin/amv072>
- Zhang, D., & Koda, K. (2012). Awareness of derivation and compounding in Chinese–English biliteracy acquisition. *International Journal of Bilingual Education and Bilingualism*, 17(1), 55–73. <https://doi.org/10.1080/13670050.2012.736949>

Appendix A

Sampling of Affixes

Affixes were sampled for the study as follows.

- Step 1: Estimates of affix type frequencies made use of the MorphoLex database (Sánchez-Gutiérrez et al., 2018). This database provides codings for the prefix, suffix, and root elements for nearly 80,000 English words.
- Step 2: The codings in MorphoLex are based on form. Therefore, the following manual adjustments were made to the codings.
- Step 2a: Different affixes that share an orthographic form were given separate codings. Examples included: (i) adjective-forming *-en* and verb-forming *-en*, (ii) adverb-forming *-ly* and adjective-forming *-ly*, and (iii) noun-forming *-th* and ordinal number-forming *-th*. To determine the proportion of words in the MorphoLex database that feature each affix, a random sample of 100 types was analysed. For example, there are 2,932 words coded as containing *-ly* in the MorphoLex database. In a random sample of 100 of these words, 94 were instances of adverb-forming *-ly* and 6 adjective-forming *-ly*. Thus, it was estimated that the MorphoLex database contains 2,756 words ($2,932 \times 0.94$) with adverb-forming *-ly* and 176 words ($2,932 \times 0.06$) with adjective-forming *-ly*.
- Step 2b: Affixes with allomorphic variants were consolidated into a single coding represented by the canonical form. This included cases where the orthographic form of an affix was changed due to the presence of another affix. Examples included: (i) Adjective-forming *-able* (canonical form) can appear as *-ible*, as well as *-ab* or *-ib* when followed by *-ly*; (ii) Agent noun-forming *-er* (canonical form) can appear as *-or*, *-yer*, *-ier*, or *-ior*; (iii) Negatory prefix *in-* (canonical form) can appear as *im-*, *il-*, or *ir-*. In these cases, the MorphoLex database counts for the variants were summed to produce a single count.
- Step 2c: Codings that required an understanding of etymology or allomorphy were based on Bauer et al. (2013) where possible and supplemented with online resources (<https://www.etymonline.com> and <https://www.affixes.org/index.html>).
- Step 3: From these adjusted codings, 50 target affixes were sampled using systemic sampling with unequal probabilities. This method allowed the type frequency of an affix to affect its likelihood of being selected in the sample, which ensured that the highest frequency affixes would be included along with a selection of mid- and low-frequency affixes.

As mentioned in the paper, after coding the 11,000 lemmas in the TV/movie subtitles word list for the presence of the 50 target affixes (see Appendix B: Affix codings), it was found that 12 occurred in fewer than 10 lemmas in the lemmas list, so metrics were not developed for these affixes (i.e., metrics were developed for 38 affixes).

Appendix B

Coding of Affixes

Procedures for coding words for affixes were as follows.

- Step 1: A corpus of subtitles from TV and movies classified as “family genre” was used as the source of words and frequency measures. Previous research by Pinchbeck et al. (2022) found that there were strong correlations between Japanese learners’ word knowledge and word frequency rankings in a TV/movie subtitles corpus.
- Step 2: Affix codings were made for the first 11,000 lemmas in the family viewing TV/movie subtitles corpus. Figures cited by Vilkaitė-Lozdienė and Schmitt (2020) suggested that 11,000 lemmas would provide 95% coverage of various genres of English. Our own calculations using the list of 11,000 lemmas and looking at various samples from COCA sub-corpora produced similar findings.
- Step 3: The 11,000 lemmas were coded for the presence of 50 target affixes by two raters, both L1 speakers of English with backgrounds in Applied Linguistics. The presence of affixes was determined using the following criteria:
 - Step 3a: If an affix was etymologically present and attached to a free baseword, and the meaning of the affixed form was similar to that of the baseword, then the affix was coded as present. For example, *reasonable* was coded as containing the suffix *-able* on the baseword *reason*.
 - Step 3b: If an affix was etymologically present and attached to a bound root, and other words containing the same bound root existed (akin to Becker’s [1993] cross-formations), then the affix was coded as present. For example, *caution* and *cautious* were coded as containing *-ion* and *-ous*, with *caut* treated as a bound root.
 - Step 3c: If an affix was etymologically present and attached to a free baseword, but the meaning of the derived form was dissimilar to that of the baseword, then the affix was coded as not present. For example, the word *hospitality* was not tagged as containing the suffix *-ity* as its meaning is dissimilar to *hospital*.
 - Step 3d: If an affix was etymologically present and attached to a bound root, but no other words containing the same bound root existed, then the affix was coded as not present. For example, the word *vicinity* was not coded as containing the suffix *-ity*.
 - Step 3e: If the affix was not etymologically present, then the affix was coded as not present. For example, the word *unite* was not coded as containing the prefix *un-*.
 - Step 3f: When determining the presence of a baseword, common allomorphs were permitted. For example, cases where the final *-e* was deleted (e.g., *nature-natural*), the final consonant was doubled (e.g., *fun-funny*), or the final *-y*

was changed to *-i* (e.g., *carry-carrier*) were still considered to represent free basewords (see Bauer et al., 2013). Similarly, allomorphs of the target affixes were treated as variations of the canonical form (i.e., *-ious* and *-itious* were treated as forms of *-ous*).

