Robert Krovetz

MORPHO-SEMANTICS AND DICTIONARY ENTRIES

Abstract This paper is about morphology and semantics, and about how their interaction is reflected in the choices made by lexicographers about dictionary entries. The paper discusses run-ons, zero-affix morphology, and the relationship between variant forms and lexical ambiguity. We compare run-on entries with variants that are headwords across different dictionaries. We found high agreement about the variants that were attested as headwords, and low agreement about the variants that were attested as run-ons. Corpus data showed differences between run-ons and headwords as well. We also compared a sample of variants that have an explicit affix with those that do not. We found many similarities with regard to the lexical semantic relationships that are involved. The paper concludes with a discussion of criteria for when derivational variants should be listed in a dictionary, and the opportunity presented by an electronic dictionary to teach a user about morphology and meaning.

Keywords morphology; lexical semantics; corpus linguistics

1. Introduction

> By trade, he's a corn-chandler [said one character]. "And what on earth *is* a corn-chandler?" Peggy asked crossly...
>
> Lady Mear said: A man who chandles corn, I suppose. Even my underrated intelligence can work that out.
>
> --- Barbara Worsley-Gough (1932, p. 149)

Morphology is about the ways that words can vary in form. It includes a wide variety of phenomena: acronyms (*FBI*), abbreviations (*corp.*), initialisms (*C.B.S.*), slashonyms (*d/b/a - doing business as*), compounds (*blackbird*), inflectional variants (*dogs, loves/loved/loving*), derivational variants (*teacher, appropriation, misread*), back-formations (*bartend*), and zero-affix variants such as *cook* as a verb and as a noun. More than one type of variation can apply, e.g., *Federal Bureau of Investigation* can be referred to as *FBI* or *F.B.I.*

Morpho-semantics is about how different word form variations relate to meaning. According to Lieber (2012, p. 2108) "the most neglected area of morphological theory in the last three decades has been derivational semantics". Similarly, Levin & Rappaport Hovav (2017) observe that "The relationship between lexical semantics and morphology has not been the subject of much study".

Dictionaries are an excellent and understudied resource for investigating morpho-semantic relationships. They allow us to evaluate which derivational variants are given explicit definition, and which have multiple meanings. We can also use corpus analysis to help identify highly-related variants. These variants often qualify as *run-ons,* which are derivational variants that are listed at the end of a homograph. We can compare such variants with those that are explicitly defined.

The notion of zero-affix morphology expresses the idea that a word can have an affix that is not visibly present. This can be seen with the word *undesirables.* Starting from the root *desire,* it transforms into *desirable,* and then *undesirable.* However, both *desirable* and *undesirable* are adjectives, and *undesirables* is a plural noun. There must be an implicit step that follows *undesirable* that converts the adjective to a noun. This is called *zero-affix morphology,* also known as *conversion*[1] or *functional shift,* because it involves a shift from one part of speech to another. Most of the literature focuses on the semantic relationships that are involved with zero-affix variants rather than their derivation. For example, *cook* as a noun is a person who cooks.

The semantic relationships between roots and derivational variants (as well as zero-affix variants) are very similar to the relationships seen with thematic roles. These relationships express different roles that are played by the noun phrases in a sentence: AGENT, INSTRUMENT, PATIENT, LOCATION, DESTINATION and others. For example, in the sentence *John ate the spaghetti with his fork,* the word *John* is an AGENT, and *fork* is an INSTRUMENT. If the sentence were *John ate the spaghetti with his wife,* John's wife plays the role of CO-AGENT rather than INSTRUMENT. These relationships are essential to understanding the meaning of the sentence. The suffix *-er* is often used to convey an AGENT or an INSTRUMENT relationship, as with *singer* or *shredder.* Similarly, *cook* as a noun is an AGENT (a person who cooks), and *ski* as a noun is an INSTRUMENT (something that is used for skiing). Morpho-semantics is linguistically interesting because it is neither entirely rule-governed nor completely idiosyncratic. We will discuss this further in Section 2.

Dictionaries differ in the types of morphological information they provide, and about how they convey this information to the user. The COBUILD Dictionary includes regular and predictable inflectional variants as part of the homograph for a word. Other dictionaries only include irregular inflectional variants (Macmillan), or define the irregular variant in a separate homograph (Collins). Dictionaries also vary about how they describe derivational variants. If the meaning is considered especially predictable, the derivational variant is listed as a *run-on* at the end of an entry. Sometimes this is expressed just as an affix (e.g, *-ly* at the end of the entry for *natural* in order to indicate that *naturally* is a word), and sometimes the variant form is spelled out.

