The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema)

John Sylak-Glassman

Center for Language and Speech Processing
Johns Hopkins University

jcsg@jhu.edu

June 2, 2016

**** Working Draft, v. 2 ****

Contents

1 Introduction

The fact that some languages extensively use suffixes and prefixes to convey grammatical meaning (e.g. subject-verb agreement) poses a challenge to most current human language technology (HLT). Suffixes and prefixes in such languages can more generally be called *morphemes*, which are defined as the meaningful subparts of words. The rules that languages use to combine morphemes, together with the actual morphemes that they use (i.e. suffixes and prefixes themselves), are both referred to as a language's *morphology*. Languages which make extensive use of morphemes to build words are said to be morphologically-rich. These include languages such as Turkish and can be contrasted with so-called *analytic* languages such as Mandarin Chinese, which does not use suffixes or prefixes at all.

In a language with rich morphology, it is less likely that a speaker will have encountered any given word before because in these languages, new words are frequently created through suffixation, prefixation, etc. In computational approaches that rely heavily on having encountered a word before in training data (or in some external resource such as a dictionary), languages with rich morphology are especially challenging since the likelihood of encountering any given word is lower. This is a kind of *data sparsity* problem. Although data sparsity is a problem for machines, it is not a problem for humans. Humans' ability to deal with rich morphology arises from two sources: The sensitivity to sub-word structure, and a knowledge that arises from this sensitivity about which morphemes exist and how they can be recombined to form new words. To ameliorate the problems associated with rich morphology in HLT systems, these systems must have knowledge about morphology.

While many approaches have focused on endowing HLT systems with knowledge about overt morphemes themselves (i.e. suffixes and prefixes, among other types) using either supervised (e.g. Dreyer and Eisner 2011; Durrett and DeNero 2013; Ahlberg et al. 2014, 2015; Nicolai et al. 2015) or unsupervised (Hammarström and Borin 2011 and references therein) discovery methods, the present work focuses on endowing HLT systems with knowledge about the meaning that morphemes convey. While knowledge about both morpheme form and meaning are useful for HLT systems, the meaning of morphemes has typically been specified in language-specific ways. This is not a problem for HLT systems that deal only with a single language, but to develop systems that can be applied across many languages, meaning needs to be defined in language-independent terms.

This paper presents the Universal Morphological Feature Schema (UniMorph Schema), which is a set of morphological features that functions as an interlingua for inflectional morphology by defining the meaning it conveys in language-independent terms. The features of the Universal Morphological Feature Schema have precise definitions based on attested cross-linguistic patterns and descriptively-oriented linguistic theory, and can capture the maximal level of semantic differentiation within each inflectional morphological category.

The goal of the Universal Morphological Feature Schema is to allow an inflected word from any language to be defined by its lexical meaning (typically carried in the root or stem) and by a rendering of its inflectional morphemes in terms of features from the schema (i.e. a vector of universal morphological features). When an inflected word is defined this way, it can then be translated into any other language since all other inflected words from all other languages can also be defined in terms of the Universal Morphological Feature Schema. Although building an interlingual representation for the semantic content of human language as a whole is typically seen as prohibitively difficult, the comparatively small extent of grammatical meanings that are conveyed by overt, affixal inflectional morphology places a natural bound on the range of meaning that must be expressed by an interlingua for inflectional morphology.

At present, the UniMorph Schema accounts for inflectional morphology only, not derivational morphology. Inflectional morphology occurs with very high frequency within languages that use it,

and across languages, the range of meaning expressed by inflectional morphology is a very small subset of the total meaning space that human languages capture (as noted, e.g., by Sagot and Walther 2013).[1]

Most importantly, the meaning that certain categories of inflectional morphology encode is useful for a range of HLT applications regardless of the language in which it occurs. For example, nominal case often correlates straightforwardly with semantic roles, which can aid information extraction by accurately identifying the relationship of actors to events. Evidentiality also aids information extraction by encoding speakers' sources of information for the propositions they assert (e.g. direct evidence, hearsay, other sensory evidence). Related to this, mood and modality also encode a speakers' state of mind, particularly uncertainty, and are helpful in sentiment analysis and assessing the level of confidence in whether an event actually occurred. Detailed tense and aspect marking help in determining when events occurred. As a final example, switch-reference morphemes overtly mark coreference between nouns in different clauses, which greatly simplifies the task of coreference resolution.

Following an overview of the schema in §2 and the principles behind its construction in §3 (p. 5), §4 (p. 7) discusses details of how to use features to specify the meanings of inflected words and morphemes. These details are the overarching annotation formatting guidelines for the schema. Next, §5 (p. 8) presents each dimension of meaning (i.e. each morphological category, such as number, person, tense, case, etc.) along with the features that compose it. After a brief conclusion in §6 (p. 8), Appendix 1 (p. 60) lists the dimensions of meaning along with their constituent features, both in alphabetical order for easy searching. Appendix 2 (p. 66) lists the features themselves with their dimension of meaning.

2 Overview

The Universal Morphological Feature Schema comprises 23 dimensions of meaning and over 212 features. The dimensions of meaning are morphological categories, such as person, number, tense, and aspect, which each represent a coherent semantic space within inflectional morphology. They include: Aktionsart, animacy, aspect, case, comparison, definiteness, deixis, evidentiality, finiteness, gender, information structure, interrogativity, mood, number, part of speech, person, polarity, politeness, switch-reference, tense, valency, and voice. These dimensions contain varying numbers of features, from just 2 for finiteness to 39 for case. Features represent the finest-grained distinctions in meaning that are possible within a given dimension. The UniMorph Schema's features are very similar to the annotation labels used in interlinear glossed text and as described by, for example, the Leipzig Glossing Rules (Comrie et al. 2008).

Each inflected word in any given language can be represented by its lemma gloss (as might appear in a dictionary, for example) and a vector (or set) of UniMorph Schema features. For example, the Spanish word *hablaste* can be represented as `speak`;FIN;IND;PFV;PST;2;SG;INFM. Note that this yields a mapping of *hablaste* $\mapsto$ `speak`;FIN;IND;PFV;PST;2;SG;INFM, which associates the entire inflected word with its meaning without any indication of the morpheme divisions (or segments) within *hablaste* nor how the UniMorph Schema features are distributed among those divisions. The Russian word *skazal* would have a very similar representation as `speak`;FIN;IND;PFV;PST;SG;MASC, differing from *hablaste* only in the fact that it does not mark person nor politeness features. Note

[1] Elements of derivational morphology typically occur with lower frequency and vary more across languages, ultimately covering a much broader and less easily specified subset of the total semantic space that human languages capture. However, the JHU team is currently researching which types of derivational morphology are cross-linguistically common and productive, since these will be most useful for annotating previously unseen forms.

that for both languages, the representation differs only by what distinctions in meaning the language marks. Because the meaning of features does not differ across languages, the featural representation of words from different languages is directly comparable. This is an essential feature of the UniMorph Schema that allows inflectional material to be faithfully translated and enhances comparability across languages.

3 Construction Methodology

3.1 Guiding Principles

The purpose of the universal morphological feature schema is to allow any given overt inflectional morpheme in any language to be given a precise, language-independent, semantically accurate definition. This influences the overall architecture of the schema in two significant ways.

First, the schema is responsible for capturing only the meanings of overt inflectional morphemes, which considerably limits the semantic space that must be formally described by the UniMorph Schema features. This limitation of the range of data that must be modeled makes an interlingual approach to the construction of the schema feasible.

Second, the schema is sensitive only to the semantic content of words, not to their surface form. This follows the insight in linguistic typology that "crosslinguistic comparison [...] cannot be based on formal patterns (because these are too diverse), but [must] be based primarily on universal conceptual-semantic concepts (Haspelmath 2010:665, and references therein). Due to the semantic focus of the schema, it contains no features for indicating the form that a morpheme takes. Instead, the schema's features can be integrated into systems and frameworks that can indicate the form of morphemes, such as the Alexina$_{Parsli}$ system (Sagot and Walther 2013) and the Leipzig Glossing Rules' theoretical framework (Comrie et al. 2008).

The UniMorph Schema features represent semantic "atoms" that are never decomposed into more fine-grained meanings in any natural language. This ensures, from both a theoretical and practical point of view, that the meanings of all inflectional morphemes in any language are able to be represented either through single features or through multiple features in combination (as described in detail in §3.2, p. 6).

The purpose of the UniMorph Schema strongly influences its relationship to linguistic theory. The features instantiated in the schema occupy an intermediate position between being universal categories and comparative concepts, in the terminology coined by Haspelmath (2010:663-7). Haspelmath defines a universal category as one that is universally available for any language, may be psychologically 'real,' and is used for both description/analysis and comparison while a comparative concept is explicitly defined by typologists, is not claimed to be 'real' to speakers in any sense, and is used only for the purpose of language comparison.

Because the purpose of the schema is to allow broad cross-linguistic morphological analysis that ensures semantic equality between morphemes in any given language and morphemes, words, or phrases in another, its features are assumed to be possibly applicable to any language. In this sense, features are like universal categories. However, like comparative concepts, the UniMorph Schema features are not presumed to be 'real' to speakers in any psychological or cognitive sense.

Like both universal categories and comparative concepts, each UniMorph Schema feature retains a consistent meaning across languages such that in every instance in which a feature is associated with a morpheme, that morpheme necessarily has the meaning captured by that feature (but may also have other meanings and serve other functions as well). This emphasis on semantic consistency across languages prevents categories from being mistakenly equated, as in the dative

case example in Haspelmath (2010:665), which highlights the problems with establishing cross-linguistic equivalence on the basis of terminology alone. The central problem is that what is glossed as 'dative,' for example, in one language does not necessarily bear much resemblance at all to what is glossed as 'dative' in another. This is the primary reason why the UniMorph Schema features are hand-engineered and assigned based on meaning: The features must transcend language-specific terminology.

3.2 Constructing the Schema

The first step in constructing the Universal Morphological Feature Schema was to identify the *dimensions of meaning* (i.e. morphological categories) that are expressed by overt inflectional morphology in the world's languages. These were identified by surveying the linguistic typology literature for common agreement features, and then by identifying the kinds of inflectional morphology that are typically associated with each part of speech. In total, 23 dimensions of meaning were identified.

To determine the feature set within each dimension, we found the finest-grained distinctions in meaning that were made within that dimension by a natural language by surveying the literature in linguistic typology. That is, we identified which meanings were "atomic" and were never further decomposed in any language. The reduction of the feature set in the universal schema to only those features whose meanings are as basic as possible minimizes the number of features and allows more complex meanings to be represented by combining features from the same dimension.

In addition to these basic features, some higher-level, superordinate features that represented common cross-linguistic groupings were also included. For example, features such as indicative (IND) and subjunctive (SBJV) represent groupings of multiple basic modality features which nevertheless seem to occur together in multiple languages and show similar usage patterns across those languages (Palmer 2001). These can be viewed as 'cover features' in which backing off to more basic features remains an option.

Each dimension has an underlying semantic space in which the features within that dimension are defined. To determine the underlying semantic space for each dimension, the literature in linguistic typology was surveyed for explanations that were descriptively-oriented and offered precise definitions for observed basic distinctions. A simple example is the dimension of number, in which six of eight features are defined as straightforward divisions of a quantificational scale of the number of entities. In addition to features that capture divisions of a semantic scale or concept, irreducible features which mark distinctions in the same semantic space, but without clear reference to the primary semantic scale or concept must also be included. An example of these features within number are greater plural (GRPL) and inverse (INVN) number marking. Greater plural indicates not only multiple entities, but an abundance ("various, many") or all possible entities (Corbett 2000:32-33).

Because an exhaustive survey of the occurrence of inflectional morphological categories across the world's languages is very difficult, the schema is likely not yet fully exhaustive and the authors invite input on dimensions or features that should be considered for inclusion. The primary criteria for inclusion are whether:

1. A proposed dimension represents a semantic space that is not already included in this schema, and

2. Any proposed features represent basic meanings that are not decomposed further in a natural language.

For example, the proposed feature 'direct' as a case feature that captures uninflected forms that are used as both subjects and direct objects in nominative-accusative languages would be rejected since it can be specified in terms of both nominative and accusative as NOM/ACC ('nominative or accusative').

4 Annotation Formatting Guidelines

The UniMorph Schema features are assigned according to the meaning of a given morpheme (or word). This is a key principle that must be adhered to in order to avoid the kind of terminological traps exemplified by the "dative" in Haspelmath (2010). The majority of morphemes or words will have features from only a limited number of dimensions of meaning specified. Moreover, each specified dimension will typically have a single, simple feature specified. For example, for the Spanish word *hablaste* 'you spoke' in (1), only 7 of the possible 23 dimensions are specified, and each of these is specified by a single feature.

(1) *hablaste* 'you (sg.) spoke' ↦ `speak`;FIN;IND;PFV;PST;2;SG;INFM

Dimension:	Finiteness	Mood	Aspect	Tense	Person	Number	Politeness
Feature:	FIN	IND	PFV	PST	2	SG	INFM

Dimensions need not have feature specifications, and can simply be left blank. For example, Spanish verbs do not mark evidentiality, and so no evidentiality feature is specified (nor is the dimension even indicated in the representation in (1)). Alternatively, if a dimension can take on any feature value, this can be indicated with an asterisk (*) as the value of the dimension. For example, if *hablaste* is unspecified for evidentiality because it is compatible with any evidentiality, this may be indicated by adding a column to the table in (1) labeled 'Evidentiality' and specifying the feature value as *. This kind of specification is unnecessary in glosses in interlinear glossed text, and may be filled in during later stages of analysis.

One consequence of defining features as only the most fine-grained, basic, irreducible semantic distinctions in a dimension of meaning is that complex feature specifications are sometimes needed to accurately specify the value of a given dimension. When a dimension has a meaning specification that cannot be captured with a single simple feature, the features that are used can either be:

1. Conjoined with + (e.g. X+Y, where X and Y are any two non-identical simple features)
2. Put in a disjunctive *or*-relationship with {X/Y}, or
3. Negated with *non*{X}.

As an example of feature conjunction, the inessive case, which occurs in Uralic languages (e.g. Finnish), marks both stationary location ('essive') and the spatial position of being *in* a given location. Both of these case parameters can be specified simultaneously with the complex feature IN+ESS (or, equivalently, ESS+IN).

Disjunction is necessary to capture the meaning of the direct case in Hindi, which is used for both the subject and direct object of inanimate nouns with non-perfect aspect verbs, but which cannot function in both roles simultaneously on a single noun. This calls for a disjunctive specification as {NOM/ACC}, with any number of disjoined features contained within braces and separated by a forward slash ('/').

Finally, negation is necessary to capture the meaning of cases described as an 'oblique' case. In Romanian, the oblique case covers any case relations outside the core functions of marking the subject and direct object (Blake 2001:176). The oblique case also occurs, for example, in English

pronouns such as *him, her*, and *them*, which are used as direct objects as well as the objects of prepositions that can assign more specific cases (e.g. *into* assigns IN+ALL). The oblique case in English can be specified as *non*{NOM} and for Romanian can be specified as *non*{NOM/ACC}. Formally, negation is a disjunction preceded by *non* that can contain any number of features. Both disjunctive and negative feature specifications are implicitly assumed to be procedurally non-final in the sense that it is assumed that context will ultimately resolve a disjunctive or negative specification down to a single feature specification, which may be either simple (e.g. NOM) or complex/conjoined (e.g. IN+ALL).

5 Dimensions of Meaning and Features

The following sections describe specific dimensions of meaning along with the features that comprise them. The dimensions are presented in alphabetical order for ease of reference. However, some of the dimensions are conceptually related and some will occur more commonly in languages than others.

With respect to conceptual relationships, tense, aspect, and Aktionsart all represent grammaticalizations of temporal relations and are typically marked on verbs.[2] Interrogativity, polarity, and evidentiality are all conceptually related to mood in that they encode (among other relevant distinctions) a speaker's degree of certainty about a proposition and, to some extent, attitude towards it. Person, number, and gender are properties of verbal actants which are marked directly on the verb in many languages. Case, valency, and voice are all conceptually related in that they mark (and shape) the relationship of nominal arguments to a verb. Although other conceptual relationships can be identified, the foregoing relationships may be helpful to the reader to tie together the dimensions that are discussed here.

Finally, before entering into a detailed discussion of the dimensions of meaning, it is worth identifying, impressionistically, some of the most frequently encountered dimensions of meaning. All languages syntactically differentiate parts of speech. Verbs and nouns are universally differentiated, and adjectives, pronouns, and adverbs are cross-linguistically common. Verbs typically distinguish tense, aspect, mood, finiteness, polarity, voice, person, and number categories. Nouns are often distinguished according to number, case, and gender. Adjectives often distinguish the same categories as nominal morphology, and may make additional distinctions to mark comparison. Pronouns typically also distinguish similar categories as nominal morphology, and third-person and/or demonstrative pronouns may make distinctions in deixis. Cross-linguistically, distinctions in the following dimensions are less common generally (although they may be common within a family or stock): Aktionsart, animacy, definiteness, evidentiality, information structure, interrogativity, politeness, possession, switch-reference, and valency. This assessment of frequency is entirely impressionistic and no doubt somewhat biased toward Indo-European languages. However, it is hoped that it will help readers, especially annotators, to identify dimensions of the meaning that are likely to be relevant to their language.

5.1 Aktionsart

Aktionsart refers to the "inherent temporal features" of a verb (Klein 1994:29-31), which can be seen as the linguistic correlates to how the action described by a verb unfolds in real life. The term *aktionsart* (plural *aktionsarten*; capitalization adapted to English) was originally used by

[2]Paraguayan Guaraní has been claimed to possess tense-like morphemes on nouns. See Tonhauser (2007) for a survey and evaluation of these claims using original data from fieldwork.

Agrell (1908) to refer to "secondary modifications of basic verb meanings by means of affixes," as in German *erblühen* 'to start flowering' from *blühen* 'to flower' (Klein 1994:17). Aktionsart is now used to refer to the kind of semantic distinctions that underlie the morphological distinctions that Agrell (1908) noticed. These kinds of semantic distinctions influence which morphological distinctions can be made on verbs and are sometimes marked with overt surface morphology. However, they are very often lexically encoded, and so not the domain of inflectional morphology.[3]

Figure 1, based on (37) in Cable (2008:19), presents a hierarchy of aktionsart distinctions that will guide the discussion of aktionsart features. The hierarchy incorporates the primary distinctions noted by Vendler (1957) and Comrie (1976a).

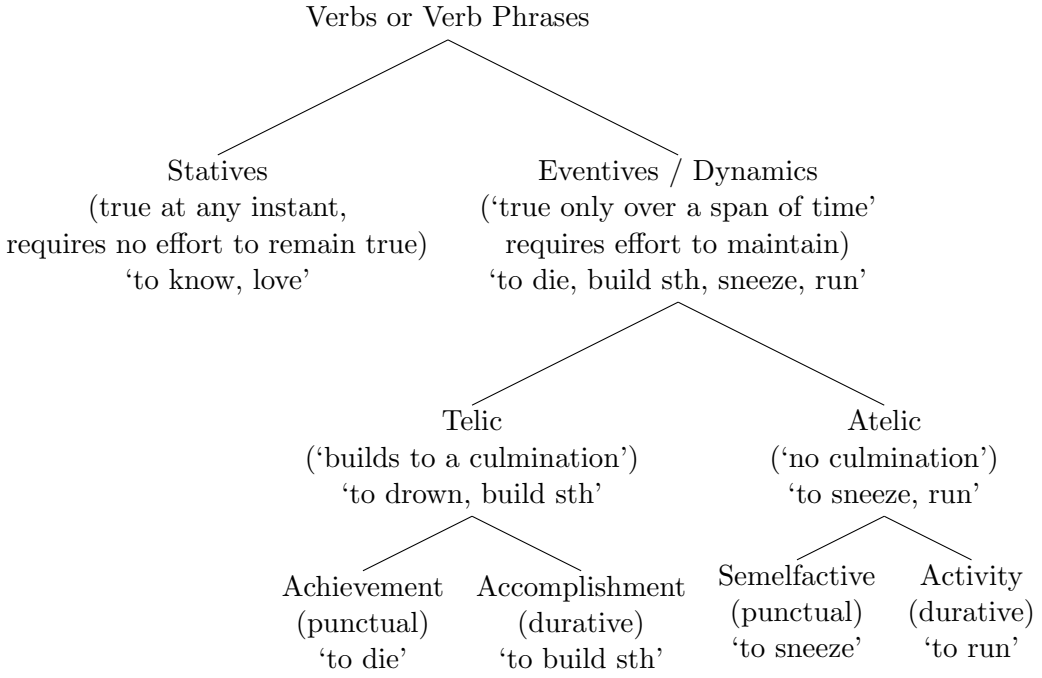


Figure 1: Hierarchy of aktionsart distinctions (adapted from Cable 2008:19)

The first distinction is between stative (STAT) and dynamic (DYN) verbs. Comrie (1976a:48-50) defines stative verbs as those whose action will continue (or continue to hold true) without any additional effort being applied. Moreover, a stative action usually continues without any internal change in the type of action that is occurring. For example, in "John knows Chris," no effort is required on John's part to continue to know someone and there is no internal dynamic to the action of knowing. However, in "John is building a shed," continued effort is required on John's part to continue building and the action of building involves different phases that progress to an endpoint.

The verb "build" is not only dynamic, it is telic (TEL). Telic verbs have a well-defined terminal point at which an action naturally terminates with a result Comrie (1976a:44-45). A test for telicity introduced by Klein (1974:106-107) is whether an action, when interrupted, can be felicitously described as having occurred. For example, if someone is drowning and they are interrupted, one cannot say "someone drowned." In contrast, if someone is playing and is interrupted, one can felicitously say "someone played." These examples, quoted in Comrie (1976a:45), illustrate the

[3]In fact, it is with some hesitation that we include this as a dimension of meaning encoded by *inflectional* morphology. These features will likely be useful for derivational morphology, and were included in previous published descriptions of the schema, so we retain a discussion of aktionsart here.

telicity of "drown" and the atelicity of "play." The action of drowning naturally terminates and yields a result whereas the action of playing need not ever necessarily terminate.

Both telic and atelic (ATEL) situations can be divided into punctual (PCT) and durative (DUR) events (Comrie 1976a:41-44). To illustrate the distinction, compare two atelic verbs, "sneeze" and "run." One can run for an hour, and for that whole time the running is uninterrupted. There are not small points in time where the action of running has ceased. However, if one says that someone sneezed for an hour, the interpretation is that multiple events necessarily had to occur, and that there were small breaks between those events in which the event was not occurring. "Run" is therefore durative because its action can extend over a time span. "Sneeze" is punctual because, lexically, it has no duration: It is a single event that takes place for an instant and the action cannot be understood as extending over a time span.[4] Forcibly extending a punctual action over a time span, for example by specifically stating a duration, forces a repetitive interpretation. Atelic punctual verbs like "sneeze" are termed semelfactive (SEMEL; Comrie 1976a:42) and atelic durative verbs are termed activities (ACTY; Vendler 1957:146). The distinction between punctual and durative telic verbs is more difficult to describe. The actions of both verbs terminate in a result, but achievements (ACH; punctual telic verbs) occur quickly and tend to mark a rapid state transition. For example, "die" represents an achievement because it marks a result that comes about instantly. A durative telic verb, such as "build," marks a result that takes place only after a span of time, and this is called, somewhat confusingly, an accomplishment (ACCMP).

Aktionsart marking is often a part of a language's derivational morphology. For example, Russian marks semelfactive verbs with the suffix *-nu-* as in *blesnut'* "to flash" (Comrie 1976a:43), which can be opposed to *blestet'* "glitter, sparkle, gleam." Aktionsart can also affect a language's overt morphology by restricting the range of inflectional possibilities. For example, stative verbs in English, such as *know*, are often unable to take the progressive form, as attested by the infelicity of the sentence, **John is knowing Russian.*

The aktionsart categories discussed here can be cast in featural terms as in Table 1.