Step 4: Words coded as not containing a target affix but which contained an orthographic string matching that of the affix (allomorphs permitted) in an appropriate position (i.e., prefixes before basewords or other elements and suffixes after basewords or other elements), were coded as “Apparent affixes” (see Metric 4).

Example 1: In the word *real*, the prefix *re-* and suffix *-al* were coded as apparent affixes.

Example 2: In the word *energy*, the prefix *en-* and diminutive-forming suffix *-y* were coded as apparent affixes (with the noun-forming suffix *-y* coded as a genuine affix).

Step 5: 12 of the sampled affixes (*-acy*, *-ify*, *co-*, *de-*, *di-*, *ex-* *micro-*, *mis-*, *pro-*, *semi-*, *sub-*, and *mono-*) were found to occur in fewer than 10 lemmas, so were removed from the sample.

Appendix C

Metrics Rationale and Methodology

Metric name	Rationale	Method
AFFIX TYPES – NUMBER (Metric 1)	The number of types featuring a given linguistic feature is thought to be a key driver of acquisition. See Bybee (2006); Bertram, Baayen and Schreuder (2000).	A count was made of lemma types containing each affix at successively larger vocabulary sizes (from 1k only to 1–11k).
AFFIX TYPE–TOKEN RATIO (Metric 2)	Word forms that are frequent are more likely to have developed meanings that are independent of their composition (i.e., non-componential semantics/semantic autonomy) (Bybee, 2006). A lower type–token ratio may indicate that the types featuring an affix have undergone this process and are thus less accessible examples of the affix.	The number of affix types at a given vocabulary size (Metric 1) was divided by the total number of affix tokens at each vocabulary size.
RELATIVE ENTROPY AMONG AFFIX TYPES (Metric 3)	Word forms that are frequent are more likely to have developed meanings that are independent of their composition (i.e., non-componential semantics/semantic autonomy) (Bybee, 2006). Relative entropy is a measure of the frequency distribution of items of a particular type (Gries & Ellis, 2015). Lower relative entropy (i.e., less evenness among the affix types) could be an indicator of greater semantic variability within the affix types, making the meaning of the affix more difficult to generalize.	This metric indicates the evenness of the frequency distribution of the types featuring the affix within a given vocabulary size. Based on Gries (2010): 1. The proportion of the total number of affix tokens at each vocabulary size accounted for by each lemma type was calculated. 2. These proportions were multiplied by the natural log of each proportion. 3. The sum of (2) for each vocabulary size was divided by the natural log of the number of affix types (Metric 1).
FORM–FUNCTION CONTINGENCY (Metric 4)	Form–function contingency indicates how consistently a linguistic form signals a linguistic function, with greater consistency likely making learning easier (Gries & Ellis, 2015).	The number of affix types (Metric 1) was divided by the number of apparent affix types (see Step 4 in Appendix B: Coding of affixes) within a given vocabulary size.
KNOWN BASEWORDS/ PRECURSORS – PROPORTION (Metric 5)	An affix is more likely to be recognised in a word if a baseword or precursor related to that word is already known.	The number of affix types that have a baseword or precursor within a given vocabulary size was divided by the number of affix types at the given vocabulary size (Metric 1).

SEMANATIC VARIABILITY –
NUMBER (Metric 6)

Fewer senses (or a single sense) may make it easier to assign a role to an affix (Goldschneider & DeKeyser, 2001).

This metric indicates the number of senses an affix has within the affix types at a given vocabulary size. The number of senses each affix has was determined by consulting the sense listings in the Oxford and Merriam-Webster dictionaries. Two raters, both L1 speakers of English with backgrounds in Applied Linguistics, first collapsed the senses into those that were broadly similar. For example, for the suffix *-able*, the senses “capable of” and “able to be” (e.g., *breakable*, *collectible*) were treated as one sense, while the senses “tending, given, or liable to” (e.g., *agreeable*, *perishable*), “subject to” (e.g., *taxable*), “relevant to” (e.g., *fashionable*), and “having a quality” (e.g., *comfortable*, *horrible*) were together treated as a second sense. Each affix type was then assigned to an affix sense based on word definitions in the Oxford and Merriam-Webster dictionaries. In cases where a word has multiple meanings, with one meaning in one affix sense and another meaning in another, 100 random concordance lines were taken from the TV and movies sub-corpus in COCA, the most frequent meaning was determined and the word assigned to this affix sense.

SEMANATIC TRANSPARENCY –
TYPES – NUMBER
(Metric 7a)

Semantic transparency may affect the likelihood of an affix being recognised as such (i.e., greater transparency = a greater chance of affix recognition). See Hay (2001, 2003); Jarmulowicz and Taran (2013).