This paper is an empirical analysis of the decisions that have been made by lexicographers about which word forms were included as run-ons. We conducted a corpus analysis using a large-scale resource of word families of derivational variants. We used the co-occurrences between roots and variants to help identify additional candidates for run-ons. We expect that those variants that co-occur will be more likely to be transparent.

[1] There are different viewpoints about conversion, and some make a distinction between conversion and zero-affix morphology (Valera, 2014).

We not only looked at run-on information, we also looked at whether derivational variants were expressed as run-ons or as headwords.

Finally, we looked at the relationship between derivational variants and word senses. Morphology is not a relationship between words, but between the senses of words. For example, *gravity* is related to *antigravity* and *microgravity,* but only in the sense of force-of-gravity. It is not related in the sense that means *serious,* as in *the gravity of the crime/offense.* We conducted an initial investigation about how often that occurs, and we make analogies between relationships that involve explicit affixes and relationships between zero-affix variants.

Our experiments were done with English. See Lieber (2012) for a discussion of derivational morphology in other languages.

The next section will discuss related work in linguistics and in lexicography. We will then describe the experiments that were done, and the results.

2. Related Work

The lexicon used to be considered "a repository of exceptions" (Bloomfield, 1933). This view changed with the publication of *Remarks on Nominalization* (Chomsky, 1970). That paper noted the similarities between verbs and nominalizations, such as *The enemy destroyed the city* compared with *The enemy's destruction of the city.* The paper revolutionized the view of the lexicon by recognizing that it must have structure in order to support such regularities. It led to work on derivational morphology as a research subject, and its importance is reflected in a recent book entitled *Nominalization: 50 Years on from Chomsky's Remarks* (Alexiadou & Borer, 2020).

Another seminal work is *The Categories and Types of Present-Day English Word Formation* (Marchand, 1969). That work is still a key reference for word formation and the interactions with semantics. In the second edition, Marchand makes analogies between zero-derivation and explicit affixes. So *clean(adj)/clean(v)* are related in the same way as *legal(adj)/legalize(v).* In both cases the verb is an action that results in the state referred to by the adjective. Lipka says "if the concept of zero is really taken seriously, it is not the noun which denotes the agent, etc., but rather the zero-morpheme, since it is the very parallelism of *cheat/Ø* vs. *swindle/er, cook/Ø* vs. *bak/er, spy/Ø* vs. *observ/er* on which the theoretical concept of zero is based" (Lipka, 1975, p. 386). We want to automatically identify the type of relationship by extracting such information from dictionary definitions. For example, *wad* (verb) can be defined as "*to make a wad of*". That is, the verb bears the semantic relationship of "formation" to the noun. This is similar to the effect that the morpheme *-ize* has on the noun *union* in order to make the verb *unionize.* Section 6 describ*es* initial efforts to make these connections.

Princeton WordNet has been augmented with morphological information (Fellbaum & Miller, 2003). Princeton WordNet is the most widely used lexicon in Computational

Linguistics, and it consists of synsets[2] that express different types of lexical semantic relationships between word senses. Each synset is based on open-class parts-of-speech (nouns, verbs, adjectives, and adverbs). The morphological information adds connections across part-of-speech based on derivational relationships (e.g., *beauty/beautify, generalize/generalization, connect/connection*) and zero-affix relationships such as *cook.* Mitetelu et al. (2023) report on analyzing words that are zero-affix variants with regard to semantic classes that are at the top of the hierarchy in WordNet. These are classes like noun.person, noun.artifact, verb.change, verb.motion, and others. There are 25 noun classes and 15 verb classes in all. They found that zero-affix variants were highly frequent in their dataset, and it was the only affix that occurred with all 14 types of semantic relationships that the resource describes (Agent, Instrument, State, and others). They also found differences between zero-affix variants and explicit variants with regard to the distribution of the classes that were at the top of the hierarchy, and the semantic relationships between the words and variants.

Lieber (2004) mentions the following research issues in morpho-semantics:

- *The polysemy question.* This is the question about why an affix can express one type of relationship with one word, and a different type of relationship with another. For example, the way *-er* can sometimes be used to indicate an AGENT, and sometimes an INSTRUMENT.

