What to compare in historical linguistics?

Jinyuan Wu

April 19, 2025

1 The (lack of) synchronic foundation for diachronic studies

The Neogrammarian hypothesis states that language changes can be explained *completely* by (a) regular sound change without exceptions, (b) analogy, and (c) borrowing. We can then use the **comparative method** and **internal reconstruction** to identify cognates and layers of borrowed words.

So far we are just repeating words you can find on standard historical linguistics books. There however is a usually unspoken caveat: what is the unit that the comparative method runs on? A historical linguist will immediately answer "the word". But what is a word, then? And *why* don't we try to determine genetic relations based on syntactic patterns, but words, whatever the term means?

A choice in methodology eventually reflects a certain underlying assumption on how things work. Choosing to apply the comparative method and internal reconstruction to "the word" means that we believe that when a language is passed to the younger generation, what are actually passed are sequences with relatively stable internal structures, which we name *words*. Now, we have to be able to identify what *historical* wordhood means *synchronically*, or otherwise in theory we will be unable to gather enough materials for diachronic studies.

Thus historical linguistics should ideally have a synchronic, and ultimately psycholinguistic foundation. Ideally, the Neogrammarian hypothesis should be explained by acquisition of phonology, and its (alleged) breakdown in dialectical continua should be explained by e.g. the psycholinguistics of how two mutually intelligible languages are perceived in the brain. Methodological disputes in historical linguistics should eventually be resolved *experimentally*, by testing their implicit assumptions on how languages are transmitted from one generation to another. Given the current status of theoretical linguistics and psycholinguistics, however, we should not expect to see this in the foreseeable future.

Still what is a word in historical linguistics is too important to be left to future biologists who will literally peak into your brain to see how language works. It is fundamental to the everyday job of historical linguists.

2 How grammar works

Let's forget about history and focus on synchronic concepts for a while here. We first go over modern theories of syntax, and point out that syntactic structures provide no definite definition for wordhood. We then turn to the linearization of abstract syntax, as well as the structure of the lexicon, and define morphological wordhood. Finally, we turn to phonological wordhood, and emphasize that phonological wordhood may have subtle differences with morphological wordhood. We conclude by a list of candidates of the synchronic counterpart of diachronic wordhood.

2.1 Abstract syntax

2.1.1 Peeling off morphophonology and focusing on abstract syntax

If you are convinced by Distributed Morphology or theories along this line of thinking, you will know that it seems we cannot define wordhood in a completely intuitive way in *abstract* or *pure* syntax.[1] Let me explain.

Consider the example *the two ugly blackbirds*. Should we bracket the noun phrase as *the [two [ugly [blackbirds]]]*? Not necessarily. The category of plural number, marked by *-s*, seems to have a scope covering at least the nominal *two ugly blackbirds*. This can first be seen from semantic interpretation: *blackbird* is a compound that denotes a certain type of birds, and *ugly blackbird* is a conjunction of being ugly and being a blackbird. Now *two ugly blackbirds* specifies a set of two ugly blackbirds, and finally, *the two ugly blackbirds* reminds the listener to recall an aforementioned or at least identifiable set of two ugly blackbirds. If we assume that the clearly hierarchical semantics has a structural origin, then we should assume that the category of number is somehow higher than adjectival modification. This head noun-adjective-number-determiner hierarchy can be found cross-linguistically. In Japhug, for instance, the number marker follows coordinated head nouns and also the numeral, highlighting its scope over the whole noun phrase (Jacques 2021, p. 368, (2-3)).

This means we are probably to analyze *two ugly blackbirds* as something like *[the*$_{\text{D}}$ *[two [-s [[ugly]*$_{\text{AP}}$ *[blackbird]*$_{\text{N}}$*]*$_{\text{FP}}$*]*$_{\text{Num'}}$*]*$_{\text{NumP}}$*]*$_{\text{DP}}$.[2] Does the compound *blackbird* get isolated from the rest of syntax and hence have a special status (sometimes called *lexical integrity*) and can be seen as a word? Not necessarily. A noun phrase is also a small world in the eyes of the clause. This doesn't make a noun phrase a "word" in any proper sense. Furthermore, derived words are indeed subject to syntactic processes. (1) shows some attested examples.