A count was made of affix types within a given vocabulary size for which there is a transparent semantic link between the affix type and its base word or precursor. Transparency of semantic links was determined by consulting dictionaries and a semantic classification database: (a) in the Oxford and Merriam-Webster dictionaries, if either a base word or precursor appeared in the definition of an affix type, then a semantically transparent relationship was coded as present; (b) in the UCREL Semantic Analysis System (USAS) (Pao et al., 2015), accessed through the WMatrix corpus analysis tool (<https://ucrel.lancs.ac.uk/wmatrix/>), if the base word/precursor and affix type received the same semantic tag, then a semantically transparent relationship was coded as present. If either approach (a or b) provided evidence of semantic transparency, a semantically transparent link was coded. Example 1: In one dictionary (Merriam-Webster), the definition of *unbelievable* is “of such a superlative degree as to be hard to believe”. Thus, the base word *believe* appears within the dictionary entry for *unbelievable*, and this was coded as semantic transparency. Example 2: The two dictionaries define *bossy* as “fond of giving people orders; domineering” and “inclined to domineer; dictatorial”. However, in USAS, the forms *boss* and *bossy* both receive the same semantic tag (S7.1: power, organizing). Therefore, the semantic tagging was taken as evidence of semantic transparency.

SEMANATIC TRANSPARENCY –
TYPES – PROPORTION
(Metric 7b)

Metric 7a was divided by the number of affix types (Metric 1) within a given vocabulary size.

(Continued)

Metric name	Rationale	Method
SEMANTIC TRANSPARENCY – TOKENS – NUMBER (Metric 7c)		The number of affix tokens for each affix type coded in Metric 7a was found. These affix token counts were then summed.
SEMANTIC TRANSPARENCY – TOKENS – PROPORTION (Metric 7d)		Metric 7c was divided by the summed total of the affix tokens within a given vocabulary size.
MORPHOLOGICAL FAMILY SIZE – TYPES – MEAN (Metric 8a)	Greater morphological family size increases the likelihood of a word being accessed via decomposition of its component parts rather than whole word access. See Kuperman and Van Dyke (2011); Ford, Davis and Marslen-Wilson (2010); Bertram, Baayen and Schreuder (2000).	This metric indicates the average size of the morphological word families of the affix types within a given vocabulary size, calculated across the affix types. For example, at the ~1000 level, for the affix <i>-ation</i> , there are three affix types: <i>attention</i> , <i>decision</i> and <i>question</i> . At the ~1000 level, no other members of the <i>attention</i> or <i>question</i> word families occur and one other member (<i>decide</i>) of the <i>decision</i> family occurs. Hence, the word families for these three affix types at the ~1000 level contain 1, 2 and 1 member respectively, giving an average size of 1.33. Morphological family membership was determined based on the supplementary data provided with Brysbaert et al. (2016). The 11,000 lemmas in our word list were matched to morphological family headwords from the Brysbaert et al. list. Because the Brysbaert et al. list is based on orthographic form only (i.e., it does not take account of part of speech), manual corrections were made to account for homographs in the lemma list. For example, the word <i>mine</i> appears in the Brysbaert et al. list as a member of the personal pronoun <i>I</i> family. This was manually corrected such that the possessive pronoun <i>mine</i> was included in the <i>I</i> family and the common noun <i>mine</i> in the same family as the verb <i>mine</i> and the nouns <i>miner</i> and <i>mining</i> .
MORPHOLOGICAL FAMILY SIZE – TOKENS – MEAN (Metric 8b)		This metric indicates the average of the summed frequencies of each affix type's word family members within a given vocabulary size.
JAPANESE LOANWORDS – TYPES – NUMBER (Metric 9a)	Since loanword status influences word difficulty (Laufer & McLean, 2016; Rogers, Webb & Nakata, 2015; Willis & Ohashi, 2012), an affix for which there are more loanwords may be easier to learn.	A count was made of affix types within a given vocabulary size that are loanwords in Japanese. Loanword status was determined following the procedure outlined in Allen (2020). A word list derived from the Balanced Contemporary Corpus of Written Japanese (nijal.ac.jp) provided a list of potential loanwords, and words that appeared in the corpus with a frequency >1 per million tokens were coded as loanwords.

JAPANESE LOANWORDS – TYPES – PROPORTION (Metric 9b)	Metric 9a was divided by the number of affix types (Metric 1) within a given vocabulary size.
JAPANESE LOANWORDS – TOKENS – NUMBER (Metric 9c)	The number of affix tokens for each affix type coded in Metric 9a was found. These affix token counts were then summed.
JAPANESE LOANWORDS – TOKENS – PROPORTION (Metric 9d)	Metric 9c was divided by the summed total of the affix tokens within a given vocabulary size.
DECOMPOSED FORMS – TYPES – NUMBER (Metric 10a)	A count was made of affix types within a given vocabulary size that have basewords or precursors with a higher frequency in the “family genre” TV/movies subtitles corpus.
DECOMPOSED FORMS – TYPES – PROPORTION (Metric 10b)	Metric 10a was divided by the number of affix types (Metric 1) within a given vocabulary size.
DECOMPOSED FORMS – TOKENS – NUMBER (Metric 10c)	The number of affix tokens for each affix type coded in Metric 10a was found. These affix token counts were then summed.
DECOMPOSED FORMS – TOKENS – PROPORTION (Metric 10d)	Metric 10c was divided by the summed total of the affix tokens within a given vocabulary size.
ORTHOGRAPHIC TRANSPARENCY, STRICT, OF BASEWORD OR PRECURSOR – TYPES – NUMBER (Metric 11a)	A count was made of affix types within a given vocabulary size for which the orthographic form of the baseword or precursor is strictly preserved in the affix type. For Metrics 11a–d, the baseword/precursor had to appear within the morphologically complex word without any changes in orthographic form (e.g., the baseword <i>comfort</i> is strictly preserved in the affix type <i>comfortable</i> , but this is not the case for the baseword <i>error</i> in the affix type <i>terrible</i>). In the case of words in which a letter is shared by the base and an affix, it was treated as being present in both elements. For example, the affix type <i>driver</i> contains the suffix <i>-er</i> , but the <i>e</i> could be considered part of the base as well as the suffix. Since the baseword orthographic form is still present, this was coded as a case of orthographic transparency of the baseword.

