Data Collection Guide for NSF Roots Project

Introduction

The goal of data collection is to observe five categories for any given root meaning and catalog both their position within the semantic paradigm and also the morphological relationships between elements in the paradigm:

- Simple state term - The term that describes the relevant state without entailment of a change over time, space, or another scale (if extant) (e.g. *red* in *The ball is red*).
- Intransitive change-of-state term ("inchoative") - The term that describes a change into the simple state, but not necessarily a cause (e.g. *redden* in *The ball reddened*).
- Transitive change-of-state term ("causative") - The term that describes a caused change into the simple state (e.g. *redden* in *John reddened the ball*).
- Derived state term - The term that describes the state of having been changed into the simple state (e.g. *reddened* in *The ball is reddened*).
- The underlying root - In some languages all forms may be derived, meaning there is an underlying root. It may be bound in some cases or possibly free (e.g. a nominal root).

The first four categories are related by monotonic semantic complexity: the inchoative is based on the simple state, the causative on the inchoative, and the derived state on the causative (or inchoative) (using lexicalized event structures in the style of Dowty 1979 and Rappaport Hovav and Levin 1998 for illustrative purposes, though nothing hinges on this):

(1) a. Simple stative: [x BE STATE]
  b. Inchoative: [BECOME [x BE STATE]]
  c. Causative: [y CAUSE [BECOME [x BE STATE]]]
  d. Derived stative: [x BE [y CAUSE [BECOME [x BE STATE]]]]

The underlying root serves as the (typically bound) root upon which any of the first four forms are derived if such a more basic form exists in a language, though when this root is bound its meaning is often harder to exactly discern and give an event structural analysis for. The default assumption is that its meaning is simply that of the relevant state-denoting lexical semantic root, though this need not necessarily be the case. However, our goal in collecting data on underlying roots when they exist is largely for coding purposes regarding the morphological paradigm.

The primary hypothesis we are testing is whether there is a distinction among the classes of roots related to change-of-state verbs between so-called "property-concept" roots (as per Dixon 1982) and "result" roots (roughly in English the difference between Levin's 1993 deadjectival vs. non-deadjectival change-of-state verb roots). The former, we propose, generally have simple state terms while the latter lack them. The root meanings we examine in this study are the following (given as an adjective or verb, with synonyms or hypernyms considered in that study given in parentheses), chosen as either the most recurrent of Dixon's meanings in his typological study for property concept roots or for result roots the meanings that seem most likely to recur in other languages:

(2) **Property concept roots**

a. *Dimension*: large (big), small, short, long, deep, wide, tall (height)

b. *Age*: old (age)

c. *Value*: bad (worse), good

d. *Color*: white, black, red, green, blue, brown

e. *Physical Property*: cool, cold, warm, hot, dirty, dry, wet, straight, hard, tough, soft, tight, clear, clean, smooth, sharp, sweet, weak, strong

f. *Speed*: fast, slow

g. *Human Propensity*: angry, calm, scare (frighten), sick, sad (depress), hurt, tire, embarrass, entertain, surprise, worry, please

(3) **Result roots**

a. *Entity-specific Change of State*: burn, melt, freeze, decay (rot), swell, grow, bloom (flower, blossom), wither (wilt), ferment, sprout (germinate), rust, tarnish

b. *Cooking Verbs*: cook (bake, fry, roast, steam), boil

c. *Breaking Verbs*: break, crack, crush, shatter, split, tear (rip), snap

d. *Bending Verbs*: bend, fold, wrinkle (crease)

e. *Verbs of Killing*: dead/die/kill, murder, drown

f. *Destroying Verbs*: destroy (ruin)

g. *Verbs of Calibratable Change of State*: go up (rise, ascend, increase, gain), go down (fall, drop, descend, decrease, decline), differ

h. *Verbs of Inherently Directed Motion*: come, go, go in (enter), go out (exit), return

In a nutshell, the primary theoretical hypothesis is that the idiosyncratic lexical semantic root of property concept verbs describes a simple state that lacks change as part of its meaning, and these will be lexicalized as simple stative forms, whereas the idiosyncratic root of result verbs will entail change as part of its meaning, and will not be lexicalized as a simple stative form.

Data Collection Methodology — Using Grammars and Dictionaries

Here we present our data collection methodology, developed and refined over the course of our actual data collection to address a variety of analytical issues that arose during the process. Our spreadsheet consists of one or more rows per each root, with one column for each form collected, a column for a gloss of each, a column for a translation of each, a column for a bibliographic reference for each, a coding column that marks its morphological relationships to all other forms in the same row, and a general notes column for general free-form grammatical information or coding/data collection choices.