- *The multiple-affix question.* This is the question about why different affixes are used to express similar relationships, as with *-ize* and *-ify* to indicate causative verbs, and *-er* and *-ant* to indicate agent nouns.

- *The zero-derivation question.* This is the question of how to account for word formation in the absence of an explicit affix.

- *The semantic-mismatch question.* This is the question of why there are morphemes that do not seem to add any meaning (e.g., *in* in *longitudinal*), and sometimes we have two affixes that seem to be redundant (e.g., *-ic* and *-al* in *dramatical*). Why doesn't adding an affix always add to the meaning? Why is it that *realistic* does not mean "pertaining to a realist"?

With regard to lexicography, DeCesaris (2021) mentions there has been relatively little research on morphological structure and dictionaries. She also notes that "the fine line between being entirely semantically transparent and only partially so is often blurred". What is added by an affix that warrants additional description?

The work of Sue Atkins with Charles Fillmore on FrameNet (Fillmore & Atkins, 2012), and the work of Patrick Hanks on Norms and Exploitations (Hanks, 2013) is especially relevant to the interaction of morphology and lexicography. FrameNet is a resource

[2] A synset is a group of sense-individuated words that share a common meaning. WordNet is a combination of a dictionary and a thesaurus.

that organizes the lexicon into a network of frames, where each frame represents a conceptual structure that captures the relationships among words and the roles they play in specific events or actions. FrameNet allows frames to inherit properties from more general frames, facilitating the capture of linguistic generalizations, while also permitting the expression of idiosyncratic information specific to particular lexical units. The work on Norms and Exploitations captures the idea that there are norms in the lexicon as well as creative ways that we can exploit those norms and therefore get irregularities.

The next section will discuss experiments to combine corpus analysis with dictionaries in order to get a better understanding of lexical semantics and the decisions made by lexicographers. This will be followed by a section about the results, and a section that compares how derivational variants are attested across three dictionaries. We then discuss our investigation of zero-affix variants. The paper will conclude with a discussion of the opportunities that are provided by electronic dictionaries for conveying morpho-semantic information.

3. Design of Experiments

We conducted an experiment with run-ons, and we made two initial investigations, one with derivational variants that were explicitly defined, and the other with zero-affix variants.

For the experiment with run-ons, we wanted to see if corpus information could be used to create such information automatically. We looked at co-occurrences within a paragraph to see how run-ons compared with headwords, and with variants that were neither run-ons nor headwords.

We used a download of the Wikipedia to compute co-occurrence information for two datasets. The first dataset used paragraphs, and the second used the entire article. We used a set of 5065 word families that we created for a different purpose for this work.[3] We manually identified the run-ons for the homographs in the *Longman Dictionary of Contemporary English* (LDOCE) for all words that started with the letters A-C. We also identified those derivational variants that were explicitly defined. Our aim was to identify new candidate run-ons based on corpus data, and to see how run-ons and headwords compared.

For the investigation of zero-affix morphology, we used a second resource. The morpho-semantic links in Princeton WordNet include both derivational and zero-affix variants.[4] We extracted the zero-affix variants from this dataset, and compared them with the first 100 derivational variants that are headwords in LDOCE. This work will be discussed in Section 6.

[3] http://lexicalresearch.com/resources/derivational-word-families.v12.tar

[4] See https://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls for the resource file. The semantic relationships are discussed in (Fellbaum et al., 2007).

4. Run-ons and Headwords

There are 5065 word families in our resource, and 13,488 derivational variants. We used a subset of 1171 roots that began with the letters A-C for our comparison. There were 3147 variants for this set. Table 1 gives a breakdown of those variants with regard to whether they appear as a run-on, or as a suffixed headword, or as a prefixed headword in the LDOCE sample.

Our aim was to compare derivational variants that were run-ons with those that were explicitly defined, and to see how well we could identify variants that were presumed to be transparent. We used a subset of the roots that began with A-C because we manually identified whether the variant was attested as a headword or as a run-on. Prefixed variants were looked up under their first letter.

We found that of the 3147 variants, 601 were run-ons (19%), 989 were headwords (31%), and 1557 were candidates (50%). Of the headwords, 769 were suffixed variants, and 220 were prefixed variants. Of the candidates, 795 were suffixed and 762 were prefixed. Some of the variants did not occur in the Wikipedia download, some of the roots did not occur, and sometimes the variant and root did not co-occur in a paragraph. The rest of the run-ons, headwords, and run-on candidates were used to determine co-occurrences, and the number of variants for each group is shown in the Table (e.g., of the 601 run-ons, 457 co-occur with their roots within the context of a paragraph).