(Continued)

Metric name	Rationale	Method
ORTHOGRAPHIC TRANSPARENCY, STRICT, OF BASEWORD OR PRECURSOR – TYPES – PROPORTION (Metric 11b)		Metric 11a was divided by the number of affix types (Metric 1) within a given vocabulary size.
ORTHOGRAPHIC TRANSPARENCY, STRICT, OF BASEWORD OR PRECURSOR – TOKENS – NUMBER (Metric 11c)		The number of affix tokens for each affix type coded in Metric 11a was found. These affix token counts were then summed.
ORTHOGRAPHIC TRANSPARENCY, STRICT, OF BASEWORD OR PRECURSOR – TOKENS – PROPORTION (Metric 11d)		Metric 11c was divided by the summed total of the affix tokens within a given vocabulary size.
ORTHOGRAPHIC TRANSPARENCY, LENIENT, OF BASEWORD OR PRECURSOR – TYPES – NUMBER (Metric 11e)	Minimal, regular changes to a baseword or precursor may still make it more likely that an affix is recognised as such.	A count was made of affix types within a given vocabulary size for which the orthographic form of the baseword or precursor is preserved though with the following minimal changes permitted (see Bauer et al., 2013): (i) truncation of final -e; (ii) shifts from -y to -i or -ie; (iii) doubling of final consonants. Thus, for example, for the affix <i>-able</i> , <i>adorable</i> was coded as an affix type that is orthographically transparent for Metrics 11e–h, but not for Metrics 11a–d.
ORTHOGRAPHIC TRANSPARENCY, LENIENT, OF BASEWORD OR PRECURSOR – TYPES – PROPORTION (Metric 11f)		Metric 11e was divided by the number of affix types (Metric 1) within a given vocabulary size.

ORTHOGRAPHIC TRANSPARENCY, LENIENT, OF BASEWORD OR PRECURSOR – TOKENS – NUMBER (Metric 11g)	The number of affix tokens for each affix type coded in Metric 11e was found. These affix token counts were then summed.
ORTHOGRAPHIC TRANSPARENCY, LENIENT, OF BASEWORD OR PRECURSOR – TOKENS – PROPORTION (Metric 11h)	Metric 11g was divided by the summed total of the affix tokens within a given vocabulary size.
ORTHOGRAPHIC TRANSPARENCY OF AFFIX – TYPES – NUMBER (Metric 12a)	Affix types were coded for whether the affix occurs in its canonical orthographic form rather than as an allomorph or having undergone changes due to co-occurrence with other affixes. Decisions on canonical and non-canonical forms were based on Bauer et al. (2013). For example, the affix <i>-able</i> appears in canonical form in <i>comfortable</i> , but in <i>possible</i> an allomorph occurs and in <i>probably</i> the form is altered due to the presence of another affix. A count was then made of affix types within a given vocabulary size in which the affix occurs in its canonical orthographic form.
ORTHOGRAPHIC TRANSPARENCY OF AFFIX – TYPES – PROPORTION (Metric 12b)	Metric 12a was divided by the number of affix types (Metric 1) within a given vocabulary size.
ORTHOGRAPHIC TRANSPARENCY OF AFFIX – TOKENS – NUMBER (Metric 12c)	The number of affix tokens for each affix type coded in Metric 12a was found. These affix token counts were then summed.
ORTHOGRAPHIC TRANSPARENCY OF AFFIX – TOKENS – PROPORTION (Metric 12d)	Metric 12c was divided by the summed total of the affix tokens within a given vocabulary size.

(Continued)