Step 1. The first step is to examine the grammar(s) of a given language, with a focus on morphology, especially verbal morphology, that encodes the relevant event structural notions, and also the more complex periphrastic means of expressing the same notions. The following questions serve as guides for learning about the relevant morphosyntactic system of the language:

(4) a. Simple stative: are there forms that directly predicate such states? How are they related morphologically to other parts of the paradigm? Is there a copula in the language that is needed? Is predication expressed possessively (e.g. *have hunger*)?

b. Inchoative: are there specific verbal forms for expressing non-caused changes? How are they related morphologically to other parts of the paradigm (e.g. is there inchoative morphology like *redd-**en*** or anticausative morphology)? Are there more or less productive ways of expressing this morphologically? Is there a periphrastic form (e.g. *become red*)? If there are multiple options, do we have information about a meaning difference?

c. Causative: are there specific verbal forms for expressing caused changes? How are they related morphologically to other parts of the paradigm (e.g. is there causative morphology as in Japanese and what does it apply to)? Are there more or less productive ways of expressing this morphologically? Is there a periphrastic form (e.g. *make red*)? If there are multiple options, do we have information about a meaning difference?

d. Derived stative: are there specific verbal forms for expressing being in a state of having undergone a change into a given state? How are they related morphologically to other parts of the paradigm (e.g. is there participial morphology as in *redden-**ed*** what does it apply to)? Are there more or less productive ways of expressing this morphologically? Is there a periphrastic form (e.g. a relative clause like *the ball that got reddened*)? If there are multiple options, do we have information about a meaning difference?

e. Underlying root: If there is one, is it free or bound?

Step 2. The next step is to look in the dictionaries for relevant data, bearing in mind the regular processes discovered in Step 1. The steps here will be the following for any cell in the spreadsheet:

- For any given form, if there is a listed, specific lexical form, we take that as the relevant datum, even if there are also other, derived forms.

- Failing that, if the grammatical resources are clear that there is a productive morphological process for making words of that function and we are confident in these resources, we construct the hypothetical form but prefix it with a @ to mark it as "Hypothetical". The point of marking hypotheticals with a diacritic is so that they can be left out of further analysis depending on the kind of analysis being done. Creating hypothetical forms, however, creates issues as to how to figure out the exact overt morphological form since there could be confounding factors such as phonology. Our interest is the existence of and paradigmatic relationship of the item within a given paradigm, so if we do not have the exact form it does not impact our goals. We thus adopt the following guidelines:

  - If the morphological process for forming the word is clear and there is a citation form for it, but there is a lot of phonology that might affect the surface form, we can code the citation form without taking the phonology into account, noting as much in the notes (e.g. *destroy-en* if we did not known that for *destroy* the suffix *-en* is realized as *-ed*).

  - If there are multiple morphological processes for a given derivation and the grammar is not very clear on which roots undergo which processes, we code the form using just a gloss (e.g. *red-CAUSE*) and indicate as much in the notes.

  - Any inflectional morphology such as agreement or tense that is irrelevant for our purposes but typically required for a citation form in that language can be either (a) whatever specific form we happened to find attested, (b) the default citation form of the language (e.g. third singular masculine agreement) if we know it, or (c) we just give the derived root without inflection, noting as much in the notes.

- Failing that, if there is a listed, specific periphrastic form, we take that as the relevant data.

- Failing that, if our grammatical resources are clear that there is a productive periphrastic process for making words of that function and we are confident in these resources, we construct the hypothetical form but prefix it with a @ to mark it as "Hypothetical" (and similar caveats as above apply if we can't figure out what the exact surface form is regarding inflectional morphology or the phonology).

- Failing that, we leave the cell blank, indicating either that no such data exists or at least that we simply have no evidence for it.

The motivation for privileging morphological processes over syntactic processes (e.g. *redden* over *make red*) is that typically when both exist, the morphological processes have meanings more like lexical forms in other languages and the periphrastics have meanings that are more general in nature; it is the former we are most interested in. Similarly, the motivation for privileging listed forms over productive morphological forms (e.g. a listed Japanese morphological causative over a derived *-sase-* causative) is that typically when both exist, the listed forms have meanings more like lexical forms in other languages and the productive forms have meanings that are more general in nature; it is the former we are most interested in (see e.g. the literature on causatives; Shibatani and Pardeshi 2001). However, a note on simple states: we will take the unmarked case of a simple stative term to either be a bare predicate (e.g. adjectival or verbal) or a bare predicate with a copula (e.g. as with adjectival forms). Yet in some languages simple states are regularly derived from nouns, and in these cases we may get a possessional encoding like *have hunger* as opposed to *is hungry*. To take this into account, we have the following guidelines for coding the simple state:

- If the simple state term is a predicate on its own we just code the actual form, and if a copula is required we don't code that. That's considered still unmarked.