There were 439 candidates that did not co-occur with their roots within a paragraph. The co-occurrences can be low because morphological variants often occur with a lower frequency than the root, and in addition a paragraph is a relatively short context. We looked at two ways to capture more co-occurrences: 1) increasing the size of the context from a paragraph to an entire Wikipedia article, 2) using additional corpora. Table 1 compares the results using co-occurrence in a paragraph with co-occurrence in the entire article. We used two domain-specific datasets for supplementing the Wikipedia paragraph dataset: Medline, and Juris.[5] In the Medline dataset the co-occurrences were within a title and abstract, and in Juris they were within a paragraph. We wanted to assess the impact of using datasets that were very different. The combined dataset increased coverage, but not as much as expanding the context from a paragraph to an entire Wikipedia article. Using the combined dataset there were 237 candidates that did not co-occur compared to 218 using the entire Wikipedia article and 439 using Wikipedia paragraphs.

[5] The Medline dataset consists of over 17 million documents from the National Library of Medicine (http://pubmed.ncbi.nlm.nih.gov). The Juris dataset is made up of cases, statutes, briefs, treaties and other legal material (https://public.resource.org/justice.gov/index.html).

Table 1: A distribution of derivational variants into different classes. The variants are from a resource of word families. The classes are run-ons, suffixed headwords and prefixed headwords sampled from the *Longman Dictionary of Contemporary English* (LDOCE). The number of words and the percentages of the population that co-occur are given for the different classes, as well as how this varies for different datasets. The table also includes prefixed and suffixed variants that co-occur with the corresponding roots, but which are not run-ons or headwords (candidates for run-ons). The Don't Co-Occur row refers to candidates that do not co-occur with their root within the dataset

	Wiki Paragraph	**Wiki Article**	**Medline**	**Juris**
Run-on	457 (76%)	520 (86%)	371 (62%)	261 (43%)
Suffix HW	716 (93%)	745 (96%)	598 (78%)	512 (67%)
Prefix HW	209 (95%)	205 (93%)	168 (76%)	141 (65%)
Don't Co-Occur	439	218	418	500
Suffix Co-Occur	532 (70%)	632 (78%)	432 (54%)	194 (24%)
Prefix Co-Occur	494 (61%)	615 (76%)	487 (60%)	306 (40%)

We will discuss these results in Section 7. Table 2 shows the results over the entire word family resource, broken down by the type of affix. This was computed over paragraphs in the Wikipedia, but the results for the entire article are similar. We found that *-ly, -ion,* and *-er* were the most highly ranked suffixes based on productivity (number of derivational variants). The most productive prefixes were *un-, re-,* and *in-.* The Frequency column reflects the subset of the words in which the derivational form with that affix co-occurs with the corresponding root in the Wikipedia.

Table 2: The 10 most productive suffixes and prefixes for a set of over 5,000 word families. The table lists the number of words with each affix (how productive it is) for the roots and variants that co-occur within a paragraph in the Wikipedia dataset

Affix	**Type**	**Frequency**	**Affix**	**Type**	**Frequency**
ly	suffix	1295	un	prefix	825
ion	suffix	1049	re	prefix	370
er	suffix	983	in	prefix	217
ness	suffix	556	non	prefix	156
al	suffix	476	sub	prefix	142
ity	suffix	412	pre	prefix	l08
ic	suffix	339	inter	prefix	106
able	suffix	318	dis	prefix	98
or	suffix	220	over	prefix	93
ist	suffix	205	de	prefix	82

We looked at the derivational variants and run-on entries in more detail. We wanted to know more about why lexicographers chose to explicitly define a variant. We also wanted to know about consistency. If the homograph had more than one sense, did the run-on apply to all of the senses? The first 100 homographs for derivational variants and the first 100 run-ons in LDOCE were assessed.