Metric name	Rationale	Method
PHONOLOGICAL TRANSPARENCY OF BASEWORD OR PRECURSOR – TYPES – NUMBER (Metric 13a)	An affix is more likely to be recognised as such if the baseword or precursor is phonologically transparent. See Jarmulowicz and Taran (2013).	A count was made of affix types within a given vocabulary size in which the baseword or precursor's phonological form is preserved in the affix type. For example, the word <i>terrific</i> has the phonological form <i>'teɪrəfɪkəl</i> , and its baseword <i>terror</i> has the form <i>'teɪə</i> . Since the form of the baseword is present in the phonological rendering of <i>terrific</i> , this was coded as a case of phonological transparency. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (https://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm) (Coltheart, 1981). When the database included multiple phonological renderings, the form listed first in the Merriam-Webster Dictionary was used. When a phonological rendering did not include syllable breaks (necessary for Metric 16), the Cambridge Dictionary was consulted. When the database did not include a word, the Cambridge Dictionary was used along with the website https://www.howtopronounce.com/ . In the case of words in which a phoneme is shared by the base and an affix, it was treated as being present in both elements. For example, the affix type <i>awfully</i> contains the suffix <i>-ly</i> , but the phoneme <i>l</i> could be considered part of the precursor <i>awful</i> as well as the suffix. Since the phonological form of the precursor is still present, this was coded as a case of phonological transparency.
PHONOLOGICAL TRANSPARENCY OF BASEWORD OR PRECURSOR – TYPES – PROPORTION (Metric 13b)		Metric 13a was divided by the number of affix types (Metric 1) within a given vocabulary size.
PHONOLOGICAL TRANSPARENCY OF BASEWORD OR PRECURSOR – TOKENS – NUMBER (Metric 13c)		The number of affix tokens for each affix type coded in Metric 13a was found. These affix token counts were then summed.
PHONOLOGICAL TRANSPARENCY OF BASEWORD OR PRECURSOR – TOKENS – PROPORTION (Metric 13d)		Metric 13c was divided by the summed total of the affix tokens within a given vocabulary size.

PHONOLOGICAL TRANSPARENCY OF AFFIX – TYPES – NUMBER (Metric 14a)	An affix is more likely to be recognised as such if it tends to be phonologically transparent. See Jamulowicz and Turan (2013).	A count was made of affix types within a given vocabulary size in which the affix is in its canonical phonological form. For example, the canonical phonological form of the suffix <i>-ation</i> was taken as eɪfən. The words <i>conversation</i> and <i>opinion</i> were phonologically rendered as kɒnvə'seɪfən and ə'pɪnjən. Since the phonological form of the affix is present in the rendering of <i>conversation</i> , it was coded as a case of phonological transparency of the affix. In the case of <i>opinion</i> , the canonical affix form is not present, so this was coded as not being a case of phonological transparency of the affix. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a).
PHONOLOGICAL TRANSPARENCY OF AFFIX – TYPES – PROPORTION (Metric 14b)		Metric 14a was divided by the number of affix types (Metric 1) within a given vocabulary size.
PHONOLOGICAL TRANSPARENCY OF AFFIX – TOKENS – NUMBER (Metric 14c)		The number of affix tokens for each affix type coded in Metric 14a was found. These affix token counts were then summed.
PHONOLOGICAL TRANSPARENCY OF AFFIX – TOKENS – PROPORTION (Metric 14d)		Metric 14c was divided by the summed total of the affix tokens within a given vocabulary size.
AFFIX SONORITY – TYPES – MEAN (Metric 15a)	Research on inflectional morphology has suggested that perceptual salience may be a factor in acquisition and the sonority of an affix has been put forward as one aspect of perceptual salience. See Goldschneider and DeKeyser (2001).	The average sonority of the affix across the affix types within a given vocabulary size was calculated. Affix sonority was calculated following Goldschneider and DeKeyser's (2001) approach to measuring "total relative sonority". In this approach, each phoneme is given a sonority hierarchy point score based on the work of Laver (1994, p. 504). The scoring system is as follows: Low vowels: 9 points Mid vowels: 8 points High vowels: 7 points Glides: 6 points Liquids: 5 points Nasals: 4 points Fricatives: 3 points Affricates: 2 points Stops: 1 point For each affix type, the sonority of the affix was calculated by totalling the points for each of its phonemes. For example, in the affix type <i>possible</i>, the affix <i>-able</i> has the phonemes ə scoring 8 points, b scoring 1 point and l scoring 5 points, producing a sonority of 14.

(Continued)

Metric name	Rationale	Method
AFFIX SONORITY – TOKENS – MEAN (Metric 15b)		The average sonority of the affix across the affix tokens within a given vocabulary size was calculated.
LOW PROBABILITY PHONOTACTICS, WITHIN- WORD – TYPES – MEAN (Metric 16a)	Low probability phonotactics is one way we determine morpheme boundaries in the stream of speech. Thus, affixes represented by more words within which there are low probability phonotactics at the affix boundary are more likely to be segmentable, and thereby to develop stronger independent representations. See Hay (2003); Hay and Baayen (2005).	The average probability of the affix boundary phonotactics was calculated, based on within-word syllable boundaries within the monomorphemic types within a given vocabulary size. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a). Using the MorphoLex database (Sánchez-Gutiérrez et al., 2018), all monomorphemic forms within the 11,000 most frequent lemmas in the “family genre” TV/movies subtitles corpus were identified. For each of these forms, the onset and coda phonemes were identified for each syllable. The probabilities of each coda–onset combination occurring at syllable boundaries within monomorphemic words for each vocabulary size were calculated based on lemma type counts. The coda–onset pair at the affix boundary in each affixed form was then identified. The mean probability (across affix types) of affix boundary coda–onset pairs appearing within monomorphemic words at each vocabulary size was then calculated.
LOW PROBABILITY PHONOTACTICS, BETWEEN- WORD – TYPES – MEAN (Metric 16b)		The average probability of the affix boundary phonotactics was calculated, based on between-word boundaries within the monomorphemic types within a given vocabulary size. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a). Using the MorphoLex database (Sánchez-Gutiérrez et al., 2018), all monomorphemic forms within the 11,000 most frequent lemmas in the “family genre” TV/movies subtitles corpus were identified. For each of these forms, the onset and coda phonemes were identified for each syllable. The probabilities of each coda–onset combination occurring between monomorphemic words for each vocabulary size were calculated based on lemma type counts. The coda–onset pair at the affix boundary in each affixed form was then identified. The mean probability (across affix types) of affix boundary coda–onset pairs appearing between monomorphemic words at each vocabulary size was then calculated.