Feature	*Label*
Stative	STAT
Dynamic	DYN
Telic	TEL
Atelic	ATEL
Punctual	PCT
Durative	DUR
Achievement	ACH
Accomplishment	ACCMP
Semelfactive	SEMEL
Activity	ACTY

Table 1: Aktionsart features based on Vendler (1957), Comrie (1976a), and Cable (2008)

5.2 Animacy

Yamamoto (1999:1) writes that animacy "can be regarded as some kind of assumed cognitive scale

[4]Scientifically, of course, any action's duration can be measured, but lexically, punctual events are treated as durationless.

extending from human through animal to inanimate." Referencing Comrie (1989), Yamamoto notes that animacy "reflects a natural human interaction amongst several different parameters," including the notions of (verbal) person, individuation, and agency. Animacy as a general organizing principle of language is typically modeled as a scale or hierarchy, such as that from Foley and Valin (1985:288; cited in Yamamoto 1999:27):

(2) speaker/addressee [1st & 2nd persons; JCS] > 3rd person pronouns > human proper nouns > human common nouns > other animate nouns > inanimate nouns

To the extent that animacy is a conceptually separate property from person, individuation, and agency, it encompasses only three principal categories: Human, animate, and inanimate (Comrie 1989:185). Animacy is the only dimension of meaning that appears not to have dedicated morphology among the world's languages, i.e. there appears to be no language with a morpheme that specifically and only means either 'animate' or 'inanimate.' However, animacy so strongly conditions the distribution of overt morphology and influences the structure of paradigms, especially with regard to case, that it is necessary to include in order to correctly analyze and generate overt morphology.

The best known example of animacy influencing morphological case comes from Russian. In Russian, "living things who/which both breathe and move" are considered to be animate (Yamamoto 1999:48 citing Corbett 1981:59). Inanimate masculine nouns have identical nominative and accusative forms which are opposed to a distinct genitive form (among other cases), but animate masculine nouns have identical genitive and accusative forms, opposed to a distinct nominative (ibid.).

(3)

	Inanimate		Animate	
Case	*city*	*table*	*pupil*	*rabbit*
nom	górod	stol	učeník	królik
gen	góroda	stolá	učeniká	królika
acc	górod	stol	učeniká	królika

It is possible for languages to make more nuanced distinctions in the use of a distinct accusative case. Comrie (1989:189) writes that "[...] we find languages that have separate accusatives only for first and second perosn pronouns (e.g. Dyirbal), only for pronouns and proper names and kin terms (e.g. Gumbainggir), only for human noun phrases (e.g. Arabana), [and] only for animate noun phrases (e.g. Thargari) [...]" (quoted in Yamamoto 1999:46). This shows the necessity to separate humanness from general animacy.

While Russian treats animals (in the colloquial sense, i.e. not including humans or microorganisms) as animate, other languages distinguish between humans and non-humans. For example, the Ryukyuan language Yuwan has two allomorphs for the nominative case, *-ga* and *-nu* (Niinaga 2010:58-59). The *-ga* allomorph is used for human pronouns, demonstratives, and elder kinship terms, while the *-nu* allomorph is used for everything else. While the source does not discuss the status of other terms for humans that fall outside the realm of pronouns and kinship (such as nouns for indicating occupations, i.e. 'worker'), the examples in (4) from Niinaga (2010:59) make clear that animals, even the most domesticated (dogs), are treated as grammatically distinct from human pronouns. Yuwan can therefore be interpreted as making a similar animacy distinction to Gumbainggir and possibly also to Arabana.

(4) a. wan=ga aik-ju-i
1.sg=nom walk-ipfv-nonpast
'I will walk.'

b. ɨn=nu aik-ju-i
dog=NOM walk-IPFV-NONPAST
'The dog will walk.'

Apart from affecting the distribution of overt case morphology, animacy affects other areas of grammar, including word order (e.g. in Navajo; Yamamoto 1999:53-54).

Four features are necessary to mark the cross-linguistically observed animacy distinctions.

Feature	*Label*
Animate	ANIM
Inanimate	INAN
Human	HUM
Non-human	NHUM

Table 2: Features necessary for encoding typologically-attested animacy distinctions

5.3 Argument Marking

Nichols (1986) discusses a fundamental distinction between head-marking languages, which mark relations between a head and dependent on the head, and dependent-marking languages, which mark such relations on the dependent. Familiar European languages are dependent-marking in that nominal case is marked on the noun itself. However, languages such as Abkhaz (Chirikba 2003) and Choctaw (Davies 1986) are head-marking, and can mark four or five distinct nominal arguments, respectively, on a verb. For example, Abkhaz can mark a subject (ergative), object (absolutive), indirect dative object, and indirect beneficiary object simultaneously on a single verb, as shown in (5).

(5) á-salam ∅- sə- z- lɔ́- šʷ- t
greeting it.3.SG.ABS 1.SG BEN 3.SG.FEM.DAT 2.PL.ERG give
'Give her my greetings!' (Chirikba 2003:39), lit. You all give greetings to her for me!

Following Kibrik (2012), the arguments marked by a verb are labeled here using case terminology, rather than, for example, grammatical terminology such as "subject, direct object, indirect object." This is especially appropriate given that head-marking languages can also show patterns similar to so-called 'quirky case' in which verbs, especially psych verbs, trigger agreement patterns in which the subject is not marked in the same way as subjects for other verbs. This shows that purely formal considerations also shape the pattern, rather than just grammatical or even semantic relations, even though these also play a role. A clear example of this is the fact that in Choctaw, the verb for 'forget' may use a quirky case pattern with a subject in the dative case in addition to a pattern in which the subject is in nominative case without any difference in meaning (Davies 1986: ch. 1, (9)).

(6) a. chim- ihaksi -li -tok
2.SG.DAT forget 1.SG.NOM PST
'I forgot you' (less 'quirky' pattern with subject in nominative case)

b. chi- am- ihaksi -tok
2.SG.ACC 1.SG.DAT forget PST
'I forgot you' (fully 'quirky' pattern, lit. to me, you [were] forgot[ten])

To capture arguments marked on the verb in head-marking languages, the UniMorph Schema employs templatic features which all begin with ARG-, to signify that a verb is marking an argument, and continue with shorthand features for case, person, number, and gender. Templatic features are also used to mark possessive morphology, as discussed in §5.21 (p. 46). Case marks the type of relation that the argument bears to the head, and can take the values NO, AC, AB, ER, DA, BE to mark the nominative, accusative, absolutive, ergative, dative, or benefactive cases, respectively. Person takes the usual values 1, 2, and 3, and number takes the values S, P for singular and plural. Gender takes the values M, F, NH for masculine, feminine, and non-human, respectively.[5] Additional values may be added to the template later as needed, for example, dual number. The following template captures how features for argument marking on the head in head-marking languages can be constructed for use in the schema.

(7) Template for Argument Marking Features

$$\text{ARG} \left\{ \begin{array}{l} \text{NO} \\ \text{AC} \\ \text{AB} \\ \text{ER} \\ \text{DA} \\ \text{BE} \end{array} \right\} \left\{ \begin{array}{l} 1 \\ 2 \\ 3 \end{array} \right\} \left\{ \begin{array}{l} \text{S} \\ \text{P} \end{array} \right\}$$

e.g. ARGNO1S, ARGNO1P, ARGNO2S, etc.

5.4 Aspect

Aspect is defined according to the framework proposed by (Klein 1994, 1995), which builds on Reichenbach (1947) and relates the concepts of Time of Utterance (TU, '|'), Topic Time (TT, '[ ]'), and Situation Time (TSit, '{ }') to define tense and aspect categories. Topic Time (TT) and Situation Time (TSit) are conceived as spans while Time of Utterance (TU) is a single point. By defining tense and aspect categories solely in terms of the ordering of these spans and TU, tense and aspect categories can be defined independent of the language under analysis in a way that facilitates cross-linguistic comparison. TU (symbolized with '|') is the time at which a speaker makes an utterance, and topic time (TT, symbolized by brackets '[ ]') is the time about which the speaker is making a claim about the action of the verb. Situation time (TSit, symbolized with braces '{ }') is the time in which the state of affairs described by the verb held true.

Aspect indicates the relationship between the time for which a claim is made (TT) and the time for which a situation actually held true (TSit). For example, in the sentence, *by lunch, Mary had drank the orange juice*, there was a time (TSit) in which Mary was drinking orange juice, and this time had come to an end before the time that the claim was made with reference to (TT, which here is lunchtime). Moreover, the topic time was before the speaker made the utterance (TU), which is the reason for the past tense form *had*. The relation of TT, TSit, and TU can be symbolized as in the diagram in (8), in which TSit is symbolized by { }, TT by [ ], and TU by |. The span symbolized with a flat line is the source state (SS) in which the orange juice has not been drunk, and the span symbolized with plus signs (+) is the target state (TS) in which the orange juice has been drunk.

(8) By lunch, Mary had drank the orange juice.
————————{——++}++[++++]++++|++

[5] 'Non-human' forms part of the gender/noun class system in Abkhaz (Chirikba 2003:39), but is better thought of as an animacy feature (§5.2).

This is an example of perfect aspect since the action (TSit, { }) occurred prior to the time about which a claim had been made (TT, [ ]).

The core aspects that can be defined by relating TSit and TT are: imperfective, perfective, perfect, progressive, and prospective. The habitual and iterative aspects are also defined this way, but require more than one TSit.

With the imperfective aspect (IPFV), TT is included fully within TSit (Klein 1994:102), as shown in the following diagram. Because time of utterance (TU) is relevant only to tense, not aspect, it is fixed in the diagrams at a point toward the end of the target state, and hence the past tense is used in all examples. The realization of the imperfective aspect overlaps with that of the progressive aspect in English due to the fact that English lacks a distinct imperfective aspectual form.

(9) Imperfective aspect
——————{—[—++]++}+++++++++|++
∼'She was leaving' (translation can only be approximate)

The perfective aspect (PFV) indicates that the TT is only partially included within TSit (108), and with 2-state verbs, it partially overlaps with TSit on its right boundary and is within the target state (TS). This leads to an interpretation of completion of the action. The following diagram illustrates the perfective aspect for a 2-state verb.

(10) Perfective aspect: Partial TT overlap with TSit within the target state
——————{——[++++}++]++++++|++
'She left.'

The perfect aspect (PRF), which is distinct from the perfective aspect, "[...] locates the TT in the post-state [TS] of the corresponding situation" (Klein 1994:9). In the perfect aspect, the TT is to the right of the TSit, and may be distantly so (Klein 1994:110), leading to examples as in (11).

(11) Perfect aspect: TT is located at a distance from TSit, yet still within its target state
——————{——+++}++++++[++]+|++
'She had left.'

Klein (1994:9) writes that "[w]ith the progressive form, the TT is properly contained in the first state of the situation (which is the only one for 1-state situations and which has no TT-contrast for 0-state situations)." This can be regarded as a subcase of the imperfective aspect.

(12) Progressive aspect (PROG): TT is located exclusively within the source state of TSit
————————{—[——]++++++}++++++|++
'She was leaving.'

The prospective aspect (PROSP) is used when the topic time (TT) precedes the situation time (TSit; 108). The prospective aspect is often confused with future tense since it refers to an action in the future. However, separating prospective aspect and future tense leads to a natural explanation for so-called 'future perfect' (which may actually be FUT;PRF) and so-called 'future-in-the-past' temporal categories. While the *going to / gonna* and *be about to* constructions are useful translations for the prospective aspect, they contain additional, separate connotations of intention for the *going to* construction and immediacy in the *be about to* construction.

(13) Prospective aspect: TT is located before TSit
——[——]—{———++++++}++++++|++
'She was going to leave.'

The iterative aspect describes the occurrence of multiple events within a single time frame. This can be represented as multiple TSit spans within a given TT. This TT and the number of TSit spans must be bounded, since if it is infinite or unbounded, it gives rise to a habitual interpretation.

(14) Iterative aspect: Multiple instances of the same TSit occur fully within a bounded TT
.....[.....{—+++}$_{x_1}$.....{—+++}$_{x_2}$.....{—+++}$_{x_n}$.....].....|.....
'He kept glancing out the window.'

The habitual aspect is essentially the iterative aspect, but with infinitely many multiple situation times within an unbounded topic time.

(15) Habitual aspect: Infinite instances of the same TSit occur fully within an unbounded TT
.....[.....{—+++}$_{x_n}$.....{—+++}$_{x_{n+1}}$.....|.....{—+++}$_{x_{n+\infty}}$.....].....
'He leaves every morning at 8.'

The features outlined for aspect are as in (3)

Feature	*Label*
Imperfective	IPFV
Perfective	PFV
Perfect	PRF
Progressive	PROG
Prospective	PROSP
Iterative	ITER
Habitual	HAB

Table 3: Aspectual features, following Klein (1994, 1995)

5.5 Case

Nominal case marks the grammatical relationship that nouns have to the verb, among other functions. More technically, "case is a system of marking dependent nouns for the type of relationship they bear to their heads" (Blake 2001:1). This encompasses two uses of the term 'case': 1) the marking of argument structure by the syntax at a deep level (such as logical form, LF), and 2) the overt morphological marking of argument structure, spatial relations, and psychological attitudes at the surface level. For our purposes, only this second type of 'case' is relevant.[6] The types of overt case that are encountered in the world's languages can be divided into three types: 1) core case, 2) non-core non-local cases, and 3) local case (following the classification of Blake 2001).

Because the number of case features is so large and have distinct uses and internal logics, we split discussion of nominal case into its three subtypes. First, core cases that indicate the roles of NPs with respect to syntactic alignment will be discussed, followed by other non-core, non-local cases. Finally, local cases, which have been shown to have an internal organization that is consistent across languages (Radkevich 2010) will be discussed.

5.5.1 Core Case

Core case is also known as 'non-local,' 'nuclear,' or 'grammatical' case (Blake 2001:119; Comrie and Polinsky 1998:97) and includes a limited set of cases that are used to indicate the role of a syntactic argument as subject, object, or indirect object. The types of core cases vary according

[6]The other type is sometimes referred to as 'deep case' or 'syntactic case.'

to the syntactic alignment that a given language uses and can be defined in terms of three "meta-arguments," S, A, and P.[7] S is the subject of an intransitive (1-argument) verb. A is the subject of a transitive (2-argument) or ditransitive (3-argument) verb. P is the direct object of a transitive or ditransitive verb.[8]

The ways in which languages map morphological cases to the meta-arguments S, A, and P are collectively called syntactic alignment systems. In a language with nominative-accusative alignment, the nominative case is used for nouns that function as S or A and accusative is used for nouns that function as P (Blake 2001:119-121). In an ergative-absolutive language, the ergative case is used for nouns that function as A only while absolutive case is used to mark nouns that function as S or P (122-125). Finally, a few languages have been claimed to possess a full, tripartite distinction between S, A, and P. These languages include Wangkumara (Breen 1976 via Blake 2001:126; Pama-Nyungan) and Yazgulyam (Èdel'man 1966:37, 167, 185, Payne 1981:176 via Bickel and Nichols 2009; Pamir, Indo-Iranian).[9,10] The core cases associated with syntactic alignment systems are summarized in Table 4.

Case Name	*Feature*	*Meta-Arguments*	*Alignment System*
Nominative	NOM	S, A	Nominative-Accusative
Accusative	ACC	P	Nominative-Accusative
Ergative	ERG	A	Ergative-Absolutive
Absolutive	ABS	S, P	Ergative-Absolutive
Nominative, S-only	NOMS	S (only)	Tripartite

Table 4: Cases associated with core syntactic argument marking

5.5.2 Non-Core, Non-Local Case

The other non-local cases include both cross-linguistically common cases, such as the dative, genitive, instrumental, comitative, and vocative, which are sometimes considered 'core,' as well as rarer cases, such as the benefactive, purposive, equative, and privative.

The dative (DAT) consistently marks indirect objects, and is often also used to mark experiencers, beneficiaries, and purposes (as well as other much less common functions; Blake 2001:144-145). The beneficiary (BEN) and purposive (PRP) functions of the dative case are split into distinct cases in a few languages such as Basque (Blake 2001:145).

Another common non-local case is the genitive (GEN), which prototypically marks the possessor (e.g. *John's*/*John* in *John's cat*, in *the cat of John*, and in *that cat of John's*). Related to the genitive is the relative case (REL), which combines the "*A* function and possessor function" and occurs in "a number of Caucasian languages and the Eskimo languages" (151).[11] The genitive case is also used in some languages to cover the uses of the partitive case (PRT), which marks the patient (P) of a verb as being only partly affected by the verb's action (e.g. *some milk* [but not all]

[7] S, A, and P are all abbreviations from the first letters of the terms 'subject,' 'agent,' and 'patient,' respectively.

[8] Indirect objects are commonly marked with the dative case and are typically not considered core arguments.

[9] Another kind of syntactic alignment, which cannot be described solely in terms of S, A, and P, is the direct-inverse system, which is described under the section on grammatical voice systems in §5.25, p. 56. These kinds of systems, which occur for example in Plains Cree, can overtly mark direct and inverse "case" on a noun.

[10] Nouns are also sometimes marked as 'antipassive' (ANTIP). This term and concept are discussed in the section on voice §5.25.

[11] Note that this case is not strictly basic, and one could resolve it in context to either ERG or GEN.

in *he drank some milk*). The partitive case occurs as an independent case in Estonian, Finnish, and Hungarian, where it "is used for the patient if it represents part of a whole or an indefinite quantity, [...] if the action is incomplete, or if the polarity of the clause is negative" (153). These uses of the partitive are captured by the genitive case in both Russian (ibid.) and French. Russian's "genitive of negation" (e.g. Russian *njet molok-á* 'no milk-GEN') can also be thought of as connected to the senses of meanings conveyed using the partitive (ibid.).

Another cross-linguistically common non-local case is the instrumental case (INS), which marks a noun as "the instrument with which an action is carried out" (Blake 2001:156). The comitative (or sociative) case (COM) expresses accompaniment ("with"; ibid.) and its function is often subsumed within the instrumental case (e.g. Russian *s Ivan-om* 'with John-INS'). The vocative case (VOC) indicates that a noun is being used as a direct form of address. For example, "master!" is rendered with the vocative in Latin as *domin-e* 'master-VOC' (4-5). The comparative case (COMPV) marks the standard of comparison, i.e. "than X," and occurs in "Dravidian and some Northeast Caucasian languages" (156). Another case that occurs in Tsez and other languages is the equative (EQTV), which indicates the standard of comparison in statements of equality and can be translated "(e.g. as much) as X" (Comrie and Polinsky 1998:101-102). The equative case is used in this form in Ancash Quechua, where it is expressed with the suffix *-naw* (Cuzzolin and Lehmann 2004). Equative case is also used in some languages to capture a similative meaning "like, resembling," and will be used as such here. Even rarer cases, which occur primarily in Australian languages, include the privative case (PRIV; called the abessive case in Uralic languages), which indicates "lacking, not having, without," and a positive counterpart, the proprietive (PROPR), which indicates "having" (Blake 2001:156). Australian languages also commonly contain an aversive case (AVR), which "indicates what is to be feared or avoided" (ibid.). Finally, the formal case (FRML), meaning "in the capacity of, as" occurs as the 'essive-formal' case in Hungarian (Spencer 2008:39). Hungarian also explicitly marks an entity as being the result of a transformation. The 'translative' case is used, for example, in contexts like 'turn into X' where X would be marked with the translative case (TRANS). The Hungarian case system also contributes the essive-modal case (BYWAY), which marks the notion of 'by way of' a location. The non-core, non-local cases are summarized in Table 5.

Case Name	*Feature*	*Definition*	*Gloss*
Dative	DAT	marks indirect object	to (indirect object)
Benefactive	BEN	marks a beneficiary of an action	(a gift, e.g.) to, for (s.o.)
Purposive	PRP	marks purpose of or reason for an action	for (profit, e.g.)[12]
Genitive	GEN	marks possessor	of s.o., s.o.'s
Relative	REL	marks possessor and A role	of s.o., s.o.'s
Partitive	PRT	marks a patient as partially affected	some of
Instrumental	INS	marks means by which an action occurred	by (means of) sth, with sth, using sth
Comitative	COM	marks accompaniment	(together) with
Vocative	VOC	indicates direct form of address	"s.o.!"
Comparative	COMPV	marks standard of comparison	than sth, s.o.[13]

[12]Note that this use of purposive is distinctive from general purposive modality (PURP) and the superordinate modality category purposive used in some Australian languages (AUPRP).

[13]Note that this use of comparative is to mark the standard of adjectival comparison, not the comparative degree of the adjective, which is expressed with the feature CMPR.

Equative	EQTV	marks equality or similarity	(as much) as s.o./sth, like s.o./sth[14]
Privative	PRIV	indicates lack of something	without, lacking sth
Proprietive	PROPR	indicates quality of possessing something	having sth
Aversive	AVR	indicates what is to be feared, avoided	(afraid) of (ghosts, e.g.), (dying) from (poison, e.g.)
Formal	FRML	indicates that sth is function as sth else	as sth, in the capacity of sth
Translative	TRANS	indicates that an entity is the result of a transformation	
Essive-modal	BYWAY	indicates that a motion event occurs 'by way of' a location	

Table 5: Non-core, non-local cases

5.5.3 Local Case

Local cases express spatial relationships typically expressed by prepositions in English (and by adpositions in general in the majority of the world's languages; Radkevich 2010:24). For example, Tabassaran (Nakh-Daghestanian) has 8 local case marking morphemes that correspond to the meanings "in (hollow space), on (horizontal), behind, under, at, near or in front of, among," and "on (vertical)," which can all appear with either essive (stationary location), allative (motion toward), or ablative (motion from) case markers (Comrie and Polinsky 1998:98).

The features for encoding local cases are compositional and reflect the structure and distinctions found in the complex local case systems of the world. Based on a survey of 111 languages with local case systems, Radkevich (2010:20-107) shows that local case morphology can be divided into four types of local case morphemes that are consistently arranged following the schema in (16).

(16) Universal Template for Arrangement of Local Cases (Radkevich 2010:5)
Noun.Lemma-Stem.Extender-Place-Distal-Motion-Aspect

In (16), the stem extender may itself be a case form such as ergative in Nakh-Daghestanian languages or genitive in Estonian (Radkevich 2010:3, 21). The types of local case morphemes include Place, Distal, Motion, and Aspect morphemes. The following local case morpheme meanings can be organized within each category (Place, Distal, etc.) through the use of abstract features (as in phonology) that are more general than the feature labels that will be specified here.

The place morphemes "roughly correspond to what [are] usually called adpositions in languages without local cases" and indicate orientation to a very precise degree (29). Consequently, of all the types of local case morphemes, Place morphemes are the most numerous. The Nakh-Daghestanian languages Tabassaran and Tsez appear to be the languages that contain the largest number of Place morphemes, which include separate morphemes for "among, at, behind, in, near, near/in front of, next to, on, on (horizontal), on (vertical)," and "under" (ibid.; Comrie and Polinsky 1998).

[14]Note that the equative case (EQTV) marks the standard of comparison in a comparative construction expressing equality while the equative marking (EQT) of an adjective indicates that the degree of the adjective itself expresses equality.

The Distal category can be seen as an elaboration of the Place category of morphemes. Only three languages in the survey of 111 languages have a distal morpheme within their local case system: Tsez (Nakh-Daghestanian), Savosavo (isolate; Solomon Islands), and Central Dizin (Omotic) (Radkevich 2010:33). Central Dizin overtly marks both distal and its opposing counterpart, proximate, but Tsez overtly marks only distal.

The motion category is composed of only three possible parameters, namely essive (ESS), allative (ALL), and ablative (ABL; 34-36). Essive indicates static location with no motion (52). Ablative indicates motion *away from* a source (ibid.). Allative indicates motion *toward* a source (ibid.).