There were several cases where it was not clear why a derivational variant was explicitly defined. For example, consider these two definitions:

(v) ***acquiesce*** **-** to agree, often unwillingly, without raising an argument; accept quietly
(adj) ***acquiescent*** **-** ready to agree without argument

The homograph for the first word has a run-on of *-escence,* and the second has a run-on of *-ly.* It is not clear why the lexicographers could not have used *-escent, -escently* and *-escence* as run-ons to the first homograph. There were also cases where most of the senses correspond, but where one sense of the root does not have a corresponding sense with the variant, or vice-versa. Such differences could be mentioned in a usage note. Section 7 gives criteria for when derivational variants should be used as headwords.

We found that run-ons were generally consistent in that they applied to all of the senses in a homograph. However, there were exceptions. For example, the homograph for *acute* had run-ons of *-ly* and *-ness.* These apply to *acute* in the sense of illness. These run-ons do not apply to the usage *acute accent,* which was another sense that was listed in the same homograph.

5. Comparison With Other Dictionaries

In our experiment with run-ons, we expected that when a variant and a root co-occur, that would be an indication of semantic transparency. We expected that the population of run-ons would have more co-occurrences with their roots than headwords co-occurring with their roots. But, that is not what we found. Both suffixed and prefixed headwords co-occur more often. The candidates for run-ons also do not co-occur as often as the headwords. We examined two other dictionaries, *Collins COBUILD English Language Dictionary* (Sinclair, 1987) and *The American Heritage® Dictionary of the English Language* (Soukhanov, 1992) to see how they compared with the Longman dictionary (LDOCE) about how derivational variants were attested. We chose COBUILD because it is a British dictionary for learners of English (as is Longman). We chose the American Heritage® dictionary because it is an American dictionary with a focus on usage. We used three sets of 100 words each for the comparison: 1) the first 100 derivational variants in LDOCE that were headwords, 2) the first 100 derivational variants that were run-ons, 3) a set of 100 words from our word family resource that were neither headwords nor run-ons (i.e., words that were candidate run-ons). We wanted to know how these three sets were described in the other two dictionaries. All of these sets involved only suffixed variants.

We found high agreement between the dictionaries with regard to the first set. Of the 100 that were headwords, 89 were also headwords in COBUILD and 91 were headwords in the American Heritage® dictionary. For COBUILD, there were 3 that were attested as run-ons (*absently, acidity,* and *adoption*), and 8 that were not defined as headwords or run-ons (*abridgment, ablative, academician, accumulator, acidify, acidulated, acquaintanceship,* and *adoptive*). For American Heritage®, the other 9 variants were all run-ons (*absently, accountancy, acquaintanceship, acrobatic, actually, addictive, additional, administrative,* and *adoption*).

We found low agreement for the second set. Of the 100 run-ons in LDOCE, 21 were attested as headwords in COBUILD, and 13 were headwords in the American Heritage® dictionary. There were 29 that were also run-ons in COBUILD, and 80 that were run-ons in American Heritage®. There were 50 that were neither headwords nor run-ons in COBUILD, and 7 for the American Heritage® dictionary.

For the third set (candidates for run-ons), 9 were attested as headwords in COBUILD, and 28 as headwords in American Heritage®. There were 2 that were attested as a run-on in COBUILD, and 61 that were attested as run-ons in the American Heritage® dictionary.

There were significant differences in ambiguity between the three sets when they were attested as headwords in the other dictionaries. For the 89 members of the first set that were attested in COBUILD as a headword, 40 (45%) were ambiguous. For the 91 members that were attested in the American Heritage® dictionary as headwords, 71 (78%) were ambiguous. In contrast, for the 21 members of the second set (run-ons in LDOCE) that were attested as headwords in COBUILD, only 5 (24%) were ambiguous. Of the 13 that were attested in the American Heritage® dictionary as headwords, 8 (61%) were ambiguous. For the third set, of the 9 that were attested in COBUILD as headwords, none were ambiguous, and of the 28 that were attested in the American Heritage® dictionary, 11 (39%) were ambiguous. With each set, the percentage of ambiguous headwords decreases, and this occurs for both dictionaries.

6. Zero-Affix Morphology

We identified the words that were indicated as having a zero-affix relationship using the morpho-semantic links in WordNet 3.0. These relationships were compared against the relationships for explicitly defined derivational forms. We used the first 100 homographs for derivational variants from LDOCE for this purpose, which is the same set we used in the comparison with run-ons. The type of relationship is given in the resource.

