

**Extrapolation of Word Meaning in English Readers:  
A Conscious Process**  
Brian N. Auriti

Much of the power of derivational morphology comes from its productivity: its capacity to create new words by combining known morphemes. Productivity gave the English language commonplace words like “strawberry” and “preview,” and continues to create words such as “Bushism” and, yet more creatively, “Googlable.” In order for this productivity to be useful, speakers must have the ability to extrapolate the meanings of the words others create. But whether this extrapolation is largely a subconscious, intuitive process or a conscious and systematic one is far from obvious. In this study I create a computer program which uses traits of English orthography to produce new “words” which appear to be morphologically complex, then use speakers’ tentative definitions for such fictitious words to examine the process of word meaning extrapolation.

Phase one of the study was to create new words which resembled those that might be created via derivational morphology, yet which the subjects were sure to be unfamiliar with. Producing such words myself seemed inadvisable, since humans tend to engage in the relatively transparent process of adding prefixes and suffixes to existing words rather than combining bound morphemes; it seemed to me that more could be learned from a study of how speakers dealt with *nontransparent* coinages along the lines of Lewis Carroll’s “slithy” and “gyre.” In Carroll’s absence, however, I chose to write a computer program called Dextrograph to generate the words for the study.

Dextrograph takes its effectiveness from the fact that English morphology by necessity gives the language certain orthographical traits which may be expressed probabilistically. For instance, -ent is a common ending for adjectives and some nouns; as a result, the bigram EN is statistically far more likely to be followed by T than by any other letter. Further, since adjectives often take the -ly suffix to form adverbs, if ENT does not end its word, the next letter may well be L. These observations prompted me to realize that a program might be able to simulate English orthography by probabilistically considering what letter was likely to follow preexisting letters in a word, and in so doing

recreate morphemes “accidentally.” Considerations of orthography (Q is always followed by U, QU is always followed by a vowel, consonant bigrams are usually followed by vowels) led me to conclude that Dextrograph would have to take into account the preceding two letters of a word in order to arrive at a reasonable choice for the next letter.

Dextrograph’s mechanics are as follows. The program starts from scratch; its only *a priori* knowledge of word structure is that words begin with word boundaries. The user then specifies a corpus in the target language for Dextrograph to read from.

Dextrograph scans the corpus, considering each trigram of each word in turn; word boundaries always count as letters for Dextrograph’s purposes. Dextrograph observes, for each trigram, what letter follows what bigram. For instance, the word “at” contains two trigrams to the program’s eyes – #AT and AT# – so Dextrograph records that #A may be followed by T, and that AT may be followed by #. Dextrograph keeps a record of the frequencies with which letters follow each bigram, as well as the frequency with which letters follow an initial word boundary.

When it has finished scanning the corpus, Dextrograph begins building a word from left to right by adding one letter at a time. The first letter is always #. This first letter is not sufficient for Dextrograph to begin applying its list of letters following bigrams, so it “jump-starts” itself by picking a next letter at random based on the frequencies with which letters follow word boundaries. It now has two letters, and uses this bigram to determine the next letter, continuing recursively until it decides that the next letter is a word boundary, at which point it stops; the result is its output.<sup>1</sup>

Once complete, Dextrograph was provided a corpus consisting of 12,879 of the first 12,880 word types from the Brown Corpus; this list was obtained from John

---

<sup>1</sup> An example might clarify this process. Suppose Dextrograph is given a two-word corpus, “crab” and “crybaby.” Dextrograph begins its output word with #, then adds a C since the words in its corpus begin with C with 100% probability. Seeing #C, it then adds R, since #C is followed by R in the corpus 100% of the time. From #CR, Dextrograph considers CR and sees that it is followed by A half the time and Y half the time. It picks the next letter according to the same proportion; suppose it picks A. Then it has #CRA, and picks B as the only possible next letter: #CRAB. But AB may be followed either with # or with Y, with equal probability. If Dextrograph picks # it is done, and outputs CRAB. If it picks Y it has formed #CRABY, and proceeds to add a # since BY is always followed with # in the corpus; its output is then CRABY. Possible outputs if Dextrograph had earlier picked Y instead of A are CRYBAB and CRYBABY.