The aspect category within case can be seen as an elaboration of the motion category. It includes four parameters, namely approximative, terminative, prolative/translative, and versative (37, 53-55). "Approximative case denotes a movement that is directed towards something but does not reach its goal (i.e. incompletive aspect). Grammatical descriptions ... point out that this case is used to emphasize that movement does not reach its goal." In contrast, terminative indicates that a goal has just been reached, and has a basic meaning of "as far as, up to." The versative morpheme indicates motion in the direction of a goal and has a meaning similar to "towards, in the direction of." Its counterpart is the prolative/translative case, which indicates motion "along" or "across" a referent point.

All the local case morphemes are listed in Table 6, which lists each case morpheme with its name (a gloss), its feature label, and its subcategory (place, distal, etc.). Note that the Distal and Proximate morphemes, REM and PROXM here, are identical to the Remote and Proximate morphemes specified for making distinctions among demonstratives in §5.8. This is not accidental. Radkevich (2010:40-42) notes that 'orientation morphemes' with meanings much like those of demonstrative pronouns are directly affixed to nouns bearing local case morphology in Tabassaran, suggesting that in some languages, these domains of morphology may overlap. Moreover, the semantics of the distal/remote and proximate morphemes are the same for local cases and for demonstratives. Distal local case and remoteness on pronouns both indicate that the thing involved is distant from the speaker while proximate morphemes in both categories indicate the opposite.

Feature	*Label*	*Case Subcategory*
Among	INTER	Pl
At	AT	Pl
Behind	POST	Pl
In	IN	Pl
Near	CIRC	Pl
Near, in front of	ANTE	Pl
Next to	APUD	Pl
On	ON	Pl
On (horizontal)	ONHR	Pl
On (vertical)	ONVR	Pl
Under	SUB	Pl
Distal	REM	Dst
Proximate	PROXM	Dst
Essive	ESS	Mot
Allative	ALL	Mot
Ablative	ABL	Mot
Approximative	APPRX	Asp
Terminative	TERM	Asp

Prolative/translative	PROL	Asp
Versative	VERS	Asp

Table 6: Local case morphemes

5.6 Comparison

Cuzzolin and Lehmann (2004) write that "[a]ll the languages of the world have at their disposal different means to express comparison and gradation," and these notions may be expressed through overt affixal morphology. Comparative constructions have "the semantic function of assigning a graded (i.e. non-identical) position on a predicative scale to two (possibly complex) objects" (Stassen 1984:145). Both comparison and gradation can be expressed via overt affixal morphology.

Languages typically assign two degrees of comparison. The comparative, such as English *-er*, Russian *-ee*, or the Georgian circumfix *u-...-es* (Cuzzolin and Lehmann 2004), relates two objects such that one exceeds the other in degree of exhibiting some quality.

The superlative relates any number of objects such that one exceeds all the others. This is specifically the relative superlative, such as that expressed by English *-est*. Another type of superlative, the absolute superlative, expresses a meaning like "very" or "to a great extent," and is used in Latin (among other languages) as in the first example in (17) from Cuzzolin and Lehmann (2004). The second example shows the Latin superlative used as a relative superlative.

(17) a. Absolute Superlative
vir felicissimus 'a very lucky man'

b. Relative Superlative
vir omnium felicissimus 'the luckiest man of all'

Although Stassen (1984) defines comparative constructions specifically as comparing entities that exhibit a quality to an unequal extent, equative constructions can also be viewed as comparative constructions in which two entities are compared, but in which they exhibit a quality to an equal extent. The standard of comparison (as much *as X*) can be marked with a special equative case morpheme (EQTV) just as the adjective itself can be marked as conveying equality (EQT), as in Estonian in (18) and in Indonesian in (19), both from Cuzzolin and Lehmann (2004).

(18) Minu õde on minu pikk-une
I.GEN sister is I.GEN tall-EQT
'My sister is as tall as me.'

(19) Ayah saya se-tinggi paman saya
father 1.SG EQT-tall uncle 1.SG
'My father is as tall as my uncle.'

The features that are necessary to encode the overt affixal morphology involved in comparative constructions is presented in Table 7. Note that absolute and relative superlatives can be distinguished by combining the superlative feature with the features for absolute or relative.

Feature	*Label*
Comparative	CMPR[15]

[15]Note that this is distinct from the comparative case, which marks the standard of comparison, e.g. *than X*. In principle, this and the equative feature for adjectives and for nouns could be collapsed to only one feature and used for both these parts of speech.

Superlative	SPRL
Absolute (for superlatives)	AB
Relative (for superlatives)	RL
Equative	EQT[16]

Table 7: Features necessary for encoding adjectival comparison morphology

5.7 Definiteness

Lyons (1999:278) defines definiteness as "the grammaticalization of identifiability," noting that while "it is to be expected that there will be other uses of definiteness which do not relate to identifiability, [...] there is always a large central core of uses relatable directly to identifiability" and this "justifies us in identifying the same category cross-linguistically." Identifiability is a pragmatic concept by which the hearer is able to be directed to uniquely identify some entity in the discourse (Lyons 1999:5-6).

Definiteness as a grammatical category can be divided into three distinct levels: Definite, specific indefinite, and non-specific indefinite. Lakhota (Siouan) contrasts these three categories, as in (20), from Lyons (1999:50,99).

(20) a. c'ą kį 'the stick' (definite)

b. c'ą wą 'a stick' (specific indefinite)
c'ą wą 'ag.li'
'He brought a stick.' (a certain stick)

c. c'ą wąži 'a stick' (non-specific indefinite)
c'ą wąži 'aų wo
'Put a stick on [the fire].' (any stick, but still only one, not many)

Note that English does not distinguish between specific and non-specific indefinites.

For the purposes of the UniMorph Schema, it is sufficient to distinguish the categories distinguished in Lakhota, which appear to be the most elaborate definite distinctions made using overt, affixal morphology. Because many languages do not distinguish beyond the level of definite and indefinite, features for these two categories are established, with two additional features for specific and non-specific. These can be combined with the indefinite feature in the same way that the features for inclusive and exclusive can be combined with first person.

Feature	*Label*
Definite	DEF
Indefinite	INDF
Specific	SPEC
Non-Specific	NSPEC

Table 8: Features necessary for encoding definiteness

[16]Note that this is distinct from the equative case, which marks the standard of comparison, e.g. as much *as X*.

5.8 Deixis

The referent of pronouns is dependent on the context of the utterance, as is the referent of words like *here, this, that, these, those.* This quality of context-dependency is called deixis, and words are sometimes systematically contrasted in terms of their deictic properties.

Demonstrative pronouns are deictic words which encode a variety of distinctions in meaning that specify the relationship between the speaker or addressee and the person or thing referred to by the pronoun, and these distinctions are relevant for determining the referent of the demonstrative. For example, the difference between *this* and *that* in English can be understood as a distinction in distance.[17] *This* is used to describe something closer to the speaker and *that* is used for something farther away. In languages that use words such as *this* and *that* as personal pronouns (e.g. *this (man)* for *he*, *that (man)* as an alternative for *he* for a person farther away, etc.), distance is therefore a deictic property. For the same reasons, reference point, visibility, and verticality are also relevant deictic properties.[18]

5.8.1 Distance

As mentioned, distance is a relevant feature in distinguishing third-person pronouns, especially in *two-person* languages in which the third-person pronouns are not fully distinct forms and are usually derived from, or identical to, demonstrative pronouns. The maximal distinction appears to be a three-way contrast between proximate, medial, and remote, roughly translating to English *this*, *that*, and *that (over there, yonder)*.[19] Basque is a two-person language that uses demonstrative pronouns for third person pronouns, and these demonstratives make a three-way distinction in distance (Hualde and Ortiz de Urbina 2003:123, 150).

(21) Singular pronouns in Basque (Hualde and Ortiz de Urbina 2003:123, 150, Bhat 2004:141)

ni	1.SG	'I'
hi	2.SG	'you'
hau	3.SG.PROXIMATE	'he, she (nearby)'
hori	3.SG.MEDIAL	'he, she (not close)'
hura	3.SG.REMOTE	'he, she (over there, yonder)'

5.8.2 Reference Point

Another deictic property is the reference point for determining the relationship of the speaker, addressee, and referent of the pronoun. This dimension often overlaps with distance distinctions, but is sometimes explicitly separated. The three primary features for the reference point feature are speaker as reference point (REF1), addressee as reference point (REF2), and a reference point not related to either speaker or addressee (distal; DIST). Another essential feature is what has been called the "anaphoric" feature, which designates a pronoun as obligatorily referring to a referent that occurs in the discourse. Strictly speaking, anaphoric pronouns must refer to something that has

[17] "Distance" here should be understood in a very broad sense, since the English demonstratives *this* and *that* can be used to express spatial, temporal, or even psychological/social distance.

[18] Other deictic properties include temporal deixis and psychological/social deixis. At present, it is unclear the extent to which these are realized by morphological alternations as opposed to different lexemes. The same can be said of deixis as a whole. Thus, this dimension is included, similar to aksionsart, to allow the encoding of regular deictic relationships and with the hope that these distinctions will be useful for characterizing regular distinctions in meaning among pronouns, even when that meaning is derivationally or lexically encoded.

[19] The term 'remote' is used here for the greatest extent of distance while the term 'distal' is reserved for a specific reference point distinction, described below.

previously occurred in discourse while a cataphoric pronoun would obligatorily refer to something that comes after it in the discourse. Because the pronouns of Lak (Nakh-Daghestanian; Bhat 2004:132-133, Friedman 2006:304) and Sinhalese (Indo-European; Gair 2003:782-783) appear not to contrast specific anaphoric and cataphoric pronouns, the feature that is used to indicate obligatory anaphoric or cataphoric reference is simply called phoric (PHOR). The pronoun system of Sinhalese (Sinhala) exemplifies all these distinctions (and makes additional distinctions not presented in this example).

(22) Reference point distinction in Sinhalese human, non-feminine pronouns (Gair 2003:782-783)

meyaa	PROX.REF1	'he (near speaker)'
oyaa	PROX.REF2	'he (near addressee)'
areya	DIST	'he (not near either speaker or addressee)'
eyaa	PHOR	'he (previously mentioned or to be disambiguated; phoric)'

The Hausa pronominal paradigm illustrates the need to distinguish distal, as a reference point feature, from remote, a distance feature.[20]

(23) The masculine pronouns of Hausa (Bhat 2004:145)

wannàn	PROX.REF1	'he (near speaker)'
wànnan	PROX.REF2	'he (near addressee)'
wancàn	REMT.NOREF	'he (not near either speaker or addressee; distal)'
wàncan	REMT	'he (far away; remote)'

5.8.3 Visibility

In addition to distance and reference point, third-person pronouns and demonstratives can also be distinguished on the basis of whether their referent is visible (VIS) or not (NVIS), as in K^{w}ak'^{w}ala (Wakashan; Beck 2000:193), Yupik (Eskimo-Aleut; Bliss and Ritter 2001), and Khasi (Austroasiatic; Bhat 2004:133). The masculine singular pronouns of Khasi illustrate distance, visibility, and (as will be explained further in the next section) verticality.

(24) The masculine singular pronouns of Khasi (Austroasiatic; Bhat 2004:133)

u-ne	PROX	'he (near)'
u-to	MED	'he (not near, not far)'
u-tay	REMT.VIS	'he (far away, visible)'
u-to	REMT.NVIS	'he (far away, not visible)'
u-tey	ABV	'he (above)'
u-thie	BEL	'he (below)'

5.8.4 Verticality

As previewed in the data from Khasi in (24), third person and demonstrative pronouns can also encode the vertical spatial relationship between the speaker and referent. Extending on a level plane from the speaker's view, a referent can either be above that plane (ABV), below it (BEL), or at the same level (EVEN) (Schulze 2003:297-300).

In addition to Khasi, many of the Nakh-Daghestanian (Northeast Caucasian) languages make verticality distinctions in their demonstrative pronouns. Agul (Kurag dialect) exhibits the typical paradigmatic structure for demonstrative pronouns in Nakh-Daghestanian (Schulze 2003:300).

[20] Moreover, the Hausa 'definite' article obligatorily refers to an entity previously mentioned in the discourse, and may therefore be better described as (ana)phoric (PHOR) rather than definite (DEF).

(25) Agul demonstrative pronouns (Schulze 2003:300, data from Magometov 1970:109-112)

me	PROX	'this'
te	REMT	'that'
le	REMT.ABV	'that (above)'
ge	REMT.BEL	'that (below)'

However, not all Nakh-Daghestanian languages possess this distinction, and in some, the original system has been subjected to restructuring. For example, in Lak, the former remote, "below" demonstrative pronoun *ga* has become an unmarked remote while the former unmarked demonstrative pronoun *ta* has come to designate something at the same level (Friedman 2006:304).

(26) Lak demonstrative pronouns (Friedman 2006:304)

		Current Usage	*Older Usage*
wa	PROX.REF1	'this (near speaker)'	
mu	PROX.REF2, PHOR	'that (near addressee; old information)'	
ta	REMT.EVEN	'that (opposite, level)'	'that (unmarked)'
ga	REMT.DIST	'that (unmarked)'	'that (below speaker)'
k'a	REMT.REF1.ABV	'that (above speaker)'	

5.8.5 Summary

The discussion above has highlighted the cross-linguistically most common features used to capture deictic distinctions that are encoded by morphological oppositions. These features are presented in Table 9.

Feature	*Label*
Proximate	PROX
Medial	MED
Remote	REMT
First Person Reference Point	REF1
Second Person Reference Point	REF2
No Reference Point, Distal	NOREF
Phoric, situated in discourse	PHOR
Visible	VIS
Invisible	NVIS
Above	ABV
Even	EVEN
Below	BEL

Table 9: Features relevant to third-person and demonstrative pronouns

5.9 Evidentiality

Evidentiality is the morphological marking of a speaker's source of information (Aikhenvald 2004). As is mentioned in §5.15, evidentiality is sometimes viewed as a category of mood and modality. The UniMorph Schema follows Aikhenvald (2004) in viewing evidentiality as a separate category from modality. Although categories of evidentiality may entail certain modal categories (such as a

hearsay or reported information evidential entailing irrealis or subjunctive mood), evidentiality is a distinct morphological category that encodes only the source of the information that a speaker is conveying in a proposition.

Aikhenvald (2004:26-60) provides a survey of evidential systems across the world's languages, and establishes the typology in Table 10.

	2-Choice Systems
A1	Firsthand (via any sense) vs. Non-firsthand: Cherokee, Yukaghir, Jarawara (26-28)
A2	Non-Firsthand vs. 'everything else': Turkish, Hunzib, Abkhaz, Mingrelian, Svan, Mansi, Khanty, Nenets, Meithei, Hare, Chipewyan, Kato (29-31)
A3	Reported vs. 'everything else': Estonian, Latvian, "some Philippine languages," Paraguayan Guaraní, Livonian, Lezgian, many South American languages, Menomini, Potawatomi, Cupeo, Shoshone, Omaha, Kham, Enga, Tauya, Mparntwe Arrernte (31-34)
A4	Sensory Evidence vs. Reported: Ngiyambaa, Diyari, Latundê/Lakondê (34)
A5	Auditory vs. 'everything else' Yuchi/Euchee (34)
	3-Choice Systems
B1	Direct (or Visual), Inferred, Reported: Qiang, Shasta, all Quechua languages, Aymara (visual, hearsay, reported), Shilluk, Mosetén, Bora, Amdo Tibetan, Ponca, Kapanawa, Maidu, Skidegate Haida, Sanuma, Koreguaje (43-46)
B2	Visual, Non-visual sensory, Inferred: Siona, Washo (46)
B3	Visual, Non-visual sensory, Reported: Oksapmin (Papuan isolate), Maricopa, Dulong (46)
B4	Non-visual sensory, Inferred, Reported: Nganasan, Enets, Retuarã (47)
B5	Reported, Quotative, 'everything else': Comanche, Dakota, Tonkawa, Chemehuevi (50)
	4-Choice Systems
C1	Visual, Non-visual sensory, Inferred, Reported: Ladakhi, Shibacha Lisu, "a number of East Tucanoan languages" (Tucano, Barasano, Tatuyo, Siriano, Macuna), Eastern Pomo, Hupa (51-54)
C2	Direct (or visual), Inferred, Assumed, Reported: Tsafiki, Shipibo-Konibo, Pawnee, Xamatauteri, Maimande (54)
C3	Direct, Inferred, Reported, Quotative: Cora, Northern Embera, SE Tepehuan (57)
	5-or-More-Choice Systems
	Tariana, Tuyuca, Desano, Hup (Hupda), Nambiquara languages of southern Amazonia (Lowe 1999:275-6), Karo (Tupí), Kashaya, traditional Wintu, Central Pomo, Foe (Papuan), Fasu (Papuan) (60)

Table 10: Cross-linguistic typology of evidential systems (Aikhenvald 2004:26-60)

The unique evidential categories, in approximate order of directness of evidence, are: Firsthand, direct, sensory, visual, non-visual sensory, auditory, non-firsthand, quotative, reported, hearsay, inferred, assumed. The degree to which these categories could be reduced using a featural analysis requires further research. Table 11 lists these categories with their suggested labels.

Feature	*Label*
Firsthand	FH
Direct	DRCT
Sensory	SEN
Visual	VISU
Non-visual sensory	NVSEN
Auditory	AUD
Non-firsthand	NFH
Quotative	QUOT
Reported	RPRT
Hearsay	HRSY
Inferred	INFER
Assumed	ASSUM

Table 11: Categories of evidentiality based on they survey and typology provided by Aikhenvald (2004)

5.10 Finiteness

Although Koptjevskaja-Tamm (1993:29) echoes the opinion of Haspelmath (1995:5) and other linguists in writing that "[t]he notion of finiteness is far from being well-defined," it is nevertheless a valuable means of distinguishing a small set of cross-linguistically common verb forms with recurring characteristics. One view is that a verbal form is finite if it takes a subject in nominative, ergative, or absolutive case and if it can take on inflectional features, such as number and gender, based on another syntactic argument (subject or object; Cowper 2002). Another view is that finite verb forms are those which can function as the predicate of independent clauses while nonfinite verb forms are always (morpho-)syntactically dependent (Koptjevskaja-Tamm 1993:29). The notion of finiteness appears to be best viewed as a language-specific constellation of syntactic properties.

Despite the difficulty of defining finiteness in a universal way along semantic or syntactic lines, the notion of finiteness remains cross-linguistically useful and is crucial for differentiating infinitives from other verb forms. For the purpose of this schema, verb forms that take the full extent of tense, aspect, and mood (TAM) marking in a language that can be used as predicates can be considered finite, while verbally-derived parts of speech such as masdars (verbal nouns), participles (verbal adjectives), and converbs (verbal adverbs) are considered nonfinite (Haspelmath 1995:4-7). The finiteness feature in the schema presented here is therefore partially redundant with the part of speech features: Masdars, participles, and converbs are always nonfinite while plain verbs may be finite or nonfinite. An example of a plain verb with the feature nonfinite (NFIN) is the infinitive in languages such as Spanish, Russian, and English.

Feature	*Label*
Finite	FIN
Nonfinite	NFIN

Table 12: Finiteness features

5.11 Gender and Noun Class

Gender is a grammatical category that includes both gender, as conventionally known from European languages like Spanish, French, German, and Russian, as well as systems with more than three categories that are typically described as noun class systems. These include systems like that of the Nakh-Daghestanian languages of the Northeast Caucasus with two to eight gender categories, as well as more extensive systems, such as those of the Bantu languages, which can include up to around twenty distinct categories (Corbett 1991:24, 43-44).

Gender systems can be organized according to either natural gender (biological sex), formal gender, or a mix (Corbett 1991:1). Natural gender systems classify nouns according to transparent semantic criteria. In formal gender systems, however, there need not be any correlation between the gender category of a noun and any of the properties of the entity to which the noun refers. Gender in these systems may either be based on phonological properties or else is an unpredictable lexical property that must be memorized.

Languages need not be organized exclusively by natural or formal gender. For instance, animate nouns are often classified by natural gender while inanimate nouns take gender based on other factors, including phonological characteristics. For example, in Russian, *ʒenʃt͡ʃina* 'woman' is feminine and *muʒt͡ʃina* 'man' is masculine because of the biological sex of their referents. This is apparent due to the phonological criteria that are used to organize most of the lexicon. Because masculine nouns typically end in consonants and feminine nouns typically end in /a/, it is apparent that, phonologically, *muʒt͡ʃina* appears to be feminine, but due to the natural gender of 'man,' it is grammatically masculine. However, phonological factors determine gender with most inanimates such that *kniga* 'book' (ending in /a/) is feminine, *stul* 'chair' (ending in a consonant) is masculine, and *moloko* 'milk' (ending in /o/) is neuter. Although phonological markers help learners determine gender, there is no other organizing principle determining which inanimate nouns are feminine, masculine, or neuter. Finally, there is a residue of nouns that end in the orthographic soft sign (*mjagkij znak*) for which gender is unpredictable and must be memorized.

The range of variation in the principles behind the organization of gender systems precludes easily finding a universal conceptual-semantic space in which gender categories can be distinguished and organized. Even among languages which use semantically transparent criteria to organize their gender systems, the same criteria are not necessarily used across languages. The same is true of gender systems organized using formal criteria, such as phonology. Therefore, unlike the other dimensions of meaning and distinctions encoded by the UniMorph Schema, it is not feasible to establish a conceptual-semantic basis for providing universally-applicable definitions for all attested gender categories.

However, it is still possible to limit the number of gender features that must be used. Within language families, gender categories and their assignment to nouns can overlap among languages enough to render the equation of gender categories appropriate. The gender features of this schema are therefore often tied to particular language families. This ensures that gender features have some generality, and do not proliferate as language-specific features.[21] For example, the eight possible genders (or noun classes) of the Nakh-Daghestanian languages are labeled as NAKH1, NAKH2, NAKH3, NAKH4, and so on for languages like Tsova-Tush (Corbett 1991:24). Similarly, the Bantu noun classes are symbolized as BANTU1-23, which encompasses the whole range of noun classes that could have been inherited in any given Bantu language from proto-Bantu (Demuth 2000:270-272).

[21] Preliminary experiments with morphological projection suggest that this approach will work well. The gender features used in this schema were applied in annotating material from Bibles in 47 languages from different families (although with a heavy Indo-European representation), and agreement for nouns aligned across languages reached approximately 75% (Sylak-Glassman et al. 2015a).

Other gender systems from other language families can be incorporated on this same model, with features of a similar format, namely a short abbreviation for the language family followed by the term used for the specific gender. The schema includes features for the masculine, feminine, and neuter distinction common in Indo-European languages.

Feature	*Label*
Masculine	MASC
Feminine	FEM
Neuter	NEUT
Bantu Noun Classes	BANTU1-23
Nakh-Daghestanian Noun Classes	NAKH1-8

Table 13: Cross-linguistically common, but language-family-specific, gender features

5.12 Information Structure

Information structure is a component of grammar that formally expresses "the pragmatic structuring of a proposition in a discourse" (Lambrecht 1994:5). More concretely, information structure directly encodes which parts of a proposition are presented as new information for the addressee and which are not (5-6).[22]

Languages may distinguish the *topic* (TOP) and the *focus* (FOC) of a sentence through overt morphology. The topic can broadly be seen as signaling what the sentence is about. Lambrecht (1994:131) defines the topic more specifically as "expressing information which is relevant to [a referent in the proposition] and which increases the addressee's knowledge of this referent." The focus can broadly be understood as signaling information that is not presupposed by the addressee (213). This information forms the core of the proposition's assertion, and typically includes the part of the proposition that is unpredictable or new to the listener (ibid.). Lambrecht (1994:213) formally defines focus as "[t]he semantic component of a pragmatically structured proposition whereby the assertion differs from the presupposition."