We did indeed find a number of cases where the semantic relationship was the same. For example:

Agent:	advocate, affiliate, ally, alternate, associate, author abortionist, accompanist, accountant, actor, adaptor, adherent
Instrument:	airbrush, aquaplane, autoclave, ax/axe abrasive, accelerator, accumulator
Theme:	abstract, ad-lib, advance, affix, award acquisition

Location:	angle, address abrasion, accommodation
Process/Result:	account, ache, advantage, aggregate, alarm, arch, average abbreviation, abridgment, accumulation
Process/State:	alert, anger, awe acquiescent, addiction, admissible

In each of the above relationships, the first line shows instances where there is no explicit affix, and the second line has instances where the affix is explicitly given. There is an analogous relationship in each group to the corresponding words. For example, an *author* is someone who *authors,* just as an *adherent* is a person who *adheres.* We obtained the semantic relationship for zero-affix variants from the morpho-semantic resource. We used our own judgment with regard to the classification of the derivational variants, although sometimes they were also found in the resource.

Not all of the zero-affix variants were found in LDOCE with more than one part-of-speech. For example, *affiliate* was only defined as a verb, and *author* was only defined as a noun. In some cases our classification for the semantic relationship for the derivational variant differed from the one given in the resource. The word *addiction* was classified as an event rather than a state (the LDOCE definition was "the state of being addicted"). The resource used the category *Undergoer* for the words we list under *Theme.*

We are exploring automatic methods to create zero-affix variants from a dictionary and from a corpus. To create the resource from a dictionary we wrote a script to create links between the definitions of word senses. If there were two or more open-class words (lemmas) in common between the definitions of a word that differ in part-of-speech, we create a link. For example, *tonsure* is defined in WordNet 3.0 as:

(n) **tonsure** (the shaved crown of a monk's or priest's head)
(n) **tonsure** (shaving the crown of the head by priests or members of a monastic order)
(v) **tonsure** (shave the head of a newly inducted monk)

The verb, *tonsure,* has the lemmas *shave, head,* and *monk* in common with the first sense of the noun, and *shave* and *head* in common with the second sense. This overlap creates a link between *tonsure* as a noun and as a verb, and that link is used to create a dataset of zero-affix candidates. For identifying candidates from a corpus, we tagged a corpus (Wikipedia), and identified cases in which a word co-occurs in a paragraph with two different parts-of-speech. The more often it co-occurs, the more likely the word meanings are to be related despite the difference. For example, we expect that *cook* as a verb and as a noun co-occur more often than *train* as a verb and as a noun.

Zero-affix morphology is a research topic. Our aim in these comparisons is to use dictionaries to explore the semantic relationships involved. We believe that corpus data will complement what we can extract from dictionaries.

7. Discussion

The interaction between morphology and semantics presents challenges as well as opportunities for lexicography. When do we need to provide a derivational variant as a headword? There are at least the following cases:

1. When there is no apparent root. These are cases like *altruism, masochism,* and *ventriloquism.* Such words are also strongly connected with *-ist* variants, which have a transparent and predictable relationship to the root. We also have *ambiguous,* with the related form *ambiguity.*

2. When there is a high semantic distance between the root and the variant. For example, *social* and *socialism.*

3. When the variant has a more specific meaning than the root and one of the arguments of the root is "hardwired". An example is the word *abstain,* which has variants *abstention* (abstain from voting), and *abstemious* (abstain from food and drink). The root form does not specify what the person is abstaining from, but this is made explicit with the variants.[6]

4. When the derived form is ambiguous. For example, *gravity* can be used in the sense *force-of-gravity,* or to mean *serious* (as in *the gravity of the crime*). This involves ambiguity between a regular word and a derivational variant. Another example is *invalid* referring to a person or to mean *not valid.* A second type of ambiguity is that between adjectival participles and tensed verbs. For example, *accomplished* can be used in the context *an accomplished pianist* or in the context *he accomplished the task.* The tensed verb use is expected, but the adjectival participle sense is not. It is not always clear how to divide a word's usages into senses, and these are some of the cases that were encountered.

We note that it is also possible to have ambiguity between a root and an inflected form, as with *bats/nuts/crackers/bananas* as an adjective that means *crazy* or as a plural noun, or with *minutes* to refer to a singular noun (the *minutes of the meeting*) or a plural noun. The unpredictable usage needs to be explicitly defined.