(1) a. [pre- and post-revolutionary] France
b. back- and tooth ache (from Internet)

Thus *the two ungly blackbirds* are probably to be analyzed as something like (2). Here following the usual Distributed Morphology assumption that *backbird* is *categorized* into what we commonly know as a noun by virtue of referring to an abstract notion of a type of objects, etc., and we describe the "categorizer phrase" as nP.

(2) *[the*$_{\text{D}}$ *[two [-s [[ugly]*$_{\text{AP}}$ *[√black √bird]*$_{\text{nP}}$*]*$_{\text{FP}}$*]*$_{\text{Num'}}$*]*$_{\text{NumP}}$*]*$_{\text{DP}}$

2.1.2 Abstract as hierarchies of grammatical categories around lexical roots

There is nothing in abstract syntax, besides lexical roots and what are known as functional heads in generative syntax, more commonly called grammatical markers or *formatives* in descriptive linguistics. And the grammatical categories, together with "arguments" they introduce (including clausal arguments, adjective phrases, etc.) are wrapped around the core lexical root of a construction (like a noun phrase or a clause)

[1] Lexicalists will push back – but I believe they are wrong (Bruening 2018).

[2] We are describing the phrase structure using *functional heads*; see the end of this section, ans § 2.1.3. We are making the Cartographic assumption that adjectival modifications are also introduced by functional heads (FPs) and not an adjunction operation radically different from complementation. The motivation is to make the primitives of syntax more simple and flexible.

layer by layer, forming an endocentric structure. The endocentric structure of layered grammatical categories in noun phrases is shown in (2). In the clause, the structure is like vP-TP-CP,[3] or in descriptive terms, a hierarchy of grammatical categories in the hierarchy of argument structure < tense, aspect and modality < speech force categories or complement clause types.

The hierarchy of functional heads (i.e. grammatical relations and categories) as is exemplified in (2) has real effects. We have already seen its semantic effects in interpreting (2). In clauses, we find that the order of tense-aspect-modality adverbs and corresponding auxiliaries seem to have a regular correspondence, which can be explained by assuming that the TP or *tense phrase* actually splits into a series of functional projections, as is what is done in Cartographic syntax (Cinque and Rizzi 2009).

In the vP layer, we can find the influence of this layered structure as well: we have several syntactic tests to show that certain arguments (usually the agentative ones) are "higher" than others. Besides commonly known phenomena of binding of reflexive pronouns (*she hates herself*), we have a good example from causativization in Japhug. We note that Japhug allows double causative, and when this happens, the meaning is always like 'X makes [[Z do sth. to W] with Y]' (we refer it by $X \to Y \to Z \to W$), and the polypersonal direct-inverse indexation on the main verb (with the form $X \to Y$) is determined by first comparing the prominence of W and Z on the empathy hierarchy, and then comparing the prominence of the winner with that of Y, and then the prominence of the final winner is compared with X, the result of which determines if the inverse marker appears. Hence a 1→3→2→3 configuration is morphologically the same as 1→2 (Jacques 2021, p. 848, (67)). Similarly, both 2→3→1 and 2→1→3 are equivalent to 2→1 in argument indexation (Jacques 2021, p. 584), and both 3→3→1 and 3→1→3 are equivalent to 3→1 in argument indexation (Jacques 2021, p. 310).[4] This strongly suggests a *[CAUSER [INSTRUMENT [AGENT PATIENT]]]* hidden structure. Actually tense and aspect can be analyzed in this way as well Wiltschko (2014, § 7.4.1).