The contrast between topic and focus can be seen in the examples in (27) from Imbabura Quichua, which uses the clitics *=ka* and *=mi* to mark topic and focus, respectively (Tellings 2014:526).

(27) Topic and focus in Imbabura Quichua (Tellings 2014:526)

a. Pi-taj Pidru-ta=ka riku-rka?
who-INT Pedro-ACC-TOP see-PST
'Who saw Pedro?'

b. Pidru-ta=ka Marya=mi riku-rka.
Pedro-ACC-TOP Maria-FOC see-PST
'Maria saw Pedro.'

In (a), the question is about Pedro, who is marked as the topic with *=ka*. In (b), the question continues to be about Pedro, and the new information, marked by the focus marker *=mi*.

[22] Lambrecht (1994:5) formally defines information structure as the "component of sentence grammar in which propositions as conceptual representations of states of affairs are paired with lexicogrammatical structures in accordance with the mental states of interlocutors who use and interpret these structures as units of information in given discourse contexts."

While most studies on information structure distinguish multiple types of topic and focus (including Lambrecht 1994) that can be distinguished via prosody or syntax, no language appears to make distinctions via overt affixal morphology to a finer degree than topic vs. focus vs. unmarked. Lambrecht (1994:119) explains the lack of any formal topic marking in many languages, or of more fine-grained topic marking in languages which do formally mark it, as a result of the difficulty of strictly limiting the topic to any particular element in a proposition. For focus marking, Lambrecht (1994:225) notes that prosodic prominence of a given syllable in the sentence occurs for all examples of focus in English, French, Italian, and Japanese, and that prosodic prominence is the only formal means which can be used by itself to indicate focus. He speculates that "the role of prosody in focus marking is in some sense functionally more important than morphosyntactic marking." This would explain the absence of morphological focus marking in many languages as well as the absence of finer distinctions in focus marking morphology. Since prosodic prominence may always be available by default as a formal means of focus marking, even one morphological means of marking focus would allow at least three types of focus to be distinguished via prosodic prominence without morphology, both prosodic prominence and morphology, and morphology without prosodic prominence.

The two features necessary for marking information structure via overt affixal morphology are given in Table 14.

Feature	*Label*
Topic	TOP
Focus	FOC

Table 14: Features necessary for encoding information structure via overt affixal morphology

5.13 Interrogativity

Interrogativity is the dimension of meaning indicating whether a verb is used to express a statement (declarative; DECL) or a question (interrogative; INT). Interrogativity is sometimes marked by overt affixal morphology while declarativity is almost never marked. One of the few possible exceptions to this generalization is the morpheme *-aš* in Kabardian (Northwest Caucasian), the absence of which on a verb "creates a neutral irrealis, or a simple interrogative" (Colarusso 1992:125). Interrogativity, on the other hand, is more commonly overtly marked, as in Turkish where interrogativity is a feature that partly defines the verbal paradigm.

Feature	*Label*
Declarative	DECL
Interrogative	INT

Table 15: Interrogativity features

5.14 Language-Specific Features

While most inflectional variation can be tied to regular semantic differences, some inflectional variation is free or subject to not-strictly-semantic conditioning factors. For example, the imperfect tense of the subjunctive mood in Spanish can be realized with an inflection that includes either *-ra-* or *-se-*. Although the choice of forms is not entirely free, the factors that condition the choice

are not tied to meaning captured by inflection. That is, the forms cannot be differentiated by a meaning captured by a feature of this schema. A similar case to that of Spanish is the two forms of the genitive singular in German nouns (e.g. the GEN.SG forms of *Buch* 'book' are *Buchs* and *Buches*). The choice of which form to use cannot be captured by a feature in this schema (nor simply one that should be added).

To capture these distinctions in features within the schema, templatic features with the base LGSPEC followed by an integer (e.g. LGSPEC1) are used. These features have different interpretations in each language, and no mapping should be assumed between the LGSPEC features and any other features across languages. This approach avoids the problem of feature proliferation and possible conflict that results if the form of the feature is closely tied to the realization in a particular language. For example, hypothetical features like *RA and *SE for Spanish would necessarily be distinct from hypothetical features for German like *S or *ES (or any alternate forms of these features). However, in the approach adopted here, *RA is instead LGSPEC1 and *SE is LGSPEC2 in Spanish, and *S is LGSPEC1 for German with LGSPEC2 for *ES.

5.15 Mood

Grammatical mood is the morphological marking of modality on a verb. "Modality is concerned with the status [from the speaker's point of view; JCS] of the proposition that describes the event" (Palmer 2001:1), and can be expressed either by affixal morphology or through periphrastic/phrasal constructions. Palmer (2001:6-10, 22) provides a classification of the various types of modality, shown in Figure 16.[23]

Propositional Modality	
Epistemic	"speakers express their judgments about the factual status of the proposition" (8)
Speculative	possible, but uncertain, "may" (6)
Deductive	only possible conclusion, "must be" (6)
Assumptive	a reasonable conclusion, "I suppose, I guess, will ... since ..."
Event Modality	"refer[s] to events that are not actualized, events that have not taken place but are merely potential" (8)
Deontic	"the conditioning factors are external to the relevant individual" (9)
Permissive	"allowed to; may" (7)
Obligative	"must do so" (7)
Commissive	"You shall have it tomorrow" (10, quoting Searle 1983)
Dynamic	"[the conditioning factors] are internal [to the relevant individual]" (10)
Abilitive	"can," ability to do (10)
Volitive	willingness to do something (10)

Table 16: Categories of modality (Palmer 2001:22)

The morphological marking of modality tends to group these primary categories of modality (along with other less commonly expressed distinctions in modal meaning) into superordinate

[23] I omit evidentiality (discussed in §5.9, p. 24), which Palmer (2001:22) and other authors classify as a type of epistemic modality. Though the meaning of evidential categories often entails certain assumptions regarding the modality of the proposition, the UniMorph Schema sides with Aikhenvald (2004) in asserting that evidentiality is a distinct category that is necessarily associated with modality, but not a subdomain of it. Specifically, evidentiality is concerned with the source of information, not the evaluation of its status as certain or uncertain.

categories. These superordinate categories represent a binary distinction, with one category expressing one set of primary categories and another expressing the rest. Most commonly, languages split mood into two superordinate categories, either indicative vs. subjunctive or realis vs. irrealis (Palmer 2001:3-5).[24] While there is some distinction in the modalities that fall under the categories within these two types of oppositions, these two sets of superordinate categories can be reduced to an underlying opposition between Realis and Irrealis. Generally, Realis modalities have the status of objective truth and reality, and Irrealis modalities are regarded as unreal, uncertain, or hypothetical.[25] To some extent, the choice of terminology is based on the tradition of description for the language, with Indo-European languages typically described in terms of indicative vs. subjunctive and Native American and Papuan languages (among others) described in terms of realis vs. irrealis. In both systems, the subjunctive or irrealis category is the non-default category in the sense that a defined set of categories of modality are assigned to it and any others that are leftover are assigned to the other category (which is therefore the default and is less likely to be indicated with overt surface morphology).

Table 17 compares the set of basic modalities that are associated with either the subjunctive or irrealis, or both.

	Subjunctive	*Irrealis*
Unique	Speculative	Future
	Concessive/Presupposed	Potential
	Permissive	Abilitive
	Reported	Interrogative
	Purpose Resultative	Inferential
	Optative/Volitive	Warnings
		Customary/Habitual
		Infrequentative ("rarely, seldom")
Shared	Imperative	
	Jussive	
	Desiderative/Volitive/Optative	
	Negation	
	Prohibitive	
	Negative Conditional	
	Counterfactual/Prescriptive	
	Simulative 'as if'	
	'Lest'	
	Timitive/Apprehensive	
	Obligative	
	Purpose	
	Condition of "if ..." clauses (protasis)	
	Verbs in subordinate clauses	

[24] The distinction in capitalization between Realis/realis and Irrealis/irrealis is intentional, following Palmer (2001), and is meant to differentiate the underlying binary distinction across all languages with such a binary division between the categories Realis/Irrealis and the language-specific surface contrasts described as realis/irrealis or indicative/subjunctive.

[25] Palmer (2001:3), citing Bolinger (1968), Terrell and Hooper (1974), Hooper (1975), and Klein (1975), notes that "[i]t has been argued that the use of the 'indicative' and the 'subjunctive', which are the traditional terms used in many European languages for the distinction between Realis and Irrealis, can be accounted for in terms of 'assertion' and 'non-assertion'." The distinction is not between "what is factual and what is not (and still less on what is true and what is not true)," but between what speakers assert is factual and what they do not (Palmer 2001:3).

Table 17: Comparison of the cross-linguistic uses of categories termed, in each language, 'subjunctive' and 'irrealis' based on discussion (data from Palmer 2001:108-139, 146-160)

Similar to the subjunctive/irrealis is the category found in Australian languages called the *purposive*. The purposive primarily "expresses obligation (and epistemic necessity) in main clauses," and can also be used in that context in Dyirbal "to suggest a result from an unknown cause" (Palmer 2001:83). In Yidiny, it can "express a natural result" and indicate purpose in a subordinate clause (ibid.). Finally, it can express indirect commands in Ngiyambaa (ibid.). Palmer (ibid.) notes that "these functions of the 'purposive' are very like those of the subjunctive in Latin, which can also be used for purpose, result and for indirect commands." Given this information, the non-purposive (AUNPRP) vs. purposive (AUPRP) opposition can be viewed as on par with the realis vs. irrealis and indicative vs. subjunctive oppositions.

These superordinate categories of modality are not only frequently encountered in language descriptions, their uses often overlap substantially across languages to the point that equating the categories across languages can be useful. For this reason, the UniMorph Schema includes features for indicative (IND), subjunctive (SBJV), realis (REAL), irrealis (IRR), and the Australian purposive (AUPRP) and non-purposive (AUNPRP).

Apart from the superordinate categories indicative vs. subjunctive, realis vs. irrealis, and non-purposive vs. purposive, the most common modal marking includes the marking of imperatives, hortatives, and jussives. Imperatives are direct commands or orders for the addressee to do some action while hortatives and jussives include more suggestive forms, such as "let them X" or "let us X" (where X is some action). These can all be understood as part of the same modal category, called here imperative-jussive, since imperatives are used with second person and jussives are complementarily used with first and third person (Palmer 2001:81, citing Lyons 1977).[26] Related to imperative-jussives are prohibitives. While these are often morphologically distinct from positive imperative-jussives, they can be analyzed underlyingly either as negative imperative-jussives ("Do not do X!"), negative potentials ("You cannot do X"), or negative permissives ("You may not do X").

Other common modal categories can be seen as expressing degrees of certainty about the factual truth of a proposition, further refining the epistemic category of modality called *speculative*. A small number of languages explicitly morphologically mark verbs for likelihood status, such as "real, likely, and potential" in Dani (Papuan; Palmer 2001:162). While marking verbs as "likely" using morphology is extremely rare, many languages exhibit a potential mood, including Finnish. The potential mood is translated "may" in the sense of "may or may not, depending on circumstances." Another modal category, the admirative, is used to express surprise, irony, or doubt, and occurs in Bulgarian, Macedonian, and other Balkan languages, as well as in Caddo (a Native American language of Oklahoma), where it marks only surprise (Palmer 2001:11). The admirative can be seen as a mood that expresses surprise or skepticism toward a fact that is acknowledged by the speaker to be true for the addressee (ibid.).

The conditional mood, familiar from Spanish, is used to express that a verb's action will occur given some condition and can generally be translated "would do X." The simulative, which occurs

[26]Some descriptions use 'imperative' to cover all persons, which is functionally equivalent to what is suggested here by the use of imperative-jussive. The term 'imperative-jussive' is suggested here to unite the range of meanings that are covered by forms such as "let us speak, speak!," and "let them speak." No language has been found to contrast imperative and jussive or hortative forms for the same grammatical person.

in Caddo (Palmer 2001:178), also expresses hypothetical action, but in the sense of "as if X." A further remark is in order on the conditional. When grammatical descriptions cite the existence of a conditional mood, for example in *if ..., then ...* statements, it may be used for, 1) the protasis, or condition itself (*if*-clause), 2) the apodosis, or the result of a condition applying (*then*-clause), or 3) both.[27]

Distinct from the Australian purposive mood, a superordinate category, is the basic purposive modality, signaling "in order to, for the purpose of." This purposive has the label PURP, while the Australian purposive is labeled AUPRP (and its non-purposive counterpart is AUNPRP).

Languages such as Tonkawa (isolate; Palmer 2001:82) also mark verbs as intentive, indicating that the speaker strongly intends for the action of the verb to be realized. This can be translated as "going to" or "will," although this renders it ambiguous with the prospective aspect or future tense, respectively, in English. Few languages mark the intentive explicitly, but as a feature of modality, it is common, lending the phrases "going to" and "will" the connotation of certainty in addition to prospective aspect and future tense, respectively.

Languages can also morphologically mark obligation, translated by "have to" or "must." Tiwi, an Australian language, has "a marker [i.e. a morpheme *-u*; JCS] that is labelled 'compulsional"' and is translated as "have to" or "must" (Palmer 2001:75). This meaning is captured by the obligative mood (OBLIG).

It is also possible for languages to use the same morpheme to mark distinct, but related, categories of modality. A case in point comes from the Tamil morphemes *-laam* "may, might" and *-ŋum* "ought to, must be" (Palmer 2001:27). The distinction between these two morphemes can be interpreted as, respectively, permissive vs. debitive/obligative (deontic modality; ibid.) or speculative vs. deductive/inferential (epistemic modality; 72).

Another commonly attested type of modality is the optative or desiderative modality, which can be used to express wishes or other notions related to desire (Palmer 2001:22, 131). For example, the suffix *-naya* in Imbabura Quechua is called the "desiderative" and is translated as "want(s) to" (Cole 1982:181). In other languages, such as Limbu, a dedicated morpheme, e.g. *-lɔ*, is used to express wishes such as "May it turn out well!" (van Driem 1987:133). This modality is traditionally called optative. It is unclear whether the desiderative morpheme in Imbabura Quechua can be used this way and whether the optative in Limbu can be used in the first person singular (it can be used in all other persons and numbers, though; van Driem 1987:133-135). No language has been found thus far to overtly contrast the desiderative and optative with distinct affixal morphology.

Finally, other marginal types of modal morphology are attested, but often they occur in only a single language or are not clearly modal. For example, Caddo contains a morphological marker that indicates that the verb occurs rarely or infrequently (Palmer 2001:175). This could be seen as marking "unlikely" within the category of speculative modality.

Table 18 shows the modal features that are necessary to capture the overtly morphologically-marked modal categories encountered in this survey.

Feature	*Label*
Indicative	IND
Subjunctive	SBJV
Realis	REAL
Irrealis	IRR
Australian Purposive	AUPRP

[27] If necessary, features for marking these distinctions might take the form COND.PROT and COND.APO for conditional marking the protasis and the conditional marking the apodosis, respectively.

Australian Non-Purposive	AUNPRP
Imperative-Jussive	IMP
Conditional	COND
General Purposive ('in order to')	PURP
Intentive	INTEN
Potential	POT
Likely	LKLY
Admirative	ADM
Obligative	OBLIG
Debitive	DEB
Permissive	PERM
Deductive	DED
Simulative	SIM
Optative-Desiderative	OPT

Table 18: Features necessary for encoding surface realizations of, and distinctions between, modal categories based primarily on data and discussion in Palmer (2001)

5.16 Number

The dimension of number is relevant for multiple parts of speech and is a common agreement feature. Its range of distinctions on nouns is most extensive, with less common categories like "greater paucal" being expressed in a small number of cases on nouns, but never on verbs.

The number categories found on nouns include singular, plural, dual, trial, paucal, greater paucal, greater plural, and inverse marking (Corbett 2000). Singular (SG) and plural (PL) are familiar from the vast majority of languages. Singular always indicates one, but plural signals "many" when the number is larger than the largest more highly specified distinction, e.g. >2 when plural is opposed to singular, but >3 when opposed to dual, and so on. Dual (DU) and trial (TRI) indicate precisely two or three entities, respectively, and these are robustly distinguished for nouns in Larike (21, 40, 45). In addition, languages can distinguish these from "a few, several," which is indicated using the the paucal number (PAUC). Finally, yet another number, greater paucal (GPAUC), for more than several, but not many, is used in Sursurunga (Austronesian), which has the most extensive nominal number distinction known to exist, distinguishing singular, dual, paucal ($\geq$3), greater paucal ($\geq$4), and plural (26-27). Some languages make further distinctions within the plural category that are based on additional meaning beyond countable number. For example, Arabic contains a "greater plural" or "plural of abundance" that indicates "various, many" (32). Banyun marks an "unlimited" plural (31) and both Hamer (Omotic) and Kaytetye (Australia) have a plural form that indicates "all possible" entities (33). A greater plural is also claimed to be present in Zulu, Setswana, Miya, and Breton (ibid.). While the precise semantics of a greater plural vary across languages, no language appears to morphologically distinguish more than one variety of greater plural with the plain plural.

Another possible number marking is called inverse marking. In inverse number systems, nouns have a default number that indicates the number with which they are "expected" to occur, based on a language-internal understanding of their occurrence in the speakers' surroundings. Variation exists across languages with inverse marking in this respect, but, for example, 'child' is by default singular while 'tree' is by default plural. Inverse number marking makes 'child' plural and 'tree'

singular, effectively inverting the number value of noun. Corbett (2000:161) explains inverse number in Kiowa as follows.

> "If we concentrate on the main classes of nouns we see that each has a basic form, without *-gɔ̀* and a 'less expected' form with it. Those denoting animates are singular/dual in their basic form, with *-gɔ̀* signalling a shift from that number, while nouns in the main class for inanimates are treated as basically dual/plural, with *-gɔ̀* and variants signalling a shift to singular."

A similar inverse system has been shown to exist in Dagaare (Grimm 2012).

On verbs, Amele (Papuan) encodes singular, dual, and plural, and Kiowa (Kiowan-Tanoan) (Corbett 2000:136-137, 159-161) encodes singular, dual, and inverse. Meriam (Meriam Mir; Trans-Fly; Torres Strait Islands) and Sursurunga add another number by marking singular, dual, paucal, and plural on verbs (Corbett 2000; Piper 1989:26-30). Kiwai also encodes four categories of number on verbs, but marks the singular, dual, trial, and plural (255). Thus, the only number category that seems not to be expressed on verbs is greater paucal and the greater plural.

Another limitation in the expression of number is that while the mass-count distinction is often grammatically relevant, it seems never to be expressed with unique morphology. That is, distinctions between mass and count nouns are typically realized via the usage of the number categories already described, not via unique morphology to indicate mass or count. This schema does not include features dedicated to expressing whether a noun is mass or count.[28]

A potentially computationally challenging phenomenon is constructed number. Constructed number refers to a specific number interpretation that arises from the combination of two or more words. Constructed number is used to convey the dual number with nouns in Zuni (Uto-Aztecan). When a plural noun is used with a verb that encodes only singular number, a dual interpretation results (Corbett 2000:170).

(28) Constructed dual in Zuni

ʔaːw-akcek(ʔi) ʔaː-kya
PL-boy go-PST

"Two boys went"

In the Talitsk dialect of Russian, a singular name used with a plural verbal inflection yields the interpretation "*name* and his family/group" (Corbett 2000:191-192).

(29) Constructed number in Talitsk Russian

Gosha priexa-l-i
Gosha.M.SG arrive-PST-PL

"Gosha and his family have arrived"

The features necessary to mark number, both on nouns and verbs, are listed in Table 19.

Feature	*Label*
Singular	SG
Plural	PL
Greater plural	GRPL

[28] Features for these distinctions, which are both cross-linguistically variable and lexical, could be MA for mass and CT for count.

Dual	DU
Trial	TRI
Paucal	PAUC
Greater paucal	GPAUC
Inverse	INVN[29]

Table 19: Number features for all parts of speech

5.17 Part of Speech

Croft (2000:89) defines the funtionally-motivated conceptual space in Table 20 for parts of speech. It is the cross-product of the concepts of *object*, *property*, and *action* with the functions of *reference*, *modification*, and *predication*. This conceptual space provides definitions for the following cross-linguistically common parts of speech, which are all captured by features in the UniMorph Schema: Nouns (N), adpositions (ADP), adjectives (ADJ), verbs (V), masdars (V.MSDR), participles (V.PTCP), converbs (V.CVB), and adverbs (ADV).

	Reference	*Modification*	*Predication*
Object	object reference: nouns	object modifier: adpositions	object predication: predicate nouns
Property	property reference: substantivized adjectives	property modifier: (attributive) adjectives, participles	property predication: predicate adjectives
Action	action reference: masdars	action modifier: adverbs, converbs	action predication: verbs

Table 20: Functionally-motivated conceptual space defining basic parts of speech, adapted from Croft (2000:89)

Nouns are a basic part of speech in the sense that they need not co-occur with any other part of speech and are not logically a subclass of any other part of speech. They typically control features such as number and gender and can bear case features, among other dimensions. Nouns, along with verbs, have been claimed to be universally instantiated across languages (Baker 2003).

A subset of nouns are the proper names, which have syntactic properties that differ from nouns as a whole (Van Langendonck 2007; Anderson 2004). For example, names in English are typically definite by default, and can only take an article if they are coerced into the role of a common noun. This illustrates the theoretical distinction between a proper name, which is a name used as such, and a proprial lemma, which is a word that functions as a proper name by default but can be coerced into functioning like a common noun, for example through the use of articles in English as in *an Alex* or *the Alex that I met yesterday* (Van Langendonck 2007; Van de Velde 2012). A special construction such as apposition or name-giving is required to coerce a common noun into the role of a proper name (more precisely, an appellative, e.g. *the metal, gold, commonly used by jewelers* or *this metal is called 'gold'*; Van Langendonck 2007:95, 171). While distinguishing proper names is motivated by linguistic properties, it is also very useful for named entity recognition tasks. In the UniMorph Schema, proper names and proprial lemmas are not given different features, and both

[29]Note that inverse number is labeled INVN while inverse for direct-inverse voice systems is labeled INV.

are captured by PROPN.

Adjectives are the prototypical modifiers of nouns. Used attributively, they must co-occur with a noun, and they often agree with features controlled by that noun. Even when used predicatively, adjectives can still be subject to the features that their head noun controls (e.g. across a copula). For example, adjectives in Russian must agree in number, gender, and case with their head noun. The phenomenon of substantivization represents a cross-linguistic exception to the generalization that adjectives must co-occur with nouns. In cases of substantivization, an adjective functions like a noun. For example, the adjective *poor* is substantivized in the English sentence *the poor face significant disadvantages*. The adjective is allowed to directly take a definite article and to control number features (here, plural). Similarly, in Russian, the word for 'animal,' *životnoe*, is adjectival in its morphological surface form, but, like a noun, controls number, gender, and case features. For example, in (30), *životnoe* causes the adjective *zdorovoe* to bear neuter gender and singular number, and the verb *bežit* bears the singular number in agreement.

(30) zdoróv-oe živótn-oe bež-ít
healthy-NEUT.SG animal-NEUT.SG run.IPFV-PRS.3.SG
'The healthy animal runs.'

Pronouns can be considered a subclass of nouns because, like nouns, they can function as the syntactic arguments of verbs. However, pronouns are distinct from nouns by virtue of being inherently deictic (see §5.8, p. 22) and by their syntactic and morphosyntactic properties. Pronouns refer to other nouns in the sentence (or in previous discourse), and unlike nouns, this reference is regulated by a variety of syntactic principles, including island constraints, crossover, and others. Pronouns may also exhibit different morphosyntactic properties from nouns. For example, pronouns are the only words in English which have morphologically marked case, e.g. *she* (nominative) from *her* (accusative/oblique) and *who* (nominative) from *whom* (accusative/oblique).