We were able to use our resource of word families to suggest additional run-ons. Variants that were not attested as headwords or as run-ons in LDOCE were attested as run-ons in two other dictionaries. But we found low agreement between dictionaries about which variants should be described as run-ons. Sometimes they were attested as headwords, and sometimes they were not attested at all. In contrast, there was high agreement that the variants that were attested as headwords in LDOCE were attested as headwords in the other two dictionaries.

We also found a difference between run-ons and headwords in the experiment with corpus data. A greater proportion of the headwords co-occur with their roots

[6] We consider it analogous to the process of Currying, in which an argument to a function is hardwired (as in converting add(x,y) into add_one(x), which adds one to x rather than y). *Abstention* can also be used in the more general sense of a voluntary decision not to act.

compared with run-on entries co-occurring with their roots. This was true across four different datasets. We looked at the corpus data in more detail to determine why this was so. The suffixed variants that were attested as headwords were, on average, four times more frequent in the Wikipedia than the variants that were attested as run-ons. The suffixed run-on candidates were even less frequent. This pattern held true across the domain-specific corpora, although the degree of variation differed by corpus. The pattern was also true for prefixed forms. The run-ons for prefixed headwords were much less frequent, on average, than the headwords they corresponded with.

There were differences in ambiguity between headwords, run-ons, and run-on candidates. The percentage of ambiguous variants is greatest in the variants that are attested as headwords, and it is the least in the run-on candidates for LDOCE that are attested as headwords in the other dictionaries. This makes sense given the findings with frequency—more frequent words also tend to be more ambiguous. We need an additional way to determine if a derived form is ambiguous. As future work, we intend to look at Transformers (Vaswani et al., 2017) for this purpose. Transformers are a type of neural-network architecture in which the representation used for a given word form is context-sensitive. We also plan to use this approach to re-assess the relatedness of variants and roots for run-ons and run-on candidates.

The availability of dictionaries in electronic form presents opportunities for teaching the user about morpho-semantics. DeCesaris (2021) noted that the Random House dictionary presented combining forms[7] as a list of words in which the combining form occurs. We can extend this to word families. When a user asks for a word to be defined, we can show the other members of the word family as well. Table 3 shows some of the largest word families in our resource. The word family for *state* is a combination of variants that are related to *state* as a verb and to *state* as a noun, and since they have different meanings they are arranged in different groups. There are different meanings for the nouns as well. The aim is to give the user an appreciation for the wealth of variation rather than just expose them to a run-on with a predictable meaning.

We should not only show the word family that the variant belongs to, we should also contrast instances from similar affixes when appropriate. For example, *non-* and *un-* both express negation, but they cannot be used interchangeably, and there are important differences in meaning: *nonchristian* refers to a person, but *unchristian* refers to a belief or a behavior. If a user asks about one, the other can be used as a contrast. The prefixes *de-* and *dis-* can also convey negation, but they are often associated with reversal of an existing state (as with *deactivate* or *disconnect*). In addition, *de-* can be used to indicate removal (e.g., *debone*), or reduction (e.g., *devalue*), among other meanings. Prefixed variants are generally not as frequent as suffixed variants, so English learners might not encounter them often enough to fully grasp the distinctions.

The contexts that are used to illustrate a word's usage can be chosen to include roots that co-occur with variants (as with our experiments). This can help inform the user about how different word forms occur in context.

[7] A *combining form* is a morpheme that comes from Latin or Greek; they can often be paraphrased with an open-class word (e.g., *geo-* (Earth), or *hydro-* (water)).

Table 3: A sample of some of the largest word families in English

interpret	interpretation misinterpret misinterpretation reinterpret reinterpretation interpreter interpretable interpretability uninterpreted uninterpretable interpretive interpretively
polar	polarity multipolarity polarize polarizer polarizable polarizability polarization depolarize depolarization repolarize repolarization unpolarized nonpolar
state	stateless statelessness stateful stative statehood tristate multistate substate superstate interstate intrastate statement overstate overstatement understate understatement misstate misstatement restate restatement unstated

8. Conclusion

Dictionaries vary a great deal with regard to the morphological information they contain, and how it is conveyed. But we found a great deal of commonality with regard to the derivational variants that were defined as headwords. About 90% of a sample of variants from the Longman dictionary were also headwords in two other dictionaries. In contrast, we found low agreement with regard to variants that were run-on entries, which are provided without definition at the end of a homograph. They were sometimes explicitly defined in other dictionaries, sometimes they were run-ons, and sometimes they were not attested as either a headword or as a run-on.