Goldsmith's website, and the hyphenated word "middle-class" excised. cursory analysis of Dextrograph's output based on this corpus confirmed the initial supposition that its technique would preserve morphemes at a relatively high frequency. Two dozen random words of its output include such as SHOMENS, RED, EIGHT, PROQUICK, JUDDER, and SHEATING, all with recognizable English morphemes. When Dextrograph was employed to create 25,000 word tokens and the results were given to Goldsmith's morpheme-finding program Linguistica as a corpus, a number of real English suffixes, including -ed, -ion, and -ent, were found in the data at high frequencies.<sup>2</sup> This confirmed that Dextrograph's results not only "looked like" pronounceable English on orthographic grounds, but indeed bore some morphological similarity to English as well.

A selection of twelve words from Dextrograph's output were chosen for the second phase of the study. Words were chosen based on their orthographical and (apparent) morphological resemblance to English in the researcher's opinion, but words bearing clear and obvious relationships to particular English words were eschewed. The words used in the study may be found on the attached sheet.

In phase two of the study, subjects were given a list containing the twelve words. They were told that they were "not likely to know" the words on the list and instructed to attempt to guess the meanings. Requests for clarification were met with instruction to choose "whatever meaning seems right for the word." Once meanings had been assigned to each word, the subjects were asked to divide each word into "meaningful chunks," and given the example of splitting "previewed" into "pre/view/ed." The experimental reason subjects were asked to divide the words was to provide a means to tell whether subjects' recognition of words' morphology was conscious or subconscious. Finally, subjects were informed that the words were in fact imitations of English constructed by computer. None expressed surprise at this, though all agreed that the words were good imitations.

Perusing the thirty-six data points reveals that subjects' definitions of words generally fell into one of three categories. Definitions were either based mostly on

---

<sup>2</sup> Linguistica also identified a number of suffixes not found in standard English, including -e, -d, and -ter. These "suffixes" are common English word endings that Dextrograph preserved at high rates.

morphological criteria, based mostly on phonetic criteria, or (more rarely) fabricated out of whole cloth. I shall consider these categories one at a time.

Definitions of the first category appeared to be constructed when subjects applied apparent morphemes found in the Dextrograph words to postulate meanings for those words. For instance, every subject used this method in defining “propsych;” indeed, all made use of the word “psychological” in their definitions. About 53% of the definitions are thought to have been constructed in this way.

Definitions of the second category were formed when subjects sought near homophones of Dextrograph’s words and assigned the homophone’s meaning to the word. The definition “a mournful song” for “allogy,” for instance, transparently comes from the near-homophone “elegy,” and “a trench” for “gulant” is thought to stem from the phonetically similar “gully.” About 28% of the definitions seem to have originated from homophones or phonetically-similar words.

The remaining 19% of the definitions fall into the third category, having apparently been constructed according to the respondent’s imagination. One clear case is one respondent’s definition of “surgeole” as a particular dance move; a different respondent defined “apiry” as “a shade of blue,” for no clear morphological or phonetic reason.

After the surveys were administered, respondents were informally interviewed about their reasons for their answers. For each of the category-2 or -3 answers on which respondents commented, they were clearly aware of having consciously used homophony or imagination respectively. Says one respondent: “I pronounced [gulant] ‘GULLant,’ and it made me think of ‘gully.’” Another says: “I had no idea [about ‘apiry’]. I just thought ‘ditzzy’ sounded good.”

Whether category-1 answers were similarly conscious, however, is far less clear. From a comparison of how respondents defined words with how they split them in category-1 cases, it is clear that respondents were not always able to elucidate where the morphemes they had used began and ended. One subject, who defined “propsych” as “in favor of psychological analysis,” split the word “pro/psy/ch;” similarly, each subject

apparently saw “calverial” morphologically as some descendant of *cheval*, but each split the word in a different way. About one quarter of category-1 definitions are accompanied by implausible splittings. This could be read to suggest that when morphology influences definitions, it sometimes does so on a subconscious level.

I believe, however, that this observation says less about a subconscious grasp of morphology than it does about the subjects’ unaccustomedness to breaking words into morphemes. It is difficult in the extreme to believe that a respondent using the word “psychological” in a definition of “propsych” could have been unaware of the presence of the root “psych.” And if subjects subconsciously use morphological data to form definitions of unfamiliar words, why was the prefix *pro-* not reflected in the (category-1) responses of those subjects who split the word *pro/psych*?