Classifiers are morphemes or words that "denote some salient perceived or imputed characteristic of the entity to which the associated noun refers (or may refer)" (Allan 1977:285). Classifiers can occur as isolated words, for example in so-called "numeral classifier" languages of East and Southeast Asia (Allan 1977:286), or as affixes, for example as with noun class markers in Bantu languages (286-287). Thai exhibits classifiers that are isolated words used to quantify nouns such that a quantified NP has the structure N NUM CLS. The examples in (31), from Allan (1977:286), exemplify variations of this structure.

(31) Classifiers in Thai (Allan 1977:286)

a. khruː lâːj khon
teacher three person.CLS
'three teachers'

b. mǎː sìː tua
dog four body.CLS
'four dogs'

c. mǎː tua nán
dog body.CLS that
'that dog'

d. tua nán
body.CLS that
'that [animal, coat, trousers, or table]'

e. sìː tua
four body.CLS
'four (of them) [animals, coats, trousers, or tables]'

Classifiers have been argued to capture a limited range of semantic contrasts (Allan 1977). If this is the case, it renders a featural analysis like that done for other dimensions of meaning feasible and would allow additional semantic information to be extracted.

Articles modify nouns and explicitly indicate whether the noun is definite or indefinite. For example, the English articles *the* and *a(n)* designate a following NP as definite and indefinite, respectively. Languages differ in whether definiteness is marked at all, and then in whether both definite and indefinite are overtly marked. For example, Russian does not mark definiteness at all, and Hebrew overtly marks only definite nouns (with the article *ha*). Articles are a subset of the larger class of determiners, which can include demonstrative adjectives and numerals as well.

Verbs, like nouns, are a basic part of speech in that they need not co-occur with any other part of speech and are not a subclass of any other part of speech. Like nouns, it has been claimed that verbs are present in every human language (Baker 2003).[30] Their typical function is as predicates which describe an action.

Verbs can also form the basis for productively deriving three parts of speech, namely participles, masdars, and converbs, that each behave similarly to adjectives, nouns, and adverbs, respectively. While these three parts of speech could be subsumed within the part of speech that they behave as (i.e. adjectives, nominals, and adverbs), the fact that they are derived from verbs typically entails other syntactic properties or restrictions that do not apply to non-derived exemplars of the part of speech they behave as. For example, these derived forms are often still able to govern nouns as direct objects (as in (32) below), and these properties motivate treating them as a class distinct from both the part of speech from which they are derived (verbs) and those which they resemble most strongly from a functional perspective.

One of these derived parts of speech is the participle, or verbal adjective, which is a form derived from a verb that modifies nouns similar to an adjective. In Russian, for example, the present active participle has adjectival inflection and agrees in number, gender, and case with the noun that it modifies. For example, the participial form *t͡ʃitajuʃt͡ʃij* "reading" is derived from the verb *t͡ʃitatʲ* "to read (IMPERFECTIVE)" and can be seen to agree in number, gender, and case in the following example.

(32) Devuʃk-a, t͡ʃitajuʃt͡ʃ-aja doklad, ne otveti-l-a
girl-F.NOM.SG reading V.PTCP-NOM.F.SG report.M NEG answer-PST-F.SG
'The girl, reading the report, did not answer.'

The participle in this example retains verbal characteristics since it governs a noun (*doklad* 'report') as a direct object. However, the participle behaves like an adjective in that it bears agreement features from the subject, *devuʃka* 'the girl.'

The second part of speech derived from verbs is the masdar, which is also called a verbal noun, gerund, and in some cases, an action nominal (Koptjevskaja-Tamm 1993).[31] In English, examples of masdars include some 'gerundive' forms ending in *-ing*, such as *running* in *the running of the race*

[30] Data from the Wakashan languages of the North American Pacific Northwest region in both Canada and the United States sparked a debate on whether nouns and verbs could be considered universally distinct. Although these parts of speech are distinguished on the surface in these languages by overt morphology, many roots can be considered ambiguous as to their verbal or nominal status.

[31] The term 'masdar' is from Arabic grammar (/masˤdar/) and is preferred by Haspelmath (1995:4,48) because, like the word 'participle,' it "consists only of a single root" (48).

happened quickly. These are also action nominals, which are taken to be the default interpretation of a masdar.

The third part of speech derived from verbs is the converb, which can be viewed as a verbal adverb, and is also commonly called a gerund or adverbial participle. Haspelmath (1995:3) defines a converb as "a nonfinite verb form whose main function is to mark adverbial subordination." Examples of converbs come from Modern Greek, Portuguese, and Huallaga Quechua, shown in (33), (34), and (35).[32]

(33) Modern Greek (Haspelmath 1995:1)

I kópela tón kítak-s-e xamojel-óndas.
the girl him look-AOR-3SG smile-CVB

'The girl looked at him smiling.'

(34) Portuguese (Haspelmath 1995:1)

Despenhou-se um avião militar, morr-endo o piloto.
crashed a plane military die-CVB the pilot

'A military plane crashed, and the pilot was killed.' (lit. '..., the pilot dying')

(35) Huallaga Quechua (Haspelmath 1995:2, citing Weber 1989:304)

Aywa-ra-yka-r parla-shun.
go-STAT-IPFV-CVB converse-1PL.INCL.IMP

'Let's converse as we go along.'

Adverbs modify verbs similarly to how adjectives modify nouns. Unlike converbs, they are often derived from adjectives or are non-derived lexemes. In some languages, such as English and Russian, adverbs can be productively derived from most adjectives, e.g. with the suffix *-ly* as in *quickly* in English and with the suffix *-o* as in *xorošo* 'well' from *xorošij* 'good.'

Similar to how pronouns are a subclass of nouns, auxiliaries are a subclass of verbs that are used for grammatical functions rather than to convey the type of lexical meaning associated with standard verbs. Examples of auxiliary verbs in English include the modals *could*, *would*, and *should*, as well as forms of *have* used to construct periphrastic perfect verb forms (e.g. *he has run*) and forms of *be* used to construct periphrastic passive verb forms (e.g. *he was hit* or *he has been hit*).

A number of other parts of speech also exist and will be illustrated here with simple examples from English. Adpositions include prepositions, postpositions, circumfixes, and infixes, which are all distinguished by where the adposition is placed in relation to its head. Prepositions, such as "in," occur before their head while postpositions, such as "ago," occur after. Note that spatial adpositions are often absent in the native vocabulary of languages with local case systems (§5.5).

Complementizers, such as English *that*, Spanish *que*, and Russian *što*, are used to embed clauses, as in *I understand that complementizers are important* or *That complementizers are important is clear*.

Conjunctions include both coordinating conjunctions and subordinating conjunctions.[33] Coordinating conjunctions, such as English *and*, link two clauses (or two arguments) in such a way that they occupy an equal syntactic position. Subordinating conjunctions link two clauses such that one

[32]The abbreviation for converb in the source, CONV, has been converted to the Leipzig Glossing Rules' standard abbreviation CVB, which is also adopted for the Universal Morphological Feature Schema.

[33]These are not distinguished in this schema, but are distinguished in Universal Dependencies (Choi et al. 2015).

clause is dependent on the other in some respect. For example, in the sentence *I did not see John because he left*, *because* is the subordinating conjunction, and it is clear that *he* refers to *John*.[34]

Other less canonical parts of speech can be identified. Following the practice of the Universal Dependencies Project (Choi et al. 2015), numerals, particles, and interjections are also treated as separate parts of speech. Numerals include cardinal numerals. Where appropriate, ordinal numerals (first, second, etc.) often behave as adjectives and should be classified as such. Particles include isolated, generally mono- or bisyllabic, words with varying grammatical and discursive meanings. Japanese is well-known for having a large number of particles, which include, among other kinds, question particles such as *ne* and *kasira* as well as particles emphasizing assertion, such as *yo*, *zo*, and *ze* (Tsujimura 2007:435-436). Interjections include words such as English *ouch*, *oof*, etc., which are expressive, but have neither grammatical nor lexical meaning.

Table 21 shows the features that are necessary to indicate part of speech.

Feature	*Label*
Noun	N
Proper Name	PROPN
Adjective	ADJ
Pronoun	PRO
Classifier	CLF
Article	ART
Determiner	DET
Verb	V
Adverb	ADV
Auxiliary	AUX
Participle (Verbal Adjective)	V.PTCP
Masdar (Verbal Noun)	V.MSDR
Converb (Verbal Adverb)	V.CVB
Adposition	ADP
Complementizer	COMP
Conjunction	CONJ
Numeral	NUM
Particle	PART
Interjection	INTJ

Table 21: Part of speech features

5.18 Person

The conventional person categories that are encoded on verbs in most languages include first person (1), second person (2), and third person (3). Apart from these common distinctions, some languages also distinguish another category of person, zero person (0), and each conventional person category is sometimes subdivided further.

Finnish has a "zero person" construction, which lacks an overt subject and is used to make "generic statements concerning human beings" (Laitinen 2006:209), as exemplified in (36):

[34]The fact that the dependency relation, and not just linear order, allows for the pronoun reference can be seen by the acceptability of the similar sentence, *Because he left, I did not see John.*

(36) Suome-ssa joutu-u sauna-an
Finland-INESSIVE get-3.SG sauna-ILLATIVE
'In Finland, you wind [JCS: one winds] up in a sauna' (ibid.).

While this construction is distinctive in Finnish, it does not give use unique morphology that would necessarily require a feature for zero person, even though the distinctive meaning is clear. However, in Santa Ana Pueblo Keres, such a zero person is morphologically distinct from other persons (Davis 1964:75) and is marked with a special pronominal affix, demonstrating the necessity for a zero person feature (0).

The conventional person categories often exhibit further distinctions within them. For example, first person plural ('we,' 1;PL) can be divided into inclusive (INCL), i.e. including the addressee, or exclusive (EXCL), i.e. excluding the addressee. This distinction is made in Ingush, which uses *vai* for first person inclusive and *txo* for first person exclusive (Nichols 2011:173-176). Bickel and Nichols (2005:53) show that the inclusive/exclusive distinction is made by 116 of 293 languages in their sample, or approximately 40%. Although the inclusive/exclusive distinction is typically viewed as a feature of the first person plural, independent features are used for inclusive (INCL) and exclusive (EXCL).[35] These can be used with first person using complex annotations such as 1+INCL and 1+EXCL.

Second person is often divided according to politeness categories, such as informal and polite. The dimension of politeness, which is manifested beyond person distinctions, is the subject of the politeness features in §5.20, p. 42.

Third person can be divided not only by gender (§5.11, p. 27), but also by hierarchical status (based on information structure, animacy, or a combination) in languages with a pragmatic voice system, in the terms of Klaiman (1991). For example, in direct-inverse systems, when the subject and argument are at the same level of the salience hierarchy, one argument is usually overtly marked as proximate and the other as obviative. The focused or otherwise more highly-ranked argument is marked as proximate and all others as obviative. This system is common in Algonquian languages, such as Plains Cree (Aissen 1997:706-709).[36] For additional detail on pragmatic voice systems, see §5.25, p. 56.

In some languages, a fourth person category is used to describe an otherwise third-person referent that is differentiated from other third-person referents by a switch-reference-like distinction (e.g. fourth person for a same-subject [SS] verb form in Central Yup'ik [Woodbury 1982] or for "disjoint reference across clauses" in Navajo [Willie 1991:108]) or, more commonly, by a distinction in obviation status (Chelliah and de Reuse 2011:306-307), as in Keres (isolate; New Mexico, USA), in which "[f]ourth person is used [...] when the subject of the action is inferior to the object, as when an animal is the subject and a human being the object" (Davis 1964:76).[37] For the purposes of morphological distinctions, these fourth person categories may call for dedicated verbal morphology. While in some cases their meaning can be captured by third person (3) plus switch-reference features (§5.22, p. 49) or features marking pragmatic voice distinctions (such as the proximate (PRX) and

[35]Note that Daniel (2005) argues vigorously that inclusive and exclusive should not be viewed as subcategories of first person.

[36]Proximate and obviative could be seen as case marking features, information structural features, or voice features. They have been included here with person since they usually serve to differentiate third person actants. Note that "proximate" is a term used both in the sense here and for indicating spatial closeness in demonstrative pronouns. These senses are differentiated in the labeling, where PRX indicates the sense of "proximate" under discussion here while PROX is reserved for proximal/proximate in the spatial sense used in pronominal deixis.

[37]In Santa Ana Pueblo Keres, the fourth person is also used "when the subject of the action is obscure, as when the speaker is telling of something that he himself did not observe." This might be marked with an evidential feature, rather than a person feature, even though the person category is morphologically distinct.

obviative (OBV)), we include a fourth person category with the feature 4 to allow for identification of a fourth person category when the semantic distinctions are complicated or not strictly inflectional in nature.

The features necessary for encoding person are given in Table 22.

Feature	*Label*
Zero person	0
First person	1
Second person	2
Third person	3
Fourth person	4
Inclusive	INCL
Exclusive	EXCL
Proximate	PRX
Obviative	OBV

Table 22: Person features

5.19 Polarity

Polarity encodes whether a statement is meant to be negative (NEG) or positive (POS; also called affirmative). Like declarativity, positive polarity is rarely overtly marked, but negativity may be, again as in Turkish, where it is a feature of the verbal paradigm alongside interrogativity. Although negation phenomena in many languages can involve subtle, complex distinctions, this complexity arises from multiword constructions at the phrase or sentence level. Morphemes and words are specified as being either negative or positive without any finer distinctions in polarity.

Feature	*Label*
Positive	POS
Negative	NEG

Table 23: Polarity features

5.20 Politeness

Politeness is defined here as a dimension of meaning that grammatically encodes social status relationships between the speaker, addressee, third parties, and the setting in which a given speech act occurs.[38] This definition is based on Comrie's (1976b) categorization of the types of honorifics. Based on the idea that honorifics are part of a language's deictic system and encode social deixis (rather than, for example, spatial deixis; Fillmore 1975), Comrie (1976b) proposes three axes to which honorific deixis is oriented. These axes relate the speaker to: 1) the referent of the linguistic expression to which the honorific is attached, 2) the addressee, and 3) bystanders to the speech act (Brown and Levinson 1987:180-181). Brown and Levinson (1987:181) add a fourth axis, relating the speaker to the setting in which the speech act occurs. Defining politeness in terms of the system used to define honorifics provides coherent definitions for the T/V pronoun distinction in Indo-European languages, honorific morphology in Japanese, Korean, and Thai (among other

[38] Corbett and the Surrey Morphology Group discuss at least what is here called 'politeness' as 'respect' (Kibort 2010; Corbett 2012).

languages), avoidance or taboo language (including "mother-in-law language" in Guugu Yimidhirr, royal language in Pohnpeian), and finally also register distinctions (colloquial vs. literary, academic, etc.).

5.20.1 Speaker-Referent Axis

Levinson (1983:90) writes that with referent honorifics, "respect can only be conveyed by referring to the 'target' of the respect" and that "the familiar *tu/vous* type of distinction in singular pronouns of address ... is really a referent honorific system, where the referent happens to be the addressee." This is in direct contrast to speaker-addressee honorifics, which confer respect without ever referring to the addressee (Comrie 1976a via Brown and Levinson 1987:180). The best known example of speaker-referent honorific morphology is the Indo-European T/V pronoun distinction, manifesting as *tu/vous* in French, *tu/voi* in Italian, *ty/vy* in Russian, and so on.[39] Another example of this kind of system appears to be Yemsa (Omotic; Corbett 2012), which has distinct second and third person honorific pronominal forms with corresponding marking of these distinctions on verbs via agreement. However, the presence of an originally tri-partite (royal, polite, common) system of referring to 100 or more lexemes leads one to the conclusion that at least the third person system (and maybe the second person system as well) may be an addressee honorific system.

No extensive typological study has revealed the maximum number of distinctions that occur in speaker-referent honorific morphology. This study therefore proposes only several levels, based on only a few examples. The T/V distinction in Indo-European gives evidence for two levels, informal and formal. The Basque second person singular pronoun *hi* has a much more limited use (Hualde and Ortiz de Urbina 2003:150) than Spanish *tu*, for example, but both are the marked intimate/informal second person pronoun within their respective languages. Similarly, Gujarati *ap*, a very formal second person plural pronoun may convey a greater level of formality than Spanish *Usted*, but, again, both serve as the marked formal second person pronoun within their respective languages. Data from Japanese motivate positing two sublevels of the formal level. Japanese uses one set of referent honorific forms in a speech style called *sonkeigo* to elevate the referent and a distinct set of referent honorific forms in a speech style called *kenjōgo* to lower the speaker's status, thereby raising the referent's status by comparison (Wenger 1982:41-43). While the features for elevating (ELEV) and humbling (HUMB) are formally independent, they should be used only in conjunction with the FORM feature in the combinations FORM+ELEV and FORM+HUMB, respectively.

Feature	*Label*
Informal	INFM
Formal	FORM
Referent Elevating	ELEV
Speaker Humbling	HUMB

Table 24: Preliminary features for labeling speaker-referent honorific morphology

[39] Some Indo-European languages retain the distinction in status levels of two second person pronouns, but have replaced the historical second person plural formal pronoun beginning with the phoneme /v/ with a form corresponding to 'your highness.' This is the case in Spanish where *vuestra merced* 'your highness' was historically contracted to *Usted* and came to fill the function of historical *vos*. Another form, *vosotros*, is a combination of *vos otros* 'you others.' While it retains the V-form, it is used as an informal second person plural, with *Usted* used to indicate formality. A similar case to Spanish *Usted* is found in Romanian second person plural *dumneavoastră*, lit. 'lordship your,' where the first element arose from Romanian *domnia* 'lordship' (Dobrovie-Sorin and Giurgea 2013:283).

5.20.2 Speaker-Addressee Axis

Brown and Levinson (1987:276) define the speaker-addressee axis as the "direct encoding[...] of the speaker-addressee relationship, *independent of the referential content of the utterance*" (emphasis JCS). Levinson (1983:90) describes a speaker-addressee honorific system by saying that "... in many languages (notably the S. E. Asian languages, including Korean, Japanese and Javanese) it is possible to say some sentence glossing as 'The soup is hot' and by the choice of a linguistic alternate (e.g. for 'soup') encode respect to the addressee without referring to him, in which case we have an addressee honorific system." Japanese *teineigo* is an example of an addressee honorific system. Javanese (Austronesian) has an addressee honorific system containing a high-level form called *krama* and a form between that and common speech that contains fewer lexical items and is called *madya* (Wenger 1982:71). Addressee honorific systems are less common than referent honorific systems (Wenger 1982), and more linguistic research needs to be done to show the full range of possible distinctions within such systems.

Feature	*Label*
Polite	POL
Medium Polite	MPOL

Table 25: Preliminary features for labeling speaker-addressee honorific morphology

5.20.3 Speaker-Bystander Axis

Levinson (1983:90) describes these systems as those in which special language is used to show respect to bystanders, i.e. "participants in audience role and [...] non-participating overhearers." Examples of these kinds of systems include "Dyirbal alternative vocabulary ... used in the presence of taboo relatives" (90) and "certain features of Pacific languages, like aspects of the 'royal honorifics' in Ponapean [Pohnpeian]" (91). These systems are typically termed "taboo speech, avoidance language," or "court language."

Avoidance language is common among Australian languages. Although avoidance language is commonly called "mother-in-law" or "brother-in-law" language, it does not differ within a single language depending on who is addressed. That is, within a language, there is a single set of avoidance lexemes that are used with anyone to whom the avoidance relationship applies (e.g. mother-in-law, brother-in-law, cross cousins, etc.) (Dixon 1980:58-65). Like "court language," as in Pohnpeian, the phonology and grammar of avoidance language does not generally differ from that of the everyday language, only the lexicon differs (59).

The maximal attested number of levels within a bystander honorific system is five in Pohnpeian (Keating and Duranti 2006:151-152). These five levels include: 1) a low level, which the speaker uses in the presence of only those having a low status; 2) a common level which is unmarked for status; 3) a general high status that can be used specifically in the presence of the secondary chief and secondary chieftess; 4) a high status form for use in the presence of the (primary) chieftess; and finally, 5) a high status form specifically for use in the presence of the (primary) chief.

Table 26 lists features that can be used for speaker-bystander systems. A single feature is used to indicate avoidance style, while three core levels are used for court language systems, with the highest level elaborated as needed for systems like Pohnpeian and Samoan and with the neutral level unspecified.

Feature	*Label*
Avoidance style	AVOID
Low status	LOW
High status	HIGH
Elevated status	STELV ("status elevated")
Supreme status	STSUPR ("status supreme")

Table 26: Features for encoding bystander honorific distinctions

5.20.4 Speaker-Setting Axis

Levinson (1983:91) notes "that while the first three kinds of [politeness axes described here; JCS] are relative strictly to the deictic centre, here specifically the social standing of the speaker, formality is perhaps best seen as involving a relation between all participant roles and the situation [or setting; JCS]." This is a way of characterizing what is referred to as 'register' in sociolinguistics. Because speech can take place in such a wide variety of settings, the best approach to defining register features is an empirical one in which only registers that are associated with distinctive morphology should be defined here.

Examples of grammaticalized "register" distinctions include Japanese's "so-called *mas*-style, and in Tamil ... a high *diglossic variant*" (Levinson 1983:91). To these can be added the distinctive literary uses of the following tenses in French: *passé simple*, *passé antérieur*, *imparfait du subjonctif*, *plus-que-parfait du subjonctif*, and *seconde forme du conditionnel passé*.

Feature	*Label*
Literary	LIT
Formal register	FOREG
Colloquial	COL

Table 27: Features for expressing common speaker-setting politeness, or register, features

5.20.5 Politeness Features

Table 28 should be considered preliminary and incomplete, but with necessary features for marking common morphological methods of expressing politeness distinctions.

Feature	*Label*
Informal	INFM
Formal	FORM
Formal, Referent Elevating	ELEV
Formal, Speaker Humbling	HUMB
Polite	POL
Avoidance style	AVOID
Low status	LOW
High status	HIGH
High status, elevated	STELEV
High status, supreme	STSUPR

Literary	LIT
Formal register	FOREG
Colloquial	COL

Table 28: Preliminary features for encoding levels of politeness

5.21 Possession

While languages often use separate possessive adjectives (such as *my*, *your*, *his*, *her*, *our*, and *their* in English) to mark possession, some languages, such as Turkish and certain Quechua languages, use overt affixal morphology to mark the possessor directly on the possessed noun.

The simplest type of marking on the possessed noun marks no characteristics of the possessor, just the fact that the noun is possessed. The morphemes that mark possessed nouns in this way have been termed 'anti-genitives' (Andersen 1991) or 'pertensives' (Dixon 2010:268). They occur in Nêlêmwa, Martuthunira, Wandala, Northeast Ambae (Aikhenvald and Dixon 2012:7), some Nilotic languages, Hausa, Wolof, and in Semitic languages as part of the 'construct state' (Creissels 2009).[40] The example in (37) from Aikhenvald and Dixon (2012:7) shows overt marking of the quality of being possessed in Northeast Ambae (Austronesian).

(37) gamali-ni Robert
club.house-PSSD Robert
'Robert's club house'

As shown in the example, the feature for marking a noun as possessed is PSSD.[41]

[40] An example of the construct state in Classical Arabic, with the possessed noun marked as the construct, is shown in (1) from Creissels (2009:73-74). The morphemic glosses in the examples have been adapted to the UniMorph Schema. For an extensive discussion of the construct state and its functions in Arabic, see Ryding (2005: ch. 8).

(1) a. Indefinite noun
daxal-a kalb-u-n
enter.PST-3.SG.MASC dog.SG-NOM-INDEF
'A dog came in.'

b. Definite noun
daxal-a l-kalb-u
enter.PST-3.SG.MASC DEF-dog.SG-NOM
'The dog came in.'

c. Noun in construct state (note the lack of overt definiteness marking)
daxal-a kalb-u l-malik-i
enter.PST-3.SG.MASC dog.SG-NOM DEF-king-GEN
'The dog of the king came in.'

d. Possessor-marked noun (see following discussion and (40))
daxal-a kalb-u-hu
enter.PST-3.SG.MASC dog.SG-NOM-PSS3SM
'His dog came in.'