- If the grammar is clear that the possessive form is the "normal" way to do it, we take it, but we code the verb with it so it's clear. This would be the marked case.

- We try to verify that the same form with a coupla would *not* have the same meaning, if we can (e.g. *have hunger* and *is hunger* are wildly different, and the latter is clearly *not* the sort of simple state reading we're interested in).

- Result root nominals (e.g. *have a fold*) are to be ignored, the degree to which we can verify that they are not really simple states.

Some cases exist where there seems to be a conflict between the semantics of the dictionary translation and the morphology, e.g. in some languages a form translated as *broken* — thus suggesting that it is the derived stative form — serves as the input for the inchoative and causative forms. Thus morphologically it does what we would think a simple stative form does but semantically it

is translated like a derived stative form. Quite likely it really is a simple stative form, and the fact that it's translated as derived is because the dictionary writer didn't have a good word for that concept since it doesn't exist in English. However, for our purposes we take the translations as literally as possible and code those as derived stative forms, so as not to conflate them with something a dictionary writer makes crystal clear is not a derived state. During analysis of the collected data we can algorithmically find such "odd" cases by looking for the weird "backwards" derivational relationship, and reclassify *en masse* later, or study these as a separate phenomenon. For more on this see below for the morphological coding procedure.

Finally, there are also a few cases where we have made special decisions about semantics based on careful thinking during the data collection. These include the following:

- We have decided that *old* and *rusty* are result states rather than simple states despite being basic adjectives in English, for the reason that semantically you just can't be old or rusty without a process over time. Thus terms translated as these are coded as result state forms.

- Some object experiencer psych-verbs, such as *frightened*, are such that their simple stative form looks deverbal. But we assume that there's a stative verb *frighten* that this is derived from (e.g. *Bears frighten me* need not have an eventive reading, but may instead describe a stative condition). This is distinct from causative *frighten* (e.g. *That bear frightened me just now*) which will be in our paradigm. Thus the two will be listed as unrelated morphologically. The result state adjetives will be derived from the causative verb.

- Some roots are glossed as both simple and result state. We will treat those as simple states, since the result state entails the simple state. In general we trust the semantics.

Data Collection Methodology — Using Grammars and Native Speakers

In some cases we employed a native speaking linguist to get the appropriate data for us. Our goal was not to make them do everything for us, but rather to have them understand the basic project and give us the forms that intuitively best fit each cell in the paradigm. Here the methodology is:

Step 1. Explain the basics of the project to them in terms of the data we're collecting and our priorities (lexical over periphrastic, etc., with naturalness also privileged), and let them fill out a survey with the relevant roots. In the meantime, the researcher reads the grammars as above to understand the language.

Step 2. A project member enters the data collected from the linguist into the spreadsheet, taking whatever analytical information they provided into account but also filling in analytical information from the grammar as needed, making some educated guesses (analytic information being morphemic breakdowns, glosses, etc.)

Step 3. The project member reviews the final data with the informant to ensure everything looks accurate.

In this case many of the caveats about phonology, inflection, and hypotheticals presumably will not arise, but to the degree that they do the same procedures above should be followed.

Data Collection Methodology — Using Fieldworkers

In some cases we asked a linguistic fieldworker to collect data while in the field. In this case we explained the basics of the project to them and provided them with the relevant root names, but otherwise left it to the fieldworker to collect the data in a manner they best saw fit given their language and their specific community and consultants. After data collection, a project member sits down with them to enter the data into our spreadsheet as above.

Coding Scheme for Relations among the Paradigm

We are interested in coding not just the forms and their semantics, but also the morphological relationships between forms, using these as a guide for relative markedness and also as a way to reconstruct a purely morphologically-based database as opposed to one that uses semantics as a basis for classification as described above. Our starting point for a coding scheme is Haspelmath (1993), which looked at pairs of causative and inchoative forms and coded them in terms of their morphological relationships to one another. However, unlike Haspelmath, we are dealing with a potentially five-way paradigm as opposed to a binary paradigm, so his relational terms are not entirely appropriate since we have more relata and also the possibility that there may be some derivational distance between items within the paradigm (e.g. the causative may be based on the inchoative, which is based on the simple state, but this means the causative is based indirectly on the simple state). To this end, our coding scheme will involve a five letter sequence that will describe the relationship of any one term to all of the others, in the following order:

(5)

1	2	3	4	5
underlying root	simple state	inchoative	causative	derived state

Thus for any given form X its associated five character sequence will include its morphological relationship to each other term in the paradigm where the first character will relate it to the underlying root, the second to the simple state, the third to the inchoative, the fourth to the causative, and the fifth to the derived state. (Since X will be one of these terms a special dummy code is introduced for relating a form to itself to maintain a consistent five character sequence for all terms.) The possible values for each position in the five character sequence will be:

(6) For a given lemma X in relation to a relatum Y:

a. i - X is the direct input to a rule forming Y (e.g. *red* is the direct input of *redden*).

b. d - X is the output of a rule with Y as direct input (e.g. *redden* is the output of a rule on *red*). Each i code for X will thus be matched up with a corresponding d code on Y and vice versa.

c. t - X is the output of a rule with a form Z as a direct input, where Z is in turn derived from Y as direct or indirect input (e.g. *reddened* is the output of a rule on *redden*, which is derived from *red*, so *reddened* is t-related to *red* and vice versa).

d. l - X and Y are a labile pair (e.g. inchoative *redden* and causative *redden* are the same surfacing form).

e. e - X and Y are equipollent (e.g. each stands in the d relation to the same other term, but are not labile).

f. u - X and Y are unrelated (e.g. they are suppletive).

g. n - Y is unattested (e.g. there is no underlying root, so nothing for X to be related to).

h. s - When X literally is Y (this is for relating something to itself, since every code involves five characters corresponding to the relationship of X to the full paradigm, and one of those is for X itself).

Note that these codes are meant to be used purely morphologically, even if it leads to potentially contradictory coding. For example, *reddened* could be derived from either the inchoative or the causative morphologically, and it's not clear which it is based purely on this (even if we know semantically that it is derived from the inchoative), and as such we code it as though it were derived from both though this is in principle not possible:

(7)

	underlying root	**simple state**	**inchoative**	**causative**	**derived state**
Form:		red	redden	redden	reddened
Code:		nsiit	ndsli	ndlsi	ntdds

The following represent examples of how the codes might use, using dummy data. Here's what the "canonical" paradigm might look like if the morphology matched the semantics perfectly, and the result state form is derived from the causative (and there is no underlying root):

(8)

	underlying root	**simple state**	**inchoative**	**causative**	**derived state**
Form:		red	red-INCH	red-INCH-CAUSE	red-INCH-CAUSE-STATE
Code:		nsitt	ndsit	ntdsi	nttds

Here's the same but supposing there is an underlying root that only serves the simple state:

(9)

	underlying root	**simple state**	**inchoative**	**causative**	**derived state**
Form:	red-	red-STATE	red-STATE-INCH	red-STATE-INCH-CAUSE	red-STATE-INCH-CAUSE-STATE
Code:	sittt	dsitt	tdsit	ttdsi	tttds

Here's the a case of an underlying root where everything is derived directly from it:

(10)

	underlying root	**simple state**	**inchoative**	**causative**	**derived state**
Form:	red-	red-STATE1	red-INCH	red-CAUSE	red-STATE2
Code:	siiii	dseee	desee	deese	deees

More normal might be this instead, though, where the result root is derived from one of the verbal forms:

(11)

	underlying root	**simple state**	**inchoative**	**causative**	**derived state**
Form:	red-	red-STATE	red-INCH	red-CAUSE	red-CAUSE-STATE
Code:	siiit	dseeu	deseu	deesi	tuuds

Here's an expected result root paradigm, where the inchoative is basic:

(12)

	underlying root	**simple state**	**inchoative**	**causative**	**derived state**
Form:			break	break-CAUSE	break-CAUSE-STATE
Code:			nnsit	nndsi	nntds

And here is one where the causative is basic:

(13)

	underlying root	**simple state**	**inchoative**	**causative**	**derived state**
Form:			break-INCH	break	break-STATE
Code:			nnsde	nnisi	nneds

And here is a case of suppletion:

(14)	**underlying root**	**simple state**	**inchoative**	**causative**	**derived state**
Form:		dead	die	kill	killed
Code:		nsuuu	nusuu	nuusi	nuuds

A language may have subparadigms (e.g. in English the derived state for *hang* is *hung* in some dialects, with an older English stem vowel alternation, rather than *hanged* via the productive participle formation). If our grammatical resources are clear that they are subparadigms, we code them as derived in the appropriate way. But if we have no idea if they are, we treat them as suppletive.

References

Dixon, R. M. W. 1982. *Where Have all the Adjectives Gone!*. Berlin: Mouton Publishers.

Dowty, David. 1979. *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In B. Comrie and M. Polinsky, eds., *Causation and Transitivity*. Amdersdam: John Bejamins.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press.

Rappaport Hovav, Malka and Beth Levin. 1998. Building verb meanings. In M. Butt and W. Geuder, eds., *The Projection of Arguments: Lexical and Compositional Factors*, pages 97–133. Stanford: CSLI Publications.

Shibatani, Masayoshi and Prashant Pardeshi. 2001. The causative continuum. In M. Shibatani, ed., *The Grammar of Causation and Interpersonal Manipulation*, pages 85–126. Amsterdam: John Benjamins.