LOW PROBABILITY
PHONOTACTICS, WITHIN-
WORD – TOKENS – MEAN
(Metric 16c)

The average probability of the affix boundary phonotactics was calculated, based on within-word syllable boundaries for the monomorphemic tokens within a given vocabulary size. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a). Using the MorphoLex database (Sánchez-Gutiérrez et al., 2018), all monomorphemic forms within the 11,000 most frequent lemmas in the “family genre” TV/movies subtitles corpus were identified. For each of these forms, the onset and coda phonemes were identified for each syllable. The probabilities of each coda–onset combination occurring at syllable boundaries within monomorphemic words for each vocabulary size were calculated based on token counts. The coda–onset pair at the affix boundary in each affixed form was then identified. The mean probability (across affix tokens) of affix boundary coda–onset pairs appearing within monomorphemic words at each vocabulary size was then calculated.

LOW PROBABILITY
PHONOTACTICS, BETWEEN-
WORD – TOKENS – MEAN
(Metric 16d)

The average probability of the affix boundary phonotactics was calculated, based on between-word boundaries within the monomorphemic tokens within a given vocabulary size. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a). Using the MorphoLex database (Sánchez-Gutiérrez et al., 2018), all monomorphemic forms within the 11,000 most frequent lemmas in the “family genre” TV/movies subtitles corpus were identified. For each of these forms, the onset and coda phonemes were identified for each syllable. The probabilities of each coda–onset combination occurring between monomorphemic words for each vocabulary size were calculated based on token counts. The coda–onset pair at the affix boundary in each affixed form was then identified. The mean probability (across affix tokens) of affix boundary coda–onset pairs appearing between monomorphemic words at each vocabulary size was then calculated.

TRANSITIONAL PROBABILITIES INDICATE THE LIKELIHOOD OF ONE EVENT GIVEN ANOTHER EVENT. THEY CAN BE FORWARDS (THE LIKELIHOOD OF B GIVEN A) OR BACKWARDS (THE LIKELIHOOD OF A GIVEN B). LOW TRANSITIONAL PROBABILITIES BETWEEN SOUNDS ARE THOUGHT TO ACT AS MARKERS OF WORD/MORPHEME BOUNDARIES, AND THUS MAY HELP IN IDENTIFYING AFFIXES AS SUCH. IT HAS ALSO BEEN FOUND THAT AFFIXES WHICH HAVE RELATIVELY LOW TRANSITIONAL PROBABILITIES ARE LIKELY TO BE MORE HIGHLY SEGMENTABLE, TO DEVELOP STRONGER INDEPENDENT REPRESENTATIONS, AND TO BE MORE PRODUCTIVE. SEE PELUCCHI, HAY AND SAFFRAN (2009); SAFFRAN, ASLIN AND NEWPORT (1996); SAFFRAN, NEWPORT AND ASLIN (1996).

The average forwards transitional probability for the phoneme combinations at the base–affix boundary across the affix types was calculated, based on the transitional probabilities of phoneme combinations within the monomorphemic types within a given vocabulary size. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a). The forwards probabilities of each combination of phoneme pairs appearing in the first 11,000 lemmas of the corpus were calculated at each vocabulary size (based on type counts). For each affixed form, the phoneme pair at the boundary between the affix and the rest of the word was identified, and the probability of that pair at the given vocabulary size checked. The mean probability across all affix types was then calculated.

(Continued)

Metric name	Rationale	Method
BACKWARDS TRANSITIONAL PROBABILITY OF BASE-AFFIX BOUNDARY – TYPES (Metric 16f)		The average backwards transitional probability for the phoneme combinations at the base–affix boundary across the affix types was calculated, based on the transitional probabilities of phoneme combinations within the monomorphemic types within a given vocabulary size. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a). The backwards probabilities of each combination of phoneme pairs appearing in the first 11,000 lemmas of the corpus were calculated at each vocabulary size (based on type counts). For each affixed form, the phoneme pair at the boundary between the affix and the rest of the word was identified, and the probability of that pair at the given vocabulary size checked. The mean probability across all affix types was then calculated.
FORWARDS TRANSITIONAL PROBABILITY OF BASE-AFFIX BOUNDARY – TOKENS (Metric 16g)		The average forwards transitional probability for the phoneme combinations at the base–affix boundary across the affix tokens was calculated, based on the transitional probabilities of phoneme combinations within the monomorphemic tokens within a given vocabulary size. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a). The forwards probabilities of each combination of phoneme pairs appearing in the first 11,000 lemmas of the corpus were calculated at each vocabulary size (based on token counts). For each affixed form, the phoneme pair at the boundary between the affix and the rest of the word was identified, and the probability of that pair at the given vocabulary size checked. The mean probability across all affix tokens was then calculated.
BACKWARDS TRANSITIONAL PROBABILITY OF BASE-AFFIX BOUNDARY – TOKENS (Metric 16h)		The average backwards transitional probability for the phoneme combinations at the base–affix boundary across the affix tokens was calculated, based on the transitional probabilities of phoneme combinations within the monomorphemic tokens within a given vocabulary size. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a). The backwards probabilities of each combination of phoneme pairs appearing in the first 11,000 lemmas of the corpus were calculated at each vocabulary size (based on token counts). For each affixed form, the phoneme pair at the boundary between the affix and the rest of the word was identified, and the probability of that pair at the given vocabulary size checked. The mean probability across all affix tokens was then calculated.