In this example, particularly sentence (c), the possessed noun is in the construct state, in which the absence of otherwise obligatory definiteness marking signals that the noun is possessed.

[41] We follow Creissels (2009) in not classifying possessed marking as a noun case. However, departing from Creissels (2009), the label PSSD is adopted because it is specifically the quality of marking a noun as possessed that must be incorporated into the schema, not the formal property of being part of a construct-state-like construction for which

Huallaga Quechua marks possession through overt affixal morphology that distinguishes the possessor's person and clusivity, yielding distinctions between morphemes with the meaning *my*, *your*, *his/her/its*, *our (inclusive)* and *our (exclusive)* (Weber 1989:54-55).

(38) Possessive suffixes in Huallaga Quechua (Weber 1989:54-55)

a. *uma-ː* 'my head' (possessive marker is additional length on final vowel)

b. *uma-yki* 'your head'

c. *uma-n* 'his/her head'

d. *uma-nchi* 'our (INCL) heads'

e. *uma-ːkuna* 'our (EXCL) heads'[42]

Turkish marks possession in a similar way, distinguishing the possessor's person, number (for first and second person), and politeness, but not clusivity as in Quechua.

(39) Possessive suffixes in Turkish (Göksel and Kerslake 2005:66)

a. *ev-im* 'my house'

b. *ev-in* 'your house (familiar; INFM)'

c. *ev-iniz* 'your house (polite; FORM)'

d. *ev-i* 'his/her/their house'

e. *ev-imiz* 'our house'

f. *ev-iniz* 'your (pl.) house'

g. *ev-leri* 'their house(s)'

Like Turkish and Huallaga Quechua, Arabic distinguishes possessive suffixes by the person and number of the possessor, but adds dual to the number distinctions made among possessors. Arabic possessive suffixes are also distinguished by the gender of the possessor (Ryding 2005:301).

(40) Possessive suffixes in Arabic (Ryding 2005:301)

Person	*Gender*	*Singular*	*Dual*	*Plural*
1		-ii	—	-naa
2	M	-ka	-kumaa	-kum
2	F	-ki		-kunna
3	M	-hu ∼ -hi	-humaa ∼ -himaa	-hum ∼ -him
3	F	-haa		-hunna ∼ -hinna

Features that mark characteristics of the possessor are composed according to a template, which begins with PSS- to mark 'possessed,' followed by a single number to mark the person of the possessor, a single letter to mark the number of the possessor, and an indication of gender, clusivity, or politeness. If a language were to mark possession by, for example, another kind of gender other than masculine or feminine (e.g. neuter), features could easily be created using this template (e.g. in this case, PSS3SN for a third person singular neuter possessive with the meaning of English *its*).

In addition to features marking characteristics of the possessor, features indicating the type of possession itself are necessary. Some languages distinguish between alienable and inalienable possession. An example of this contrast is the difference in the type of possession involved in 'my

Creissel's label, CSTR, would be better suited.

[42]This form is not explicitly given in the source, but is formed via the description given for forming the first person plural exclusive.

house' vs. 'my back,' in which the first type of possession is possession of property whose ownership can change while the second indicates inherent ownership. In the Mande language Kpelle, this example is rendered as in (41) from Welmers (1973:279).

(41) a. Alienable possession

ŋa pɛrɛi
I house

'my house'

b. Inalienable possession

m-pôlu
1.SG-back

'my back'

The features necessary to encode the overt morphological marking of possession on the possessed noun are presented in Table 29.

Feature	*Label*
Alienable possession	ALN
Inalienable possession	NALN
Possessed	PSSD
Possession by 1.SG	PSS1S
Possession by 2.SG	PSS2S
Possession by 2.SG.MASC	PSS2SM
Possession by 2.SG.FEM	PSS2SF
Possession by 2.SG.INFM	PSS2SINFM
Possession by 2.SG.FORM	PSS2SFORM
Possession by 3.SG	PSS3S
Possession by 3.SG.MASC	PSS3SM
Possession by 3.SG.FEM	PSS3SF
Possession by 1.DU	PSS1D
Possession by 1.DU.INCL	PSS1DI
Possession by 1.DU.EXCL	PSS1DE
Possession by 2.DU	PSS2D
Possession by 2.DU.MASC	PSS2DM
Possession by 2.DU.FEM	PSS2DF
Possession by 3.DU	PSS3D
Possession by 3.DU.MASC	PSS3DM
Possession by 3.DU.FEM	PSS3DF
Possession by 1.PL	PSS1P
Possession by 1.PL.INCL	PSS1PI
Possession by 1.PL.EXCL	PSS1PE
Possession by 2.PL	PSS2P
Possession by 2.PL.MASC	PSS2PM
Possession by 2.PL.FEM	PSS2PF
Possession by 3.PL	PSS3P
Possession by 3.PL.MASC	PSS3PM

Possession by 3.PL.FEM	PSS3PF

Table 29: Features necessary for encoding characteristics of the possessor on a possessed noun

5.22 Switch-Reference

Switch-reference is a type of anaphoric linkage that disambiguates the reference of subjects and other NPs across clauses (Stirling 1993:1). Although switch-reference has a functional basis, disambiguating the reference of subjects and other NPs, it is a fully grammaticalized phenomenon and is used in the languages in which it occurs even when the reference of subjects or other NPs is already fully disambiguated by other means. The example in (42) from Usan, a Papuan language, cited in Stirling (1993:4) illustrates both the phenomenon of switch-reference and the fact that in languages that use it, it need not be functionally motivated in every context (i.e. it is a fully grammatical phenomenon). Since the second verb in each example already uses person marking to indicate an identical and different referent, respectively, from the first verb, the same subject (SS) and different subject (DS) switch-reference marking is redundant. Its presence, however, is obligatory because it is a fully grammaticalized part of the language, much as subject-verb agreement is obligatory in English even when the subject's number is clear.

(42) Switch-reference in Usan (Papuan; Stirling 1993:4, citing Haiman and Munro 1983:xi(3,4))

a. ye nam su-**ab** isomei
I tree cut-**SS** I.went.down
'I cut the tree and went down.'

b. ye nam su-**ine** isorei
I tree cut-**DS** it.went.down
'I cut the tree down' (i.e. 'I cut the tree and it went down'; JCS)

The switch-reference markers are in bold, as are their glosses, which indicate same subject (SS) and different subject (DS). Note that the difference between the verbal forms *isomei* and *isorei* is due to person (first and third, respectively) and would disambiguate the subjects of the second clause without switch-reference marking. Thus, switch-reference marking is not necessary here, but must be present because it is a part of Usan's grammar.

Switch-reference (SR) marking is concentrated in languages of North America (notably in the Southwest, Great Basin, and coastal Northern California), Australia, Papua New Guinea, and the Bantu languages of Africa (Stirling 1993:5). It also occurs in languages of South America, and switch-reference-like phenomena have been identified in the Northeast Caucasus (Nichols 1983).

A typical and basic distinction in SR systems is between same subject, SS, and different subject, DS. This type of system occurs in Usan (as in (42)), Imbabura Quichua (Cohen 2013:55-56), and Mojave (Yuman; Munro 1980), to name a few. (43) illustrates an example from Mojave. Note that unlike in Usan, switch-reference marking is the only thing that disambiguates the subjects of the two clauses.

(43) Switch-reference in Mojave (Yuman; Stirling 1993:3 citing Munro 1980:145(4); morpheme glosses as in original source)

a. nya-isvar-**k** iima-k
when-sing-**SS** dance-Tns
'When he$_i$ sang, he$_i$ danced.'

b. nya-isvar-**m** iima-k
when-sing-**DS** dance-Tns
'When he$_i$ sang, he$_j$ danced.'

This basic distinction can be enriched by adding a third underspecified value, an 'open reference' SR marker. "Nichols (1983:247 etc.) says that in a number of languages of the Northeast Caucasus, such as Chechen and Ingush, DS marking verbs have what she calls 'Open Reference', signalling indifference as to the referential relation between the two pivots rather than specified non-identity" (Stirling 1993:34). While for Chechen and Ingush, the DS marker is the marker that is actually open reference, in Lak and Dargwa (also Nakh-Daghestanian), the SS marker is the one that is open reference and the DS marker calls for strict non-identity.

While the basic distinction between SS and DS is common, larger systems exist. Some switch-reference systems, such as those in Imbabura Quichua (Cohen 2013:55-56), Chickasaw, Choctaw, and Hopi have separate sets of switch-reference markers for separate grammatical contexts (Stirling 1993:16). In these languages, no more than two sets of SR markers exist, given the description in Stirling (1993:16), and all use one set for adverbials (as well as complement and relative clause constructions in Choctaw and Chickasaw) and one set for another function, such as paratactic clause combinations (Choctaw and Chickasaw), relative clauses (Hopi), and subjunctive contexts (Imbabura Quichua). (44) shows Imbabura Quichua's two-set system.

(44) Switch-reference marking suffixes in Imbabura Quichua (Cohen 2013:55(9))

CONTEXT	SS	DS
Adverbial	*-shpa*	*-xpi*
Subjunctive	*-ngapax*	*-chun*

Other systems are larger by virtue of allowing coreference between subjects and NPs in other grammatical roles, such as direct and indirect object. For example, in Capanahua (Panoan), there are "six DS suffixes, two of which imply the identity of the subject of the [morphologically] marked clause with the object of the controlling clause, and one of which implies the identity of the object of the marked clause with the subject of the controlling clause" (Stirling 1993:25-26 citing Jacobsen 1967:257).[43] Another system that has expanded switch-reference due to indexing subjects with non-subject NPs is attested in Warlpiri (Pama-Nyungan). Warlpiri has 4 overt SR markers that relate subjects to subjects, subjects to direct objects, and subjects to indirect objects (Stirling 1993:25).

(45) Switch-reference morphemes in Warlpiri (Stirling 1993:25 citing Simpson 1983)

a. *karra* - Same subject marking
b. *kurra* - "The subject of the marked clause is coreferential with the object of the controlling clause"
c. *rlajinta* - Same subject marking, but "the event described by the controlling clause is an 'accidental' consequence of the event described by the marked clause"
d. *rlarni* - "The subject of the marked clause, if non-overt, is the same as the oblique dative argument of the controlling clause"

In addition, other SR marking systems combine subject (or other NP) disambiguation with other dimensions of meaning, such as simultaneity of actions, as in Kâte (Papuan), and aspect, as in Kashaya (Pomoan). In Kâte, SS and DS SR morphemes also mark simultaneous or sequential

[43] Stirling (1993) does not discuss the other three DS markers.

actions. There are four total morphemes, with the DS sequential morpheme being phonologically null (an unpronounced zero morph; Stirling 1993:31, 40).[44]

(46) SR morphemes in Kâte (Stirling 1993:31 citing Longacre 1983:187)
huk = SS simultaneous
ra = SS sequential
ha = DS simultaneous
∅ = DS sequential

There is a tendency for languages to associate SS with sequentiality and DS with simultaneity by default (Stirling 1993:44). This is explicitly the case in Tunebo (Chibchan), but in that language, "simultaneity and sequentiality [are] more basic than the SS/DS distinction" (ibid.). Examples in Stirling (1993:44) demonstrate that the SS and DS markers can be used with different and identical subjects, respectively, as long as the correct temporal relationship is maintained (simultaneous and sequential, respectively; ibid.).

In Kashaya, six pairs of suffixes mark SS-DS switch-reference (Stirling 1993:41 citing Oswalt 1983:269). Of these, "one pair indicates simultaneous or alternating action, one indicates that the eventuality of the marked clause sequentially precedes the eventuality of the controlling clause in the present or past, and a third indicates that the marked clause eventuality sequentially precedes the controlling clause eventuality in the future or conditional. The other three suffixes are normally taken to be past tense, but may be specified as future by co-occurring with the future tense suffix already mentioned" (ibid.).

This system can be interpreted using the system from Klein (1994: discussed in detail in §5.4, p. 13) if 'eventuality' is understood to mean TSit and if aspect is understood to be applicable between multiple situation times ('TSits'), not just between TSit and TT. This particular kind of aspect could be termed 'multiclausal aspect' to differentiate it from conventional verbal aspect (as in §5.4). The pair indicating "simultaneous or alternating action" could be modeled as in (47).[45]

(47) Kashaya "simultaneous or alternating action" multiclausal aspect
Marked clause (m) ——{——}——
Controlling clause (c) ——{——}——

This multiclausal aspect is most similar to imperfective in which TT is included within TSit. To indicate its relationship with simultaneous marking in SR systems in languages like Kâte, it will be termed 'simultaneous multiclausal aspect.'

Where "the marked clause sequentially precedes the eventuality of the controlling clause in the present or past," the relationships in (48) and (49) might obtain, depending on the tense of the controlling clause.

(48) When the controlling clause is present tense
Marked clause (m) —{——}————————————
Controlling clause (c) ————————————{–[—|—]–}———

(49) When the controlling clause is past tense
Marked clause (m) —{——}————————————
Controlling clause (c) ————————————{–[——]–}——|—

These multiclausal aspects, as well as those that are to follow, are similar to the perfect aspect, in which TSit precedes TT, except here, the TSit of the marked clause precedes the TSit of the con-

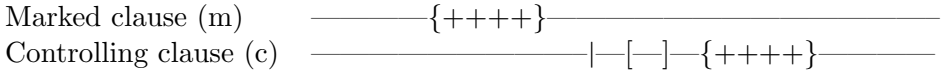
[44] Normally, if anything is phonologically null in an SR system, it is SS marking (Stirling 1993:30-31).

[45] Recall that braces { } symbolize the TSit span, brackets [ ] symbolize the TT span, and the pipe | symbolizes the Time of Utterance (TU).

trolling clause. This will be termed 'sequential multiclausal aspect,' again to indicate its connection with sequential marking in systems like that of Kâte.

Finally, when "the marked clause eventuality sequentially precedes the controlling clause eventuality in the future or conditional," the model in (50) might obtain.[46]

(50) When the controlling clause is in the future tense or conditional mood



As a final note, some West African languages have what have been called "logophoric" systems in which pronouns are explicitly coreferential with a pronoun in a previous clause. Here, the marking occurs on a noun rather than a verb, which is part of the reason these systems have not traditionally been considered switch-reference. This may be a trivial distinction, but other considerations favor treating logophoricity and switch-reference as separate phenomena. Languages that are logophoric often have syntactic and semantic restrictions on where logophoricity must be marked, e.g. in subordinate clauses which represent reported speech or are governed by the verb "say" (Stirling 1993:52-53). The obligatoriness of marking logophoricity also depends on person (53). In addition, there are unexpected co-reference patterns (ibid.). These can be seen in Gokana, which marks logophoricity on the verb and has thus been interpreted as a switch-reference language.

(51) Interpretation of logophoric clauses in Gokana

Lébɑ̀rè kɔ ɑè de-è ɑ gɪ́ɪ́ɑ́
Lebare said he ate-LOG he yams

1. 'Lebare$_i$ said he$_i$ ate his$_i$ yams.' (he himself ate his own yams)
2. 'Lebare$_i$ said he$_j$ ate his$_i$ yams.' (Lebare said someone else ate Lebare's yams)
3. 'Lebare$_i$ said he$_i$ ate his$_j$ yams.' (Lebare said he himself ate someone else's yams)
4. *'Lebare$_i$ said he$_j$ ate his$_j$ yams.' (*Lebare said someone else ate their own yams)

Because these interpretations (especially 2.) are not expected in the case of standard same subject SR marking, logophoricity should be marked separately from standard same subject SR marking.

To capture all these possible types of SR morphology, the features in Table 30 are necessary. To capture systems like those in Warlpiri and Capanahua, it is necessary to have a schematic morpheme that relates the role of the NP in the controlling clause to that of the NP in the marked clause. This takes the form ControllingClauseNP_Relation_MarkedClauseNP (abbreviated CN_R_MN), where the roles of the NPs would be marked using case relations, such that same subject in a nominative-accusative language would be CNOM_S_MNOM, where S in the R slot stands for 'same.' Different subject marking might be CNOM_D_MNOM.

Feature	*Label*
SS	SS
SS Adverbial	SSADV
DS	DS
DS Adverbial	DSADV

[46] This multiclausal rendering of the conditional may shed light on how to model conditionals as a whole. Note that the TT here is after the TSit of one clause but before that of another. In referring to the TT, then, it is possible to treat the preceding TSit as realis and in the past while that of the TSit of the controlling clause is yet to come in the future. The preceding TSit can be seen as the condition on which the action of the following TSit is predicated. A multiclausal rendering of the conditional may be necessary for accurately capturing its semantics in all contexts, even when only one clause is being analyzed.

Open Reference	OR
SR among NPs in any argument position	CN_R_MN
Simultaneous Multiclausal Aspect	SIMMA
Sequential Multiclausal Aspect	SEQMA
Logophoric	LOG

Table 30: Features necessary for capturing switch-reference morphology

5.23 Tense

Tense and aspect are defined according to the framework in Klein (1994, 1995), which builds on Reichenbach (1947) and uses the concepts of Time of Utterance (TU, '|'), Topic Time (TT, '[ ]'), and Situation Time (TSit, '{ }') to define tense and aspect categories. Topic Time (TT) and Situation Time (TSit) are conceived as spans while Time of Utterance (TU) is a single point. By defining tense and aspect categories solely in terms of the ordering of these spans and TU, tense and aspect categories can be defined independent of the language under analysis in a way that facilitates cross-linguistic comparison.

TU (symbolized with '|') is the time at which a speaker makes an utterance, and topic time (TT, symbolized by brackets [ ]) is the time about which the claim in the utterance is meant by the speaker to hold true. Situation time (TSit, symbolized with braces { }) is the time in which the state of affairs described by the speaker actually held true. Tense is the relationship of TU to TT, and aspect is the relationship of TT to TSit (for which, see §5.4, p. 13).

Another parameter that affects the definition of tense and aspect categories is whether the verb in question is 1-state or 2-state. A 1-state verb is a verb like 'sleep,' which lexically encodes only a single state (symbolized as '———'). From experience, speakers understand that the time period of that state is finite, but this is not lexically encoded, only pragmatically inferred (Klein 1995:682). In a 2-state verb, on the other hand, there is a source state (SS, symbolized as '———') and a target state (TS, symbolized as '++++++') and the verb lexically encodes those two states. The verb 'leave' is a 2-state verb, since it is impossible to leave without going through a transition of being somewhere (the source state) and then being gone from that place (the target state). The definition of tense and aspect categories can depend on these internal properties of the verb. The internal temporal properties of a verb below the level of the relationships between TT, TSit, and TU can be considerably more complex and these properties form the category of Aktionsart, discussed in §5.1, p. 8.

Tense is the relationship of the time of utterance (TU, '|') to the topic time (TT, '[ ]'). For example, in the sentence, "The book was lying on the table," the speaker is making a claim about a time period (TT) that occurred prior to the time of utterance (TU). Past tense indicates that the time for which the claim is meant to be true, TT, occurred before TU. Present tense indicates that the situation holds true during the time of utterance. Future tense indicates that at some point after the time of utterance, a situation will hold true. In the examples, TSit is not indicated, but can be assumed to hold true while the single state of 'to lie' holds true (indicated by '——').

(52) Past tense: TT precedes TU
.....——[————]—.....|.....
'The book was lying on the table.' ('to lie, be in a supine position' is a 1-state verb)

(53) Present tense: TU is within TT

.....—[——|——]—.....
'The book is lying on the table.'

(54) Future tense: TU precedes TT
.....|.....—[————]—.....
'The book will be lying on the table.'

Although past, present, and future are the core tense relationships that can be establishing by the relative positioning of TU and TT, languages use morphological marking to indicate the temporal distance between TU and TT, leading to distinctions like recent vs. remote past. A survey of such systems in Comrie (1985) reveals that while the most common split is a two-way split in the past tense between events that happened on the same day (hodiernal, from Latin *hodie* 'today'; 87) from those that did not, languages can make up to six such temporal distinctions.

Haya (Bantu) distinguishes hodiernal (today), hesternal (yesterday), and remote (before yesterday; 90). Kalaw Lagaw Ya (also called Western Torres Strait; classification disputed) makes these distinctions and adds another for events that happened the past night (96). Bamileke-Ngyemboon (Bantu) distinguishes four levels of temporal distance symmetrically in the past and future, such that for the past there is hodiernal, hesternal, recent past (in the last few days), and remote past while for the future there is later today, tomorrow, within the next few days (recent future), and farther ahead yet (remote future; 96). The related language Bamileke-Dschang (Bantu) also has a symmetrical system, but adds another step, an 'immediate' step indicating 'just now' or 'coming up in a moment' (97). This results in five distinctions in temporal distance both in the past and future. This points out the need to define distinctions these distinctions as distance between TU and TT, irrespective of the relationship between those two concepts.

Two other languages, Yandruwandha (Australia; 98) and Yagua (Peba-Yaguan; Peru; 99) have five level distinctions, but only in the past. The most elaborated system occurs in the Chinookan language Upper Chinookan (also called Wasco-Wishram, Wishram, Kiksht), which provides for at least six, possibly seven, distinctions in the past and two in the future (Comrie 1985:99-100). The distinction that is most notable is that between 'from 1-10 years ago' and 'remote past,' which is uncommon and is usually subsumed under the category of remote past.

1.	ga(l)...u-	remote past (remote past)
2.	ga(l)...t-	from 1-10 years ago (distant past?)
3.	ni(g)...u-	from a week to a year ago (recent past?)
4.	ni(g)...t-	last week (pre-hesternal?)
5.	na(l)-	yesterday or preceding couple of days (hesternal?)
6.	i(g)-	earlier today, with the possible refinements (hodiernal)
6a.	i(g)...u-	earlier on today, but not just now (hodiernal)
6b.	i(g)...t-	just now (immediate)

Table 31: The six to seven level past tense system of Upper Chinookan (Comrie 1985:99-100)

The dimensions necessary to encode the core tenses and these elaborations are listed in Table 32. Because the levels of temporal distance can be symmetrical, terms for these levels are meant to apply to both past and future, such that a hodiernal past would be indicated as PST+HOD and a future hodiernal would be FUT+HOD. These are preliminary categories since it may be better to encode some of the temporal distinctions using a feature-like analysis. For example, it may be possible to encode some of the finer distinctions in Upper Chinookan with combinations of temporal

distance features. Preliminarily, I propose that Upper Chinookan's 'last week' distinction could be PST+1DAY+RCT while 'from a week to a year ago' would be simply PST+RCT and similarly the 'from 1-10 years ago' level may be PST+RCT+RMT while the truly remote past would be simply PST+RMT.

Feature	*Label*
Present	PRS
Past	PST
Future	FUT
Immediate	IMMED
Hodiernal (today)	HOD
Within 1 day	1DAY
Recent	RCT
Remote	RMT

Table 32: Features for tense based on the relations described by Klein (1994) and the temporal distance distinctions exemplified in Comrie (1985)

5.24 Valency

Valency (or arity) refers to the number of arguments a verb can govern (i.e. select for). For example, a typical transitive verb takes two arguments, a subject and direct object, and therefore has a valency of 2, i.e. it is bivalent. Verbs that occur without any arguments, which in many languages include words for weather activity such as 'rain,' are often called *impersonal* verbs and have a valency of 0. Verbs that take a single argument, such as intransitive verbs, have a valency of 1. Ditransitive verbs, such as 'give,' take a subject, direct object, and indirect object, and therefore have a valency of 3.

The valency of a verb is often a lexical property, but both the valency and the relationship between arguments that are already present can be changed by specific morphology in many languages. Some altered valency configurations that can result are reflexive, reciprocal, causative, and applicative.

Reflexive morphemes indicate that the action performed by the subject is performed on itself (to a greater degree than might be expected from middle voice marking). Note the contrast between (55a) and (55b). Similarly, reciprocal morphemes indicate that with a plural subject, non-identical participants perform the action of the verb mutually on each other, as in (55c).

