Grammatical Properties

David R. Mortensen

February 15, 2024

Words can be inflected according to different dimensions[1]:

(1)
a. Part of speech
b. Switch-reference
c. Information structure
d. Politeness
e. Argument marking (agreement, etc.)
f. Valency
g. Voice
h. Finiteness
i. Mood
j. Interogativity
k. Evidentiality
l. Polarity
m. Tense
n. Aspect
o. Case
p. Animacy
q. Person
r. Number
s. Definiteness
t. Deixis
u. Gender and noun class
v. Possession
w. Comparison

Each of these dimensions, or properties, can take on different attributes. For example, Mood can take the values

(2)
a. Indicative
b. Subjunctive
c. Realis
d. Irrealis
e. Purposive
f. Non-Purposive
g. Imperative-Jussive
h. Conditional
i. Intentive
j. Potential
k. Likely
l. Admirative
m. Obligative
n. Debitive
o. Permissive
p. Deductive
q. Simulative
r. Optative-Desiderative

In any given language, only a subset of the dimensions and (for any given dimension) only a subset of the values are going to be expressed through morphology.[2]

A paradigm, simply, corresponds to the Cartesian product of the sets relevant values from one or more relevant dimensions. Take Case × Number. In German, this would be, roughly,

$$\{\text{Nom}, \text{Gen}, \text{Acc}\} \times \{\text{Sg}, \text{Pl}\}$$

.

[1] John Sylak-Glassman. The composition and use of the universal morphological feature schema (unimorph schema). Published online, June 2016. URL `https://unimorph.github.io/doc/unimorph-schema.pdf`

[2] This fact is one of the motivations for the Sapir-Whorf hypothesis, which is widely held among non-linguists and states that the structure of language constrains the structure of thought. The linguists retort is that—even when a language cannot express one of these attributes through one grammatical means, it can always express it through another.

In the resulting matrices, there will often be rows or columns that are logically impossible. For example, assume a matrix

$$\{1, 2, 3\} \times \{\text{Sg}, \text{Pl}\} \times \{\text{Incl}, \text{Excl}\}.$$

In addition to the "good" cells, this would yield tuples like $\langle 1, \text{Sg}, \text{Incl}\rangle$. It is logically anomalous to imaging a person speaking for themselves alone and simultaneously including the INTERLOCUTOR (the person they are speaking to), which is what $\langle 1, \text{Sg}, \text{Incl}\rangle$ would mean.

But when we speak of a paradigm, we are usually not referring just to the properties (the values in these dimensions). We usually mean the forms of a particular lexeme for each of the cells in the matrix of properties. The questions we are asking, morphonologically speaking, are like the following:

(3) **Representation**

a. What is the proper formulation of the properties. For example, how should person and clusivity be represented?

i. Person: 1, 2, 3; Clusivity: Incl, Excl

ii. ±me, ±you

Treating all of the properties as PRIVATIVE (present or absent) requires breaking a dimension into two dimensions. It also means that there are more features with less principled combinations. Introducing BINARY features makes the formalism more complicated but also allows for more elegant formulation of some paradigms.

(4) **Generation**

a. Given a tuple (we will often treat these as sets) of morphological properties and a lexeme, what is the corresponding word form? In other words,

$$\textsc{Inflect}(x) = \arg\max_{\varphi} p(\varphi | x, \ell) \tag{1}$$

where x is a set of properties, ℓ is a lexeme (more later), and φ is a word form.

b. What information is needed to compute INFLECT?

i. Is ℓ just a lemma or underlying form?

ii. Is ℓ another form of the lexeme?

iii. Is ℓ a set of other word forms (so-called PRINCIPAL PARTS[3])?

(5) **Analysis**

a. Given a word form, what are the lexeme and properties that most likely correspond to it,

$$\textsc{Analyze}(\varphi) = \arg\max_{x,\ell} p(x, \ell | \varphi) \tag{2}$$

[3] PRINCIPLE PARTS are the cells in a set of paradigms that are sufficient for predicting the word forms in the rest of the cells in the paradigms.

or, in another formulation, what are the lexeme and properties that are compatible with φ? This formulation is necessitated by the phenomenon of syncretism[4].

Note that this way of approaching words is different than the one from the first unit. There, we built up complex signs by combining simple signs, with signifier and signified being constructed, piece by piece, in lockstep. This may be called a compositional approach to morphology. The alternative version presented here is to *realize* a signifier based on an already complete signified consisting of a lexical meaning and a collection of properties. This kind of approach is termed realizational.

Representation

In order to understand the issue of representation, let’s condisider the case of Totonac verbs. Table 1 presents a partial paradigm for ‘open’. Certain formatives are color-coded. These are related to the first and second persons. Note that *k-* occurs always in 1sg and 1pl.excl. Except in the perfective, *-á:w’* occurs in the 1pl.incl and 1pl.excl. In the imperfective, *-á:’* occurs in the 2sg and 2pl. In the 2pl, *-ti’t* always occurs.