ORTHOGRAPHIC AFFIX LENGTH – TYPES – MEAN (Metric 17a)	Greater length is thought to increase the likelihood of an affix being recognized as such, meaning word access is more likely to be via decomposition rather than whole word. See Kuperman and Van Dyke (2011).	The average length of the affix in letters across the affix types within a given vocabulary size was calculated. In cases where a letter was shared between two word elements, the letter was treated as belonging to both elements. Example 1: The word <i>probably</i> was coded as containing the suffixes <i>-able</i> and <i>-ly</i> . In this case, the letter <i>l</i> was treated as being present in both affixes, meaning that the affix length of <i>-able</i> in this word was three letters (<i>a</i> , <i>b</i> , and <i>l</i>) and the affix length of <i>-ly</i> was two letters. Example 2: The word <i>computer</i> was coded as containing the suffix <i>-er</i> . In this case, the letter <i>e</i> was treated as being present in both the base word and the suffix, meaning that the affix length of <i>-er</i> was two letters.
ORTHOGRAPHIC AFFIX LENGTH – TOKENS – MEAN (Metric 17b)		The average length of the affix in letters across the affix tokens within a given vocabulary size was calculated.
PHONOLOGICAL AFFIX LENGTH – TYPES – MEAN (Metric 18a)		The average length of the affix in phonemes across the affix types within a given vocabulary size was calculated. In cases where a phoneme was shared between two word elements, the phoneme was treated as belonging to both elements. Phonological transcriptions of words were obtained from the MRC Psycholinguistic Database (see Metric 13a).
PHONOLOGICAL AFFIX LENGTH – TOKENS – MEAN (Metric 18b)		The average length of the affix in phonemes across the affix tokens within a given vocabulary size was calculated.
AFFIX AS INITIAL/FINAL MORPHEME – TYPES – NUMBER (Metric 19a)	The beginnings and ends of words are better remembered (Aitchison, 2012). Thus, an affix that tends to occur at the beginning or at the end of a word might be more salient. In addition, in the case of suffixes, the final suffix is POS-defining.	A count was made of affix types within a given vocabulary size in which the affix is the initial morpheme (in the case of prefixes) or the final morpheme (in the case of suffixes).
AFFIX AS INITIAL/FINAL MORPHEME – TYPES – PROPORTION (Metric 19b)		Metric 19a was divided by the number of affix types (Metric 1) within a given vocabulary size.
AFFIX AS INITIAL/FINAL MORPHEME – TOKENS – NUMBER (Metric 19c)		The number of affix tokens for each affix type coded in Metric 19a was found. These affix token counts were then summed.
AFFIX AS INITIAL/FINAL MORPHEME – TOKENS – PROPORTION (Metric 19d)		Metric 19c was divided by the summed total of the affix tokens within a given vocabulary size.

Appendix D

Worked Examples of Metric Calculations

Example 1:

JAPANESE LOANWORDS – TYPES – NUMBER (Metric 9a)

This metric concerns whether lemmas containing a given affix are loanwords in Japanese or not (see Appendix C for details on how loanword status was determined). Taking the affix *-al* (adjectives) as an example, the basic data for this metric is as follows.

Vocabulary size	Lemma containing the affix	Loanword status
~1000	normal_jj	1
~2000	magical_jj	0
~2000	natural_jj	1
~2000	personal_jj	1
~2000	usual_jj	0
~2000	usually_rb	0
~3000	familiar_jj	0
~3000	medical_jj	1
~3000	naturally_rb	0
~3000	official_jj	1
~3000	original_jj	1
~3000	personally_rb	0
~3000	practically_rb	0
~3000	professional_jj	1
~3000	unusual_jj	0
~4000	emotional_jj	0
~4000	equal_jj	1
~4000	musical_jj	1
~4000	national_jj	1
~4000	normally_rb	0
~4000	officially_rb	0
~4000	personality_nn	1
~4000	physical_jj	0
~4000	practical_jj	0
~4000	technically_rb	0

(Continued)

Vocabulary size	Lemma containing the affix	Loanword status
~5000	accidentally_rb	0
~5000	criminal_jj	0
~5000	electrical_jj	0
~5000	instrumental_jj	0
~5000	international_jj	1
~5000	mechanical_jj	0
~5000	nuclear_jj	0
~5000	traditional_jj	0
~6000	central_jj	1
~6000	central_np	1
~6000	classical_jj	1
~6000	critical_jj	0
~6000	financial_jj	1
~6000	logical_jj	0
~6000	political_jj	0
~6000	spectacular_jj	0
~6000	technical_jj	1
~7000	commercial_jj	1
~7000	essential_jj	1
~7000	global_jj	1
~7000	ideal_jj	0
~7000	molecular_jj	0
~7000	physically_rb	0
~7000	polar_jj	0
~7000	sentimental_jj	0
~7000	sexual_jj	1
~7000	visual_jj	1
~8000	additional_jj	0
~8000	brutal_jj	0
~8000	educational_jj	0
~8000	emotionally_rb	0
~8000	equally_rb	0
~8000	federal_jj	0