Causative morphemes add an additional participant (and therefore syntactic argument) and indicate that the additional participant was somehow forced to perform the action of the verb. For example, *Mark* is added as an additional participant in the causativized sentence in (55d).

(55) a. I washed the shirt.
b. *Reflexive*: I washed myself (i.e. I bathed (myself)).
c. *Reciprocal*: They washed each other.
d. *Causative*: I made Mark wash the shirt.

Applicative morphemes increase the number of oblique arguments (that is, arguments other than the subject or object) that are selected by the predicate (Polinsky 2013). For example, in Tukang Besi (Austronesian), "the verb 'fetch' takes one theme object in the basic construction

(as shown in 56a), but with the applicative marker it takes two objects, theme and benefactive" (Polinsky 2013 citing Donohue 1999:256).

(56) a. no-ala te kau
3.REALIS-fetch the wood
"She fetched the wood."

b. no-ala-ako te ina-su te kau
3.REALIS-fetch-APPL the mother-my the wood
"She fetched the wood (as a favor) for my mother" (similar to "She fetched my mother wood."; JCS).

Feature	*Label*
Impersonal	IMPRS
Intransitive	INTR
Transitive	TR
Ditransitive	DITR
Reflexive	REFL
Reciprocal	RECP
Causative	CAUS
Applicative	APPL

Table 33: Features necessary for indicating the valency of a verbal form, along with the valency they represent

5.25 Voice

Voice is the dimension of meaning that "expresses relations between a predicate [i.e. a verb] and a set of nominal positions - or their referents - in a clause or other structure" (Klaiman 1991: front matter). In the view of Fillmore (1968), "the function of voice marking, or overt verbally encoded manifestations of voice, is to signal the intactness or disruption of the basic relation(s) of a verb to its core nominal(s)" (Klaiman 1991:6). For example, the alternation between an active sentence *he broke the window* versus its passive equivalent *the window was broken* changes the subject of the sentence, a core nominal, from *he* to *the window* and omits the semantic agent *he* in the passive variant.

Klaiman (1991:2) defines three types of grammatical voice:

1. Derived voice, in which changes in the assignment of semantic roles to nouns or changes in their structural positions are used to signal a non-default, or marked, relationship between the predicate and each of the nominals (e.g. active/passive alternations);

2. Basic voice, which represents "a particular pattern of organization of a language's verbal lexicon" (e.g. in the lexicon of Fulani); and

3. Pragmatic voice, in which "alternations in verbal marking signal the variable assignment to sentential arguments of some special pragmatic status or salience" (31-32). Within pragmatic voice, the two main types are direct-inverse systems and so-called 'Austronesian voice.'

Derived voice includes two voice categories familiar from Indo-European languages, active and passive. For propositions marked with active voice, "the action notionally devolves from the standpoint of the most dynamic, or active, party involved in the situation, typically the Agent" (3).

Passive voice is used to mark "action which notionally devolves from the standpoint of a nondynamic, typically static participant in the situation, such as the Patient of a transitive verb" (ibid.). In the example alternation of *he broke the window* (active), *he* is the agent while in the passive variant, *the window was broken*, the action appears to originate from the grammatical subject, *the window*, even though it is semantically the patient. In ergative-absolutive languages, an ergative subject is demoted to an absolutive subject in what is termed an antipassive construction (230). Derived voice can also include middle voice in languages like Sanskrit, in which verbs can alternate in being marked for active, middle, or passive voice. However, middle voice is more often part of basic voice systems.

In Klaiman's terms, systems, rather than individual verbs, are considered to represent basic voice if the choice between active and middle voice does not reflect a rearrangement or change in structural or semantic roles, but rather a choice of lexical items (which may have still have overt voice-marking morphology). Modern Fula (Fulani) "has three voices [active, middle, passive], each associated with a distinct inflectional paradigm of the verb" and "about a fifth of the lexical verbs [...] can inflect in all three" (26). For many others, however, verbal lexical items have an inherent voice associated with them.

Before proceeding to describe pragmatic voice systems, the category of middle voice must be defined. The middle voice is used when "the viewpoint [of the predicate; JCS] is active in that the action notionally devolves from the standpoint of the most dynamic (or Agent-like) participant in the depicted situation. But the same participant has Patient-like characteristics as well, in that it sustains the action's principal effects" (3). The example cited for this is the Classical Greek sentence *loúomai khitô:na* 'I wash (MIDDLE) the shirt (for myself),' i.e. 'I am washing my shirt' (Lyons 1968:373).

Pragmatic voice systems include what have been called direct-inverse systems, which are common in North American languages, as well as complex voicing systems in Austronesian languages (so-called 'Austronesian voice'). In languages that possess direct-inverse systems, a salience hierarchy exists such that, as a hypothetical example, first person is higher (more "salient") than second person, which is higher than third person, and all these human pronouns are higher than any non-human animate nouns (or pronouns referring to them), and all these in turn are higher than inanimate nouns or pronouns, yielding the hierarchy: $1 > 2 > 3 >$ non-human animate $>$ inanimate. When the argument of the verb that is the most 'salient' in the sentence functions as the subject, the verb is either morphologically unmarked or is marked with a morpheme indicating direct voice (e.g. *-a:* in Plains Cree; Klaiman 1992:230). When the argument of the verb that is lower in the hierarchy functions as the subject, it is marked with a morpheme indicating inverse voice (e.g. *-iko* in Plains Cree; ibid.).

In general, pragmatic voice marking affects the prominence of nominals associated with specific semantic roles or with positions on a salience hierarchy. One striking example of this is the alignment system commonly found in Austronesian languages, particularly those of the Philippines, such as Tagalog, Cebuano, and Ilocano. In the pragmatic voice system of Cebuano, a different voice is used to focus nouns occupying four semantic roles, agent (A), goal (G), directional (D), and instrumental (I) (Klaiman 1991:247). Data from Cebuano (quoted in Klaiman 1991:247) illustrates this voice system for all the semantic roles except instrumental (I).

(57) a. Ni- hatag si Juan sa libro sa bata
A.VOICE give FOCUS Juan G book D child
"*Juan* gave the book to the child"

b. Gi- hatag ni Juan ang libro sa bata
G.VOICE give A Juan FOCUS book D child

"Juan gave *the book* to the child"

c. Gi- hatag -an ang bata ni Juan sa libro
D.VOICE give D.VOICE FOCUS child A Juan G book
"Juan gave the book *to the child*"

Here, a voice marker that is tied to the semantic role of the focused noun is used on the verb and the overt marker of the semantic role on the focused noun is replaced by a marker that indicates both its semantic role and its status as focused. The Austronesian language that makes the most distinctions in semantic role marking in its voice system is Iloko (Ilocano). The semantic roles it marks are given dedicated features in the UniMorph Schema since they are used by other Austronesian languages. Those roles are: Agent (AGFOC), patient (PFOC), location (LFOC), beneficiary (BFOC), accompanier (ACFOC), instrument (IFOC), and conveyed (CFOC; either by actual motion or in a linguistic sense, as by a speech act) (Rubino 2005:336-338).

The minimal features needed to encode voice morphology are given in Table 34.

Feature	*Label*
Active	ACT
Middle	MID
Passive	PASS
Antipassive	ANTIP
Direct	DIR
Inverse	INV
Agent Focus	AGFOC
Patient Focus	PFOC
Location Focus	LFOC
Beneficiary Focus	BFOC
Accompanier Focus	ACFOC
Instrument Focus	IFOC
Conveyed Focus	CFOC

Table 34: Grammatical voice features

6 Conclusion

The UniMorph Schema is intended to capture the full range of meaning that can be expressed by inflectional morphology across the world's languages. As instantiated in this document, the schema contains 23 dimensions of meaning and over 240 features. The UniMorph Schema provides a way to annotate very rich, fine-grained representations of the meaning encoded in fully inflected word forms, thereby allowing HLT applications to extract meaning from inflectional morphology across languages with strong confidence that the inflectional categories are semantically equivalent.

Through techniques such as projection (Yarowsky et al. 2001; Sylak-Glassman et al. 2015a; among many others), inflected words may receive a fully exhaustive representation, with values for all 23 dimensions (or a maximum number applicable for the given part of speech). However, sparser representations are appropriate for annotating a single language since some of the dimensions will not be expressed in each language (e.g. it is difficult to determine an evidentiality value for most verbs in English using only contextual evidence). While the linguistic principles which define the feature values may give the UniMorph Schema a steep learning curve for those without any

experience with linguistics, those with even a small amount of linguistics training should be able to translate other annotation schemas to the UniMorph Schema (e.g. the Penn Treebank tag VBZ to V;PRS;3;SG or native grammatical terminology such as 'non-past' to *non*{PST}).

The UniMorph Schema is designed to achieve extremely broad cross-linguistic coverage, and has proven useful in universalizing inflected forms in over 350 languages on Wiktionary (Sylak-Glassman et al. 2015b,a). However, it is a work-in-progress, and would benefit from user input, especially in case any other dimensions or features should be included.

7 Appendix 1: Full Alphabetical Listing of Dimensions and Features

The following table lists the dimensions of meaning, sorted alphabetically, along with their features, alphabetized on the feature labels.

Dimension	*Feature*	*Label*
Aktionsart	Accomplishment	ACCMP
Aktionsart	Achievement	ACH
Aktionsart	Activity	ACTY
Aktionsart	Atelic	ATEL
Aktionsart	Durative	DUR
Aktionsart	Dynamic	DYN
Aktionsart	Punctual	PCT
Aktionsart	Semelfactive	SEMEL
Aktionsart	Stative	STAT
Aktionsart	Telic	TEL
Animacy	Animate	ANIM
Animacy	Human	HUM
Animacy	Inanimate	INAN
Animacy	Non-human	NHUM
Argument Marking	3.SG Object (from feature template)	ARGAC3S
Aspect	Habitual	HAB
Aspect	Imperfective	IPFV
Aspect	Iterative	ITER
Aspect	Perfective	PFV
Aspect	Perfect	PRF
Aspect	Progressive	PROG
Aspect	Prospective	PROSP
Case	Ablative	ABL
Case	Absolutive	ABS
Case	Accusative	ACC
Case	Allative	ALL
Case	Near, in front of	ANTE
Case	Approximative	APPRX
Case	Next to	APUD
Case	At	AT
Case	Aversive	AVR
Case	Benefactive	BEN
Case	Essive-modal	BYWAY
Case	Near	CIRC
Case	Comitative	COM
Case	Comparative	COMPV
Case	Dative	DAT
Case	Equative	EQTV
Case	Ergative	ERG
Case	Essive	ESS
Case	Formal	FRML

Case	Genitive	GEN
Case	In	IN
Case	Instrumental	INS
Case	Among	INTER
Case	Nominative	NOM
Case	Nominative, S-only	NOMS
Case	On	ON
Case	On (horizontal)	ONHR
Case	On (vertical)	ONVR
Case	Behind	POST
Case	Privative	PRIV
Case	Prolative/translative	PROL
Case	Proprietive	PROPR
Case	Proximate	PROX
Case	Purposive	PRP
Case	Partitive	PRT
Case	Relative	REL
Case	Distal	REM
Case	Under	SUB
Case	Terminative	TERM
Case	Translative	TRANS
Case	Versative	VERS
Case	Vocative	VOC
Comparison	Absolute	AB
Comparison	Comparative	CMPR
Comparison	Equative	EQT
Comparison	Relative	RL
Comparison	Superlative	SPRL
Definiteness	Definite	DEF
Definiteness	Indefinite	INDF
Definiteness	Non-Specific	NSPEC
Definiteness	Specific	SPEC
Deixis	Above	ABV
Deixis	Below	BEL
Deixis	Even	EVEN
Deixis	Medial	MED
Deixis	No Reference Point, Distal	NOREF
Deixis	Invisible	NVIS
Deixis	Phoric, situated in discourse	PHOR
Deixis	Proximate	PROX
Deixis	First Person Reference Point	REF1
Deixis	Second Person Reference Point	REF2
Deixis	Remote	REMT
Deixis	Visible	VIS
Evidentiality	Assumed	ASSUM
Evidentiality	Auditory	AUD
Evidentiality	Direct	DRCT

Evidentiality	Firsthand	FH
Evidentiality	Hearsay	HRSY
Evidentiality	Inferred	INFER
Evidentiality	Non-firsthand	NFH
Evidentiality	Non-visual sensory	NVSEN
Evidentiality	Quotative	QUOT
Evidentiality	Reported	RPRT
Evidentiality	Sensory	SEN
Finiteness	Finite	FIN
Finiteness	Nonfinite	NFIN
Gender	Bantu Noun Classes	BANTU1-23
Gender	Feminine	FEM
Gender	Masculine	MASC
Gender	Nakh-Daghestanian Noun Classes	NAKH1-8
Gender	Neuter	NEUT
Information Structure	Focus	FOC
Information Structure	Topic	TOP
Interrogativity	Declarative	DECL
Interrogativity	Interrogative	INT
Language-Specific Features	varies by language	LGSPEC1
Language-Specific Features	varies by language	LGSPEC2
Mood	Admirative	ADM
Mood	Australian Non-Purposive	AUNPRP
Mood	Australian Purposive	AUPRP
Mood	Conditional	COND
Mood	Debitive	DEB
Mood	Deductive	DED
Mood	Imperative-Jussive	IMP
Mood	Indicative	IND
Mood	Intentive	INTEN
Mood	Irrealis	IRR
Mood	Likely	LKLY
Mood	Obligative	OBLIG
Mood	Optative-Desiderative	OPT
Mood	Permissive	PERM
Mood	Potential	POT
Mood	General Purposive	PURP
Mood	Realis	REAL
Mood	Subjunctive	SBJV
Mood	Simulative	SIM
Number	Dual	DU
Number	Greater paucal	GPAUC
Number	Greater plural	GRPL
Number	Inverse	INVN
Number	Paucal	PAUC
Number	Plural	PL
Number	Singular	SG

Number	Trial	TRI
Part of Speech	Adjective	ADJ
Part of Speech	Adposition	ADP
Part of Speech	Adverb	ADV
Part of Speech	Article	ART
Part of Speech	Auxiliary	AUX
Part of Speech	Classifier	CLF
Part of Speech	Complementizer	COMP
Part of Speech	Conjunction	CONJ
Part of Speech	Determiner	DET
Part of Speech	Interjection	INTJ
Part of Speech	Noun	N
Part of Speech	Numeral	NUM
Part of Speech	Particle	PART
Part of Speech	Pronoun	PRO
Part of Speech	Proper Name	PROPN
Part of Speech	Verb	V
Part of Speech	Converb	V.CVB
Part of Speech	Masdar	V.MSDR
Part of Speech	Participle	V.PTCP
Person	Zero person	0
Person	First person	1
Person	Second person	2
Person	Third person	3
Person	Fourth person	4
Person	Exclusive	EXCL
Person	Inclusive	INCL
Person	Obviative	OBV
Person	Proximate	PRX
Polarity	Positive	POS
Polarity	Negative	NEG
Politeness	Avoidance style	AVOID
Politeness	Colloquial	COL
Politeness	Formal, Referent Elevating	ELEV
Politeness	Formal register	FOREG
Politeness	Formal	FORM
Politeness	High status	HIGH
Politeness	Formal, Speaker Humbling	HUMB
Politeness	Informal	INFM
Politeness	Literary	LIT
Politeness	Low status	LOW
Politeness	Polite	POL
Politeness	High status, elevated	STELEV
Politeness	High status, supreme	STSUPR
Possession	Alienable possession	ALN
Possession	Inalienable possession	NALN
Possession	Possession by 1.DU	PSS1D

Possession	Possession by 1.DU.EXCL	PSS1DE
Possession	Possession by 1.DU.INCL	PSS1DI
Possession	Possession by 1.PL	PSS1P
Possession	Possession by 1.PL.EXCL	PSS1PE
Possession	Possession by 1.PL.INCL	PSS1PI
Possession	Possession by 1.SG	PSS1S
Possession	Possession by 2.DU	PSS2D
Possession	Possession by 2.DU.FEM	PSS2DF
Possession	Possession by 2.DU.MASC	PSS2DM
Possession	Possession by 2.PL	PSS2P
Possession	Possession by 2.PL.FEM	PSS2PF
Possession	Possession by 2.PL.MASC	PSS2PM
Possession	Possession by 2.SG	PSS2S
Possession	Possession by 2.SG.FEM	PSS2SF
Possession	Possession by 2.SG.FORM	PSS2SFORM
Possession	Possession by 2.SG.INFM	PSS2SINFM
Possession	Possession by 2.SG.MASC	PSS2SM
Possession	Possession by 3.DU	PSS3D
Possession	Possession by 3.DU.FEM	PSS3DF
Possession	Possession by 3.DU.MASC	PSS3DM
Possession	Possession by 3.PL	PSS3P
Possession	Possession by 3.PL.FEM	PSS3PF
Possession	Possession by 3.PL.MASC	PSS3PM
Possession	Possession by 3.SG	PSS3S
Possession	Possession by 3.SG.FEM	PSS3SF
Possession	Possession by 3.SG.MASC	PSS3SM
Possession	Possessed	PSSD
Switch-Reference	SR among NPs in any argument position	CN_R_MN
Switch-Reference	DS	DS
Switch-Reference	DS Adverbial	DSADV
Switch-Reference	Logophoric	LOG
Switch-Reference	Open Reference	OR
Switch-Reference	Sequential Multiclausal Aspect	SEQMA
Switch-Reference	Simultaneous Multiclausal Aspect	SIMMA
Switch-Reference	SS	SS
Switch-Reference	SS Adverbial	SSADV
Tense	Within 1 day	1DAY
Tense	Future	FUT
Tense	Hodiernal	HOD
Tense	Immediate	IMMED
Tense	Present	PRS
Tense	Past	PST
Tense	Recent	RCT
Tense	Remote	RMT
Valency	Applicative	APPL
Valency	Causative	CAUS
Valency	Ditransitive	DITR

Valency	Impersonal	IMPRS
Valency	Intransitive	INTR
Valency	Reciprocal	RECP
Valency	Reflexive	REFL
Valency	Transitive	TR
Voice	Accompanier Focus	ACFOC
Voice	Active	ACT
Voice	Agent Focus	AGFOC
Voice	Antipassive	ANTIP
Voice	Beneficiary Focus	BFOC
Voice	Conveyed Focus	CFOC
Voice	Direct	DIR
Voice	Instrument Focus	IFOC
Voice	Inverse	INV
Voice	Location Focus	LFOC
Voice	Middle	MID
Voice	Passive	PASS
Voice	Patient Focus	PFOC

Table 35: Dimensions of meaning presented alphabetically with their features sorted alphabetically by feature label

8 Appendix 2: Full Alphabetical Listing of Features and Dimensions

This appendix contains feature labels sorted alphabetically, with a short gloss of the feature and the dimension in which it belongs.

Label	*Feature*	*Dimension*
0	Zero person	Person
1	First person	Person
2	Second person	Person
3	Third person	Person
4	Fourth person	Person
1DAY	Within 1 day	Tense
AB	Absolute	Comparison
ABL	Ablative	Case
ABS	Absolutive	Case
ABV	Above	Deixis
ACC	Accusative	Case
ACCMP	Accomplishment	Aktionsart
ACFOC	Accompanier Focus	Voice
ACH	Achievement	Aktionsart
ACT	Active	Voice
ACTY	Activity	Aktionsart
ADJ	Adjective	Part of Speech
ADM	Admirative	Mood
ADP	Adposition	Part of Speech
ADV	Adverb	Part of Speech
AGFOC	Agent Focus	Voice
ALL	Allative	Case
ALN	Alienable Possession	Possession
ANIM	Animate	Animacy
ANTE	Near, in front of	Case
ANTIP	Antipassive	Voice
APPL	Applicative	Valency
APPRX	Approximative	Case
APUD	Next to	Case
ARGAC3S	3.SG Object (from feature template)	Argument Marking
ART	Article	Part of Speech
ASSUM	Assumed	Evidentiality
AT	At	Case
ATEL	Atelic	Aktionsart
AUD	Auditory	Evidentiality
AUNPRP	Australian Non-Purposive	Mood
AUPRP	Australian Purposive	Mood
AUX	Auxiliary	Part of Speech
AVOID	Avoidance style	Politeness
AVR	Aversive	Case
BANTU1-23	Bantu Noun Classes	Gender

BEL	Below	Deixis
BEN	Benefactive	Case
BFOC	Beneficiary Focus	Voice
BYWAY	Essive-modal	Case
CAUS	Causative	Valency
CFOC	Conveyed Focus	Voice
CIRC	Near	Case
CLF	Classifier	Part of Speech
CMPR	Comparative	Comparison
CN_R_MN	SR among NPs in any argument position	Switch-Reference
COL	Colloquial	Politeness
COM	Comitative	Case
COMP	Complementizer	Part of Speech
COMPV	Comparative	Case
COND	Conditional	Mood
CONJ	Conjunction	Part of Speech
DAT	Dative	Case
DEB	Debitive	Mood
DECL	Declarative	Interrogativity
DED	Deductive	Mood
DEF	Definite	Definiteness
DET	Determiner	Part of Speech
DIR	Direct	Voice
DITR	Ditransitive	Valency
DRCT	Direct	Evidentiality
DS	DS	Switch-Reference
DSADV	DS Adverbial	Switch-Reference
DU	Dual	Number
DUR	Durative	Aktionsart
DYN	Dynamic	Aktionsart
ELEV	Formal, Referent Elevating	Politeness
EQT	Equative	Comparison
EQTV	Equative	Case
ERG	Ergative	Case
ESS	Essive	Case
EVEN	Even	Deixis
EXCL	Exclusive	Person
FEM	Feminine	Gender
FH	Firsthand	Evidentiality
FIN	Finite	Finiteness
FOC	Focus	Information Structure
FOREG	Formal register	Politeness
FORM	Formal	Politeness
FRML	Formal	Case
FUT	Future	Tense
GEN	Genitive	Case
GPAUC	Greater paucal	Number

GRPL	Greater plural	Number
HAB	Habitual	Aspect
HIGH	High status	Politeness
HOD	Hodiernal	Tense
HRSY	Hearsay	Evidentiality
HUM	Human	Animacy
HUMB	Formal, Speaker Humbling	Politeness
IFOC	Instrument Focus	Voice
IMMED	Immediate	Tense
IMP	Imperative-Jussive	Mood
IMPRS	Impersonal	Valency
IN	In	Case
INAN	Inanimate	Animacy
INCL	Inclusive	Person
IND	Indicative	Mood
INDF	Indefinite	Definiteness
INFER	Inferred	Evidentiality
INFM	Informal	Politeness
INS	Instrumental	Case
INT	Interrogative	Interrogativity
INTEN	Intentive	Mood
INTER	Among	Case
INTJ	Interjection	Part of Speech
INTR	Intransitive	Valency
INV	Inverse	Voice
INVN	Inverse	Number
IPFV	Imperfective	Aspect
IRR	Irrealis	Mood
ITER	Iterative	Aspect
LFOC	Location Focus	Voice
LGSPEC1	Varies by language	Language-Specific Features
LGSPEC2	Varies by language	Language-Specific Features
LIT	Literary	Politeness
LKLY	Likely	Mood
LOG	Logophoric	Switch-Reference
LOW	Low status	Politeness
MASC	Masculine	Gender
MED	Medial	Deixis
MID	Middle	Voice
N	Noun	Part of Speech
NAKH1-8	Nakh-Daghestanian Noun Classes	Gender
NALN	Inalienable Possession	Possession
NEG	Negative	Polarity
NEUT	Neuter	Gender
NFH	Non-firsthand	Evidentiality
NFIN	Nonfinite	Finiteness
NHUM	Non-human	Animacy