	Imperfective	Perfective	Perfect	Progressive
1sg	ktalajkí:y	ktalájki:lh	ktalajki:ní:’t	ktalajkí:mah
2sg	talajkí:ya’	talájki’	talajki:ní:’ta’	talajkí:pa:’t
3sg	talajkí:y	talájki:lh	talajki:ní:’t	talajkí:mah
1pl.incl	talajki:yá:’w	talajkí:’w	talajki:ni:’tá’w	talajki:má:’w
1pl.excl	ktalajki:yá:’w	ktalajkí:’w	ktalajki:ni:’tá’w	ktalajki:má:’w
2pl	talajki:yá:’ti’t	talajkí:’ti’t	talajki:ni:’tá’nti’t	talajki:pá:’ti’t
3pl	talajkí:qó:’y	talajki:qó:’lh	talajki:qo:’ní:’t	talajki:ma:qó:’lh

Table 1: Paradigm for the Totonac very *talajkí:y* ‘open’

One way of representing person, number, and clusivity is as follows: This treats properties as sets of atomic features that can be present or absent (privative features). But properties can also be represented as binary feature

1sg	{1, S}
2sg	{2, S}
3sg	{3, S}
1pl.incl	{1, P, INCL}
1pl.excl	{1, P, EXCL}
2pl	{2, P}
3pl	{3, P}

Table 2: Totonac person-number-clusivity as privative properties

[4] Syncretism is the phenomenon in which two or more cells of a paradigm contain the same word form.

vectors. Here we use linguists notation, in which the values (+ for 1 and − for 0 occur before names for the dimensions in the vector) as shown in Table 3.

Table 3: Totonac person-number-clusivity as binary feature fectors

1sg	[+me, −you, −pl]
2sg	[−me, +you, −pl]
3sg	[−me, −you, −pl]
1pl.incl	[+me, +you, +pl]
1pl.excl	[+me, −you, +pl]
2pl	[−me, +you, +pl]
3pl	[−me, −you, +pl]

This representation is convenient, because it allows us to say that

(6) a. *k-* occurs in all cells that are [+me, −you]
 b. *-á:'w* occurs in all cells that are [+me, +pl]
 c. *-a'* occurs in all cells that are [−me, +you, +imperfective]
 d. *-ti't* occurs in all cells that are [−me, +you, +pl]

and so on. You can express all of these rules using the privative notation, but you sometimes have to have redundant rules. For example, you would have to have separate rules to express the fact that *k-* realizes 1sg and 1pl.excl.

Generation

There are at least two ways to view generation based on morphological properties:

- The input is an identifier for a lexeme and a set of properties. A cascade of rules spell out (realize) the root, then the properties. The rules are conditioned on the properties in the set (and possibly the word form that has been spelled out so far).
- The input is a lemma and a sequence of tokens, each representing a property. A sequence-to-sequence model transduces this input sequence into a fully inflected word form.

At some level, these two approaches are versions of the same thing; they differ in implementation. We will talk about both of these approaches in greater depth in subsequent lectures.

The big differences:

(7) Rule-based
 a. Interpretable
 b. No supervision is necessary
 c. Difficult to learn empirically

d. Tends to overgeneralize
e. String rewrite rules

(8) Seq2seq

a. Uninterpretable
b. Necessarily supervised (at least partially)
c. Easy to learn from data
d. Tends to overfit
e. Encoder-decoder architectures

In most formulations, the rule-based approach relies crucially upon rules being ordered into ordered blocks. The blocks are like `case` statements in many programming languages. The rules are ordered within a block based on specificity. Execution proceeds through the block, checking whether each rule's conditions are met. When they are, that rule applies and execution continues to the next block. If none of the rules with conditions can apply, a fall-back (or default) rule applies. Sometimes this rule does nothing.

For example, in the Totonac example, we must assume that their are two rules in the last block that add suffixes to [+me, +pl] verbs:

(9) $$\begin{bmatrix} +\text{me} \\ +\text{pl} \\ +\text{perf} \end{bmatrix} \text{X} \rightarrow \text{Xw}$$

(10) $$\begin{bmatrix} +\text{me} \\ +\text{pl} \end{bmatrix} \text{X} \rightarrow \text{Xá:’w}$$

Rule (9) applies just in case all of its conditions are satisfied. It is more specific than (10), therefore it must be ordered first. For first personal plurals that are not perfective, the morphology always falls back to (10).

Analysis

Analysis is just the inverse of generation. It can be viewed as either a sequence-to-sequence task or as a parsing task. As a seq2seq task, it is simply generation with the input and the output flipped: the input is a fully-inflected word form and the output is a lemma followed by a sequence of property tokens. The rule-based approach can be solved with either bottom-up and top-down parsing algorithms. We will talk more about this in subsequent lectures.

References

John Sylak-Glassman. The composition and use of the universal morphological feature schema (unimorph schema). Published online, June 2016. URL `https://unimorph.github.io/doc/unimorph-schema.pdf`.