(Continued)

Vocabulary size	Lemma containing the affix	Loanword status
~8000	historical_jj	0
~8000	hysterical_jj	0
~8000	occasionally_rb	0
~8000	originally_rb	0
~8000	sensational_jj	1
~8000	tropical_jj	0
~9000	cultural_jj	0
~9000	dental_jj	0
~9000	exceptional_jj	0
~9000	incidentally_rb	0
~9000	operational_jj	0
~9000	spiritual_jj	0
~9000	universal_jj	1
~10000	bridal_jj	1
~10000	experimental_jj	0
~10000	industrial_jj	0
~10000	intellectual_jj	0
~10000	maniacally_rb	0
~10000	serial_jj	0
~10000	supernatural_jj	0
~10000	thermal_jj	0
~11000	colossal_jj	0
~11000	confidential_jj	0
~11000	environmental_jj	0
~11000	facial_jj	0
~11000	inspirational_jj	0
~11000	mystical_jj	0
~11000	occasional_jj	0
~11000	regional_jj	0
~11000	tribal_jj	0

The above data thus produces the following metric for this affix, where the value is the sum of all affix types that are loanwords up to a given vocabulary size:

Vocabulary size	JAPANESE LOANWORDS – TYPES – NUMBER (Metric 9a)
~1000	1
~2000	3
~3000	7
~4000	11
~5000	12
~6000	17
~7000	22
~8000	23
~9000	24
~10000	25
~11000	25

Example 2:

KNOWN BASEWORDS/PRECURSORS – PROPORTION (Metric 5)

This metric concerns whether a baseword or precursor of an affix type is within a particular vocabulary size (in which case it is considered “known” at that level). That is, for affix types at the ~1000 vocabulary size level, the coding indicates whether the baseword/precursor is also within the first 1,000 lemmas; for affix types through to the ~2000 vocabulary size level, the coding indicates whether the baseword/precursor is also within the first 2,000 lemmas, etc. Taking the affix *-able* as an example, below are some of the basic data for this metric (only through to the ~5000 vocabulary size level, for convenience). As can be seen, this metric is somewhat more complex than that of Example 1 since the coding for a given affix type can change across vocabulary sizes. For example, *terrible_jj* is coded as 0 at the ~1000, ~2000, ~3000 and ~4000 vocabulary size levels, but is coded as 1 at the ~5000 level, since its baseword, *terror*, is at the ~5000 level.

Vocabulary size	Lemma containing the affix	Baseword/Precursor is within . . .				
		~1000 vocabulary level	~2000 vocabulary level	~3000 vocabulary level	~4000 vocabulary level	~5000 vocabulary level
~1000	impossible_jj	0	0	0	0	0
~1000	possible_jj	0	0	0	0	0
~1000	probably_rb	0	0	0	0	0
~1000	terrible_jj	0	0	0	0	1
~2000	comfortable_jj	/	0	0	0	0
~2000	horrible_jj	/	0	0	1	1
~2000	incredible_jj	/	0	0	0	0

(Continued)

Vocabulary size	Lemma containing the affix	Baseword/Precursor is within ...				
		~1000 vocabulary level	~2000 vocabulary level	~3000 vocabulary level	~4000 vocabulary level	~5000 vocabulary level
~2000	possibly_rb	/	0	0	0	0
~2000	responsibility_nn	/	0	0	0	0
~2000	responsible_jj	/	0	0	0	0
~3000	adorable_jj	/	/	0	0	1
~3000	available_jj	/	/	0	0	0
~3000	capable_jj	/	/	0	0	0
~3000	miserable_jj	/	/	0	0	0
~3000	terribly_rb	/	/	0	0	1
~3000	unbelievable_jj	/	/	1	1	1
~3000	valuable_jj	/	/	0	0	0
~4000	incredibly_rb	/	/	/	0	0
~4000	invisible_jj	/	/	/	0	0
~4000	possibility_nn	/	/	/	0	0
~4000	reasonable_jj	/	/	/	1	1
~4000	remarkable_jj	/	/	/	0	0
~5000	sensible_jj	/	/	/	/	1
~5000	uncomfortable_jj	/	/	/	/	0

The above data thus produces the following metric for this affix, where 4 types are found within the ~1000 vocabulary size level (see above: “impossible_jj”, “possible_jj”, “probably_rb”, “terrible_jj”), but none has a baseword/precursor within the ~1000 vocabulary level. Similarly, there are 10 types found within the ~2000 vocabulary size level, but none with a baseword/precursor up to this level.

Vocabulary size	Calculation	KNOWN BASEWORDS/ PRECURSORS – PROPORTION (Metric 5)
~1000	0 ÷ 4	0.000
~2000	0 ÷ 10	0.000
~3000	1 ÷ 17	0.059
~4000	3 ÷ 22	0.136
~5000	7 ÷ 24	0.292