We analyzed corpora for variants attested as headwords to see how they compared with variants attested as run-ons. We found that a greater proportion of headwords co-occurred with their roots compared with run-ons co-occurring with their roots. Headwords were much more frequent than run-ons, and those were more frequent than run-on candidates. We also found that there were differences in ambiguity when headwords, run-ons, or run-on candidates were attested as headwords in other dictionaries. Variants that were attested as headwords had the highest percentage of words that were ambiguous, and variants that were run-on candidates had the lowest percentage of ambiguous words.

While it is a common practice to include run-on information in dictionary entries, it is unclear that this is useful to the user. A comparison with a resource of word-families indicated many variants that could have been included as run-ons, but were not, and a sample of the variants that *were* defined included a number of cases where the meaning seems transparent. We provided a set of criteria regarding when a variant should be explicitly defined. We propose that it is beneficial to expose a user to entire word families rather than just run-ons, and that dictionaries should be set up to convey distinctions between variants with semantically similar affixes.

References

Alexiadou, A., & Borer, H. (Eds.). (2020). *Nominalization: 50 Years on from Chomsky's Remarks.* Oxford University Press.

Bloomfield, L. (1933). *Language.* Henry Holt.

Chomsky, N. (1970). Remarks on Nominalization. In R. Jacobs, & P. Rosenbaum (Eds.), *Readings in English Transformational Grammar* (pp. 184–221). Ginn & Co.

Choudhari, D., Damani, O., & Laxman, S. (2011). Lexical Co-occurrence, Statistical Significance, and Word Association. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1058–1068). Association for Computational Linguistics.

DeCesaris, J. (2021). Dictionaries and Morphology. In Z. Gavrillidou, L. Mitits, & S. Kiosses, (Eds.), *Lexicography for Inclusion: Proceedings of the 19th Euralex International Conference* (pp. 577–584). Democritus University of Thrace.

Fellbaum, C., & Miller, G. A. (2003). Morphosemantic links in WordNet. *Traitement automatique de langue, 44*(2), 69–80.

Fellbaum, C., Osherson, A., & Clark, P. E. (2007). Putting Semantics into WordNet's "Morphosemantic" Links. In *Proceedings of the Third Language and Technology Conference,* Poznan, Poland.

Fillmore, C., & Atkins, B. T. (2012). Towards a frame-based lexicon: the semantics of RISK and its neighbors. *Frames, fields, and contrasts* (pp. 75–102). Routledge.

Hanks, P. (Ed.). (1979). *The Collins English Dictionary.* Collins.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations.* MIT Press.

Levin, B., & Rappaport Hovav, M. (2017). Morphology and Lexical Semantics. In A. Spencer, & A. Zwicky (Eds.), *The Handbook of Morphology* (pp. 248–271). Wiley.

Lipka, L. (1975). Review of: An Introduction to Modern English Word-Formation. *Lingua, 37*(4), 382–389.

Lieber, R. (2004). *Morphology and Lexical Semantics.* Cambridge University Press.

Lieber, R. (2012). Semantics of Derivational Morphology. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language meaning,* Vol. 3, (pp. 2098–2119).

Marchand, H. (1969). *The Categories and Types of Present-Day English Word Formation: A Synchronic-Diachronic Approach,* second edition. Beck.

Mititelu, V. B., Leseva S., & Stoyanova, I. (2023). Semantic analysis of verb–noun zero derivation in Princeton WordNet. *Zeitschrift für Sprachwissenschaft*, *42*(1), 181–207.

Procter, P. (Ed.). (1978). *The Longman Dictionary of Contemporary English.* Longman.

Rundell, M. (2002). *Macmillan English Dictionary for Advanced Learners.* Macmillan.

Schone, P., & Jurafsky D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem?. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.

Sinclair, J. (Ed.). (1987). *Collins COBUILD English Language Dictionary.* Collins.

Soukhanov, A. (Ed.). (1992). *The American Heritage Dictionary of the English Language,* Third Edition. Houghton-Mifflin.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*(1), 261–272.

Valera, S. (2014). Conversion. In R. Lieber, & P. Štekauer (Eds.), *The Oxford Handbook of Derivational Morphology* (pp. 154–168). Oxford University Press.

Worsley-Gough, B. (1932). *Public Affaires.* Victor Gollancz.

Contact information

Robert Krovetz
Lexical Research
rkrovetz@lexicalrsearch.com