Hierarchies like this are actually one of the best criterion that tell a grammatical marker from a lexical root. In English, auxiliaries and suffixes in *have been being consulted* shows a passive < progressive < perfect < present hierarchy, which is fixed in its semantics and in its linear order. Therefore we are *not* observing complement clause constructions. On the other hand, *want to be able to do sth.* and *be able to want to do sth.* are both valid: the latter is less frequently attested but is attested anyway[5]. Therefore *be able to* and *want* are not auxiliaries – yet.

2.1.3 A note for panicking descriptive linguists

In (2), we have *determiner phrase* or *number phrase* or *nominal categorizer phrase*, but we do not have *noun phrase*. This is related to how the idea of functional heads was historically developed in generative syntax. At first we only had lexical heads, i.e. the lexical root at the center of a construction. Later it was found that certain phenomena are better captured if we assume that the functional markers have their own "phrases" as well, like DP, nP, TP, vP, etc., and finally it is found that we can keep the concept of *head* to functional markers only.

[3] See standard Chomskyan generative syntax textbooks.

[4] On the other hand, 3→1→2 becomes 3→1, and 3→2→1 becomes 3→2. But this just means that when both inner arguments are speech participants, then agentivity leads to a higher prominence. Still predictable on structural basis once we refine the empathy hierarchy.

[5] You can check yourself by searching it in COCA.

Still a descriptive linguist wants to avoid (a) explicitly mentioning (too many) functional heads, (b) using constituency relations to representing all hierarchical grammatical information, and To be fair, we *can* always do away with these. Constituency and dependency are formally equivalent, and we can always replace sentences like "in a CP, ..." by "in a full clause that allows information structure marking, ...", i.e. avoid functional heads by focus on the grammatical environments they create. What is being done here is quite similar to how in physics, virtual photons are integrated out, leaving an effective Coulomb interaction. Thus (2) may be replaced by something like

(3) $[\text{the}_{\text{definiteness}}\ \text{two}_{\text{plural}}\ [\text{ugly}\ [\text{black-bird}]_{\text{nominal compound}}\text{-s}]_{\text{modification}}]_{\text{noun phrase}}$

2.2 There is no wordhood in abstract syntax

2.2.1 Wordhood as small constituency?

Now this causes a problem. If we insist on defining a word as a small constituent, then *blackbird is* a word – but *blackbirds* isn't, because the latter has an affix with a quite high position in the structure attached to it. Which goes against the common notion of wordhood.

Similar problems occur in clausal syntax. The abstract syntax of *he sleepwalked into this frustrative situation* can be described as follows:

but *sleepwalked* is *not* recognized as a word.

2.2.2 Wordhood as phase in generative syntax?

Now, another way to define wordhood is by the concept of *phase* in generative syntax. Recall that we have said that arguments are first finished on their own and then sent to clausal syntax, so clausal syntax doesn't have a lot to do with their internal structures. On the other hand, the vp-TP-CP projections are finished in one batch. So markers of grammatical categories of valency, tense-aspect-modality, and speech forces (imperative, interrogative etc.) are *closer* to the verb in this sense. This perspective actually gives us a way to flatten the deeply hierarchical syntax tree in

But now it seems we have to recognize that *have been being performing* is also a *word* under this definition.

2.3 Abstract syntax is intricate and volatile

Abstract syntax probably is not the best

Still,

2.4 The organization of the lexicon, and morphological wordhood

If there is no good definition of wordhood in abstract syntax, then what is recorded in the lexicon?

2.5 Phonological wordhood

3 Historical wordhood and morphological or phonological wordhood

References

Bruening, Benjamin. "The lexicalist hypothesis: Both wrong and superfluous". In: *Language* 94.1 (2018), pp. 1–42.

Cinque, Guglielmo and Rizzi, Luigi. "The cartography of syntactic structures". In: (2009).

Jacques, Guillaume. *A grammar of Japhug*. Vol. 1. Language Science Press, 2021.

Wiltschko, Martina. *The universal structure of categories*. Vol. 142. Cambridge University Press, 2014.