Indeed, there is no case in the data where a respondent gave a morphologically-justifiable definition of a word, yet was clearly unaware of the presence of the morphemes. The two subjects giving category-1 definitions of “*afféal*,” for instance, both cited “*fealty*” as inspiration for their answers – in spite of one subject not having preserved the morpheme “*feal*” in her splitting of the word. This strongly suggests that implausible splitting of words should not be taken to mean that respondents have subconscious knowledge of morphology.

Unfortunately, splitting of words was originally intended as the means by which subconsciousness of morpheme extraction could be ascertained. Barring its use as a criterion, the interviews would seem to be the only way that a lack of consciousness of use of morphemes could be ascertained. The interviews failed to bring to light any such instances, and therefore it is reasonable to assume by default that when morphological criteria are used to determine word meanings, they are used consciously.

I conclude that whichever of the three means an English speaker uses to infer word meaning out of context, the inference is made consciously and without influence from subconscious recognition of morphemes. This conclusion stems primarily from the failure of this study to find positive evidence in favor of the alternative hypothesis (that assignment of word meaning can occur subconsciously). The objection could be made,

however, that as all respondents were college-aged, their conscious recognition of familiar morphemes and phone sequences in Dextrograph's words was inevitable. I do not entirely agree with this objection; there were clearly morphemes and homophones in the data which respondents could have recognized but failed to<sup>3</sup>. However, I nonetheless believe that this study could be bolstered by application to a younger set of subjects, whose vocabulary (and hence the set of morphemes and phone sequences they might recognize consciously) would be smaller.

---

<sup>3</sup> For instance, "gulant" contains *gules*, a heraldic term for red; "behorts" contains *hort*, which is used in words such as "horticulture" to refer to plant life; and "cativer" may be pronounced as a close homophone of "cadaver."

## Data<sup>4</sup>

**AFFEAL:** *To affect with charisma; aff/eal. To pledge loyalty to; af/feal. Appealing to the emotions; af/feal <appeal>.*

**ALLOGY:** *A confession of truth; al/log/y. Mournful song; al/logy <elegy>. A big collection of data; al/log/y.*

**APIRY:** *Blissfully ignorant; ditzzy; apir/y. To aspire to a goal; a/piry <aspire>. A shade of blue; apir/y.*

**BEHORTS:** *Diminishes value; be/hort/s. Acts vocal in support of someone; be/hort/s. Like a cohort, only someone who is against your conspiracies; be/horts.*

**CALVERIAL:** *Chivalrous; cal/ver/ial. Pertaining to mounted combat; calver/ial. Battle term; cal/verial.*

**CATIVER:** *A dip, as in some surface; cat/iv/er. One who organizes objects into common sections; cat/iver. Military term; ca/tiv/er.*

**GULANT:** *Proud; gul/ant <gallant>. An agent promoting adherence; gul/ant <gluant>. A trench; gu/lant <gully>.*

**PROPSYCH:** *In favor of psychological analysis; pro/psy/ch. To perform a battery of tests to determine the psychological state of the subject; pro/psych. Impeding psychological processes; pro/psych.*

**RATIVERNEOUS:** *Engulfing; overwhelming; rat/ive/r/ne/ous <cavernous>. Transmitted into a vein with a controlled flow; rati/ver/neous <intravenous>. Being so ravenous that one has the characteristics of a rat, such as hording [sic], stuffing cheeks with food, and scurrying; rativ/ern/eous <rat + ravenous>.*

**STRATICE:** *Strategy; strat/ice. An isolated segment of a single stratum; strat/ice. A person's valor; stra/tice <status>.*

**SURGEOLE:** *A dance move, in specific one arm is thrown in the air, a step is taken with the symmetrical foot, and the other foot follows; sur/ge/ol/e. A mechanized microrobot programmed to assist in a medical procedure from inside the patient; surg/eole. Anything having to do with surgical procedures; surg/eole.*

**SYMPARY:** *Characterized as having sympar; vengeful; sym/par/y. Equality; sym/pary. A rebuff to a show of sympathy; symp/ary.*

---

<sup>4</sup> Italicized responses are category-1, morphemic; underscored responses are category-2, phonetic, and accompanied by the apparent near-homophone; other responses are category-3, imaginative.