NOM	Nominative	Case
NOMS	Nominative, S-only	Case
NOREF	No Reference Point, Distal	Deixis
NSPEC	Non-Specific	Definiteness
NUM	Numeral	Part of Speech
NVIS	Invisible	Deixis
NVSEN	Non-visual sensory	Evidentiality
OBLIG	Obligative	Mood
OBV	Obviative	Person
ON	On	Case
ONHR	On (horizontal)	Case
ONVR	On (vertical)	Case
OPT	Optative-Desiderative	Mood
OR	Open Reference	Switch-Reference
PART	Particle	Part of Speech
PASS	Passive	Voice
PAUC	Paucal	Number
PCT	Punctual	Aktionsart
PERM	Permissive	Mood
PFOC	Patient Focus	Voice
PFV	Perfective	Aspect
PHOR	Phoric, situated in discourse	Deixis
PL	Plural	Number
POL	Polite	Politeness
POS	Positive	Person
POS	Positive	Polarity
POST	Behind	Case
POT	Potential	Mood
PRF	Perfect	Aspect
PRIV	Privative	Case
PRO	Pronoun	Part of Speech
PROG	Progressive	Aspect
PROL	Prolative/translative	Case
PROPN	Proper Name	Part of Speech
PROPR	Proprietive	Case
PROSP	Prospective	Aspect
PROX	Proximate	Case
PROX	Proximate	Deixis
PRP	Purposive	Case
PRS	Present	Tense
PRT	Partitive	Case
PRX	Proximate	Person
PSS1D	Possession by 1.DU	Possession
PSS1DE	Possession by 1.DU.EXCL	Possession
PSS1DI	Possession by 1.DU.INCL	Possession
PSS1P	Possession by 1.PL	Possession
PSS1PE	Possession by 1.PL.EXCL	Possession

PSS1PI	Possession by 1.PL.INCL	Possession
PSS1S	Possession by 1.SG	Possession
PSS2D	Possession by 2.DU	Possession
PSS2DF	Possession by 2.DU.FEM	Possession
PSS2DM	Possession by 2.DU.MASC	Possession
PSS2P	Possession by 2.PL	Possession
PSS2PF	Possession by 2.PL.FEM	Possession
PSS2PM	Possession by 2.PL.MASC	Possession
PSS2S	Possession by 2.SG	Possession
PSS2SF	Possession by 2.SG.FEM	Possession
PSS2SFORM	Possession by 2.SG.FORM	Possession
PSS2SINFM	Possession by 2.SG.INFM	Possession
PSS2SM	Possession by 2.SG.MASC	Possession
PSS3D	Possession by 3.DU	Possession
PSS3DF	Possession by 3.DU.FEM	Possession
PSS3DM	Possession by 3.DU.MASC	Possession
PSS3P	Possession by 3.PL	Possession
PSS3PF	Possession by 3.PL.FEM	Possession
PSS3PM	Possession by 3.PL.MASC	Possession
PSS3S	Possession by 3.SG	Possession
PSS3SF	Possession by 3.SG.FEM	Possession
PSS3SM	Possession by 3.SG.MASC	Possession
PSSD	Possessed	Possession
PST	Past	Tense
PURP	General Purposive	Mood
QUOT	Quotative	Evidentiality
RCT	Recent	Tense
REAL	Realis	Mood
RECP	Reciprocal	Valency
REF1	First Person Reference Point	Deixis
REF2	Second Person Reference Point	Deixis
REFL	Reflexive	Valency
REL	Relative	Case
REM	Distal	Case
REMT	Remote	Deixis
RL	Relative	Comparison
RMT	Remote	Tense
RPRT	Reported	Evidentiality
SBJV	Subjunctive	Mood
SEMEL	Semelfactive	Aktionsart
SEN	Sensory	Evidentiality
SEQMA	Sequential Multiclausal Aspect	Switch-Reference
SG	Singular	Number
SIM	Simulative	Mood
SIMMA	Simultaneous Multiclausal Aspect	Switch-Reference
SPEC	Specific	Definiteness
SPRL	Superlative	Comparison

SS	SS	Switch-Reference
SSADV	SS Adverbial	Switch-Reference
STAT	Stative	Aktionsart
STELEV	High status, elevated	Politeness
STSUPR	High status, supreme	Politeness
SUB	Under	Case
TEL	Telic	Aktionsart
TERM	Terminative	Case
TOP	Topic	Information Structure
TR	Transitive	Valency
TRANS	Translative	Case
TRI	Trial	Number
V	Verb	Part of Speech
V.CVB	Converb	Part of Speech
V.MSDR	Masdar	Part of Speech
V.PTCP	Participle	Part of Speech
VERS	Versative	Case
VIS	Visible	Deixis
VOC	Vocative	Case

Table 36: Features sorted alphabetically by their label, with their short gloss and dimension of meaning indicated

9 References

Agrell, Sigurd. 1908. *Aspektänderung und Aktionsartbildung beim polnischen Zeitworte. Ein Beitrag zum Studium der indogennanischen Pratverbia und ihrer Bedeutungsfunktionen.* Lund, Sweden: Ohlsson.

Ahlberg, Malin; Markus Forsberg; and Mans Hulden. 2014. Semi-supervised learning of morphological paradigms and lexicons. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden: Association for Computational Linguistics, 569–578.

Ahlberg, Malin; Markus Forsberg; and Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, Denver, CO: Association for Computational Linguistics, 1024–1029.

Aikhenvald, Alexandra Y. 2004. *Evidentiality.* Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. and Robert M. W. Dixon. 2012. *Possession and Ownership.* Oxford: Oxford University Press.

Aissen, Judith. 1997. On the syntax of obviation. *Language* 73(4):705–750.

Allan, Keith. 1977. Classifiers. *Language* 53(2):285–311.

Andersen, Torben. 1991. Subject and topic in Dinka. *Studies in Language* 15(2):265–294.

Anderson, John M. 2004. On the grammatical status of names. *Language* 80(3):435–474.

Baker, Mark. 2003. *Lexical Categories.* Cambridge, UK: Cambridge University Press.

Beck, David. 2000. Grammatical convergence and the genesis of diversity in the Northwest Coast Sprachbund. *Anthropological Linguistics* 42(2):1–67.

Bhat, D. N. Shankara. 2004. *Pronouns.* Oxford: Oxford University Press.

Bickel, Balthasar and Johanna Nichols. 2005. Inclusive-exclusive as person vs. number categories worldwide. *Clusivity*, edited by Elena Filimonova, Philadelphia: John Benjamins, 49–72.

Bickel, Balthasar and Johanna Nichols. 2009. Case marking and alignment. *The Oxford Handbook of Case*, edited by Andrei L. Malchukov and Andrew Spencer, Oxford: Oxford University Press, 304–321.

Blake, Barry J. 2001. *Case.* Cambridge, UK: Cambridge University Press, 2nd edn.

Bliss, Heather and Elizabeth Ritter. 2001. Developing a database of personal and demonstrative pronoun paradigms: Conceptual and technical challenges. *Proceedings of the ICRS Workshop on Linguistic Databases*, edited by Steven Bird; Peter Buneman; and Mark Lieberman, Philadelphia: Institute for Research in Cognitive Science.

Bolinger, Dwight L. 1968. Postposed main phrases: An English rule for the Romance subjunctive. *Canadian Journal of Linguistics* 14:3–33.

Breen, J. G. 1976. Ergative, locative, and instrumental inflections in Wangkumara. *Grammatical Categories in Australian Languages*, edited by Robert M. W. Dixon, Canberra: Australian Institute of Aboriginal Studies, 336–339.

Brown, Penelope and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics, Cambridge, UK: Cambridge University Press.

Cable, Seth. 2008. Tense, aspect and Aktionsart. Unpublished handout from "Proseminar on Semantic Theory" for *Theoretical Perspectives on Languages of the Pacific Northwest*. Available at: http://people.umass.edu/scable/PNWSeminar/handouts/Tense/Tense-Background.pdf.

Chelliah, Shobhana L. and Willem J. de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Dordrecht, Netherlands: Springer.

Chirikba, Viacheslav. 2003. *Abkhaz*. Munich: Lincom Europa.

Choi, Jinho; Marie-Catherine de Marneffe; Tim Dozat; Filip Ginter; Yoav Goldberg; Jan Hajič; Christopher Manning; Ryan McDonald; Joakim Nivre; Slav Petrov; Sampo Pyysalo; Natalia Silveira; Reut Tsarfaty; and Dan Zeman. 2015. Universal Dependencies. Accessible at: http://universaldependencies.github.io/docs/.

Cohen, Clara. 2013. Hierarchies, subjects, and the lack thereof in Imbabura Quichua subordinate clauses. *Structure and Contact in Languages of the Americas*, vol. 15, edited by John Sylak-Glassman and Justin Spence, Berkeley, CA: Survey of California and Other Indian Languages, 51–68.

Colarusso, John. 1992. *A Grammar of the Kabardian Language*. University of Calgary Press.

Cole, Peter. 1982. *Imbabura Quechua*, *Lingua Descriptive Studies*, vol. 5. Amsterdam: North-Holland Publishing Company.

Comrie, Bernard. 1976a. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge, UK: Cambridge University Press.

Comrie, Bernard. 1976b. Linguistic politeness axes: Speaker-addressee, speaker-referent, speaker-bystander. *Pragmatics Microfiche* 1.7(A3). Department of Linguistics, University of Cambridge.

Comrie, Bernard. 1985. *Tense*. Cambridge, UK: Cambridge University Press.

Comrie, Bernard. 1989. *Language Universals and Linguistic Typology*. Oxford: Basil Blackwell, 2nd edn.

Comrie, Bernard and Maria Polinsky. 1998. The great Daghestanian case hoax. *Case, Typology, and Grammar: In Honor of Barry J. Blake*, edited by Anna Siewierska and Jae Jung Song, Amsterdam: John Benjamins, 95–114.

Comrie, Bernard; Martin Haspelmath; and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Http://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Corbett, Greville G. 1981. Syntactic features. *Journal of Linguistics* 17:55–76.

Corbett, Greville G. 1991. *Gender*. Cambridge, UK: Cambridge University Press.

Corbett, Greville G. 2000. *Number*. Cambridge, UK: Cambridge University Press.

Corbett, Greville G. 2012. Politeness as a feature: So important and so rare. *Linguistik Online* 51(1/2):9–27.

Cowper, Elizabeth. 2002. Finiteness. Ms. University of Toronto. Available at: http://www.chass.utoronto.ca/∼cowper/Cowper.finiteness.pdf.

Creissels, Denis. 2009. Construct forms of nouns in African languages. *Proceedings of the Conference on Language Documentation and Linguistic Theory 2*, edited by Peter K. Austin; Oliver Bond; Monik Charette; David Nathan; and Peter Sells, London: School of Oriental and African Studies (SOAS), 73–82. Available at: http://www.hrelp.org/publications/ldlt2/papers/ldlt2_08.pdf.

Croft, William. 2000. Parts of speech as language universals and as language-particular categories. *Approaches to the Typology of Word Classes*, edited by Petra M. Vogel and Bernard Comrie, New York: Mouton de Gruyter, 65–102.

Cuzzolin, Pierluigi and Christian Lehmann. 2004. Comparison and gradation. *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung / An International Handbook on Inflection and Word-Formation*, vol. 2, edited by Geert Booij; Christian Lehmann; Joachim Mugdan; and Stavros Skopeteas, Berlin: Mouton de Gruyter, 1212–1220.

Daniel, Michael. 2005. Understanding inclusives. *Clusivity*, edited by Elena Filimonova, Philadelphia: John Benjamins, 3–48.

Davies, William D. 1986. *Choctaw Verb Agreement and Universal Grammar*. Dordrecht, Netherlands: D. Reidel Publishing Company.

Davis, Irvine. 1964. The language of Santa Ana Pueblo (anthropological papers, no. 69). *Smithsonian Institution Bureau of American Ethnology, Bulletin 191: Anthropological Papers, Numbers 68-74*, Washington, DC: United States Government Printing Office, 53–190.

Demuth, Katherine. 2000. Bantu noun classes: Loanword and acquisition evidence of semantic productivity. *Classification Systems*, edited by G. Senft, Cambridge, UK: Cambridge University Press, 270–292.

Dixon, Robert M. W. 1980. *The Languages of Australia*. Cambridge, UK: Cambridge University Press.

Dixon, Robert M. W. 2010. *Basic Linguistic Theory, Volume 2: Grammatical topics*. Oxford: Oxford University Press.

Dobrovie-Sorin, Carmen and Ion Giurgea. 2013. *A Reference Grammar of Romanian: Volume 1: The Noun Phrase*. Philadelphia: John Benjamins.

Donohue, Mark. 1999. *A Grammar of Tukang Besi*. Berlin: Mouton de Gruyter.

Dreyer, Markus and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. *Proceedings of EMNLP 2011*, Edinburgh: Association for Computational Linguistics, 616–627.

Durrett, Greg and John DeNero. 2013. Supervised learning of complete morphological paradigms. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta: Association for Computational Linguistics, 1185–1195.

Èdel'man, D. I. 1966. *Jazguljamskij jazyk*. Moscow: Nauka.

Fillmore, Charles J. 1968. The case for case. *Universals in Linguistic Theory*, edited by Emmon Bach and Robert T. Harms, New York: Holt, Rinehart, and Winston, 1–88.

Fillmore, Charles J. 1975. *Santa Cruz Lectures on Deixis*. Bloomington, IN: Indiana University Linguistics Club.

Foley, William A. and Robert D. Van Valin. 1985. Information packaging in the clause. *Language Typology and Syntactic Description*, vol. 1, edited by Timothy Shopen, Cambridge, UK: Cambridge University Press.

Friedman, Victor A. 2006. Lak. *Encyclopedia of Language and Linguistics*, vol. 6, Elsevier, 303–305.

Gair, James W. 2003. Sinhala. *The Indo-Aryan Languages*, edited by George Cardona and Dhanesh Jain, New York: Routledge, 766–817.

Göksel, Asli and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. New York: Routledge.

Grimm, Scott. 2012. Inverse number marking and individuation in Dagaare. *Count and Mass Across Languages*, edited by Diane Massam, Oxford: Oxford University Press, 75–98.

Haiman, John and Pamela Munro. 1983. Introduction. *Switch-Reference and Universal Grammar*, edited by John Haiman and Pamela Munro, Philadelphia: John Benjamins, Typological Studies in Language, ix–xv.

Hammarström, Harald and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics* 37(2):309–350.

Haspelmath, Martin. 1995. The converb as a cross-linguistically valid category. *Converbs in Cross-Linguistic Perspective: Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerunds –*, edited by Martin Haspelmath and Ekkehard König, Berlin: Mouton de Gruyter, Empirical Approaches to Language Typology, 1–56.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3):663–687.

Hooper, Joan Bybee. 1975. On assertive predicates. *Syntax and Semantics 4*, edited by P. Kimball, New York: Academic Press, 91–124.

Hualde, José Ignacio and Jon Ortiz de Urbina. 2003. *A Grammar of Basque*. The Hague: Mouton de Gruyter.

Jacobsen, William H. 1967. Switch-reference in Hokan-Coahuiltecan. *Studies in Southwestern Ethnolinguistics*, The Hague: Mouton, 238–263.

Keating, Elizabeth and Alessandro Duranti. 2006. Honorific resources for the construction of hierarchy in Samoan and Pohnpeian. *The Journal of the Polynesian Society* 115(2):145–172.

Kibort, Anna. 2010. A typology of grammatical features. Online overview by Surrey Morphology Group. Available at http://www.grammaticalfeatures.net/inventory.html.

Kibrik, Andrej A. 2012. What's in the head of head-marking languages? *Argument Structure and Grammatical Relations: A Cross-Linguistic Typology*, edited by Pirkko Suihkonen; Bernard Comrie; and Valery Solovyev, Amsterdam: John Benjamins, 211–240.

Klaiman, M. H. 1991. *Grammatical Voice*. Cambridge, UK: Cambridge University Press.

Klaiman, M. H. 1992. Inverse languages. *Lingua* 88:227–261.

Klein, F. 1975. Pragmatic constrains in distribution: The Spanish subjunctive. *Papers from the 11th Regional Meeting of the Chicago Linguistic Society*, edited by Robin E. Grossman; L. James San; and Timothy J. Vance, Chicago: Chicago Linguistic Society, 353–365.

Klein, Horst G. 1974. *Tempus, Aspekt, Aktionsart*. Tübingen: Max Niemeyer Verlag.

Klein, Wolfgang. 1994. *Time in Language*. New York: Routledge.

Klein, Wolfgang. 1995. A time-relational analysis of Russian aspect. *Language* 71(4):669–695.

Koptjevskaja-Tamm, Maria. 1993. *Nominalizations*. London: Routledge.

Laitinen, Lea. 2006. Zero person in Finnish: A grammatical resource for construing human reference. *Grammar from the Human Perspective: Case, Space and Person in Finnish*, edited by Marja-Liisa Helasvuo and Lyle Campbell, Amsterdam: John Benjamins, 209–232.

Lambrecht, Knud. 1994. *Information Structure and Sentence Form: Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge, UK: Cambridge University Press.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge, UK: Cambridge University Press.

Longacre, Ronald. 1983. Switch reference systems from two distinct linguistic areas: Wojokeso (Papua New Guinea) and Guanano (Northern South America). *Switch-Reference and Universal Grammar*, edited by John Haiman and Pamela Munro, Philadelphia: John Benjamins, 185–207.

Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.

Lyons, John. 1968. *Introduction to Theoretical Linguistics*. Cambridge, UK: Cambridge University Press.

Lyons, John. 1977. *Semantics*. Cambridge, UK: Cambridge University Press. 2 vols.

Magometov, Aleksandr Amarovič. 1970. *Agul'skij jazyk*. Tbilisi, Georgia: Mecniereba.

Munro, Pamela. 1980. On the syntactic status of switch-reference clauses: The special case of mojave comitatives. *Studies of Switch-Reference*, Los Angeles: University of California, Los Angeles, UCLA Papers in Syntax, 144–159.

Nichols, Johanna. 1983. Switch-reference in the Northeast Caucasus. *Switch-Reference and Universal Grammar*, edited by John Haiman and Pamela Munro, Philadelphia: John Benjamins, 245–265.

Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language* 62(1):56–119.

Nichols, Johanna. 2011. *Ingush Grammar*, *University of California Publications in Linguistics*, vol. 143. Berkeley, CA: University of California Press.

Nicolai, Garrett; Colin Cherry; and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, Denver, CO: Association for Computational Linguistics, 922–931.

Niinaga, Yuto. 2010. Yuwan (Amami Ryukyuan). *An Introduction to Ryukyuan Languages*, edited by Michinori Shimoji and Thomas Pellard, Tokyo: Research Institute for Languages and Cultures of Asia and Africa, 35–88.

Oswalt, Robert. 1983. Interclausal reference in Kashaya. *Switch-Reference and Universal Grammar*, edited by John Haiman and Pamela Munro, Philadelphia: John Benjamins, 267–290.

Palmer, Frank R. 2001. *Mood and Modality*. Cambridge, UK: Cambridge University Press, 2nd edn.

Payne, John R. 1981. Iranian languages. *The Languages of the Soviet Union*, edited by Bernard Comrie, Cambridge, UK: Cambridge University Press, 158–179.

Piper, Nick. 1989. *A Sketch Grammar of Meryam Mir*. Master's thesis, Australian National University, Canberra.

Polinsky, Maria. 2013. Applicative constructions. *The World Atlas of Language Structures Online*, edited by Matthew Dryer and Martin Haspelmath, Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at http://wals.info/chapter/109.

Radkevich, Nina V. 2010. *On Location: The Structure of Case and Adpositions*. Ph.D. thesis, University of Connecticut, Storrs, CT.

Reichenbach, Hans. 1947. *Elements of Symbolic Logic*. New York: Macmillan and Co.

Rubino, Carl. 2005. Iloko. *The Austronesian Languages of Asia and Madagascar*, London: Routledge, 326–349.

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge, UK: Cambridge University Press.

Sagot, Benoît and Géraldine Walther. 2013. Implementing a formal model of inflectional morphology. *Systems and Frameworks for Computational Morphology*, edited by Cerstin Mahlow and Michael Piotrowski, Berlin: Springer, 115–134.

Schulze, Wolfgang. 2003. The diachrony of demonstrative pronouns in East Caucasian. *Current Trends in Caucasian, East European and Inner Asian Linguistics: Papers in Honor of Howard I. Aronson*, edited by Dee Ann Holisky and Kevin Tuite, Philadelphia: John Benjamins, 291–348.

Searle, John R. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge, UK: Cambridge University Press.

Simpson, Jane H. 1983. *Aspects of Warlpiri Morphology and Syntax*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Spencer, Andrew. 2008. Does Hungarian have a case system? *Case and Grammatical Relations: Studies in Honor of Bernard Comrie*, edited by Greville G. Corbett and Michael Noonan, Philadelphia: John Benjamins, 35–56.

Stassen, Leon. 1984. The comparative compared. *Journal of Semantics* 3(1-2):143–182.

Stirling, Lesley. 1993. *Switch-Reference and Discourse Representation*. Cambridge Studies in Linguistics, Cambridge, UK: Cambridge University Press.

Sylak-Glassman, John; Christo Kirov; Matt Post; Roger Que; and David Yarowsky. 2015a. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. *Proceedings of the 4th Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, edited by Cerstin Mahlow and Michael Piotrowski, Berlin: Springer, Communications in Computer and Information Science, 72–93.

Sylak-Glassman, John; Christo Kirov; David Yarowsky; and Roger Que. 2015b. A language-independent feature schema for inflectional morphology. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Beijing: Association for Computational Linguistics, 674–680.

Tellings, Jos. 2014. 'Only' and focus in Imbabura Quichua. *Proceedings of the 40th Annual Meeting of the Berkeley Linguistics Society*, edited by Herman Leung; Zachary O'Hagan; Sarah Bakst; Auburn Lutzross; Jonathan Manker; Nicholas Rolle; and Katie Sardinha, Berkeley, CA: Berkeley Linguistics Society, 523–544.

Terrell, Tracy and Joan Bybee Hooper. 1974. A semantically based analysis of mood in Spanish. *Hispania* 57:484–494.

Tonhauser, Judith. 2007. Nominal tense? The meaning of Guaraní nominal temporal markers. *Language* 83(4):831–869.

Tsujimura, Natsuko. 2007. *An Introduction to Japanese Linguistics*. Oxford: Blackwell Publishing, 2nd edn.

Van de Velde, Mark. 2012. Agreement as a grammatical criterion for proper name status in Kirundi. *Onoma* 37:127–139.

van Driem, George. 1987. *A Grammar of Limbu*. Berlin: Walter de Gruyter.

van Langendonck, Willy. 2007. *Theory and Typology of Proper Names*. Berlin: Mouton de Gruyter.

Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66(2):143–160.

Weber, David J. 1989. *A Grammar of Huallaga (Huánuco) Quechua*, *University of California Publications in Linguistics*, vol. 112. Berkeley, CA: University of California Press.

Welmers, William E. 1973. *African Language Structures*. Berkeley, CA: University of California Press.

Wenger, James R. 1982. *Some Universals of Honorific Language with Special Reference to Japanese*. Ph.D. thesis, University of Arizona, Tucson, AZ.

Willie, MaryAnn. 1991. *Navajo Pronouns and Obviation*. Ph.D. thesis, University of Arizona, Tucson, AZ.

Woodbury, Anthony C. 1982. Switch reference, syntactic organization, and rhetorical structure in Central Yup'ik Eskimo. Tech. Rep. 98, Southwest Educational Development Laboratory, Austin, TX. Available at: https://archive.org/details/ERIC_ED252059.

Yamamoto, Mutsumi. 1999. *Animacy and Reference*. Amsterdam: John Benjamins.

Yarowsky, David; Grace Ngai; and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. *Proceedings of the First International Conference on Human Language Technology (HTL)*, Stroudsburg, PA: Association for Computational Linguistics, 1–8.