What to compare in historical linguistics?

Jinyuan Wu

April 20, 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 (§ 2.1), and point out that syntactic structures provide no definite definition for wordhood (§ 2.2). 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.

2.1 Abstract syntax

We aim to provide *a* theory of abstract syntax of language and demonstrate that it is possible to do syntactic analysis without mentioning *words*. Note that this is only half of the study: we still need to put various abstract function items (affixes or clitics or particles) on lexical roots and get everything phonologically realized, which is discussed in § 2.3 and does give us more solid definitions of wordhood.

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]$_{\text{nP}}$]$_{\text{FP}}$]$_{\text{Num'}}$]$_{\text{NumP}}$]$_{\text{DP}}$

[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.

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) 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 denote 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.

[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.

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.

Still a descriptive linguist wants to avoid (a) explicitly mentioning functional heads, (b) using constituency relations to representing certain grammatical information, which intuitively would be better represented by dependency relations, and (c) assuming a tree that is too deep, containing many layers, constituents, etc. (nP, TP-splitting, multiple functional projections for different types of adjectives in Cartography). 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.

Dixon (2009, p. 49)[6] complains about using constituency to represent the relation between lexical items and grammatical items. He is right: for consistency, in *to the fat man*, either we write *[fat man]* as a functional projection as well (according to Cartographic syntax), or we de-emphasize the status of *the* and *to* as constituents and only treat them as markers of certain *syntactic environments* or **constructions**.[7] Thus (2) may be replaced by something like (3). This analysis avoids problem (a) by replacing concepts like DP or NumP by "a definite noun phrase with a numeral". Now since functional heads are eliminated, the term *head* can be kept to the lexical *core* of a construction, which is *blackbird* here.

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

(b) and (c) are not huge problems in (3). (c) is a problem that occurs when describing e.g. the *have been being consulted* split TP projections. These split TP projections are what are *newly* introduced into the clause, when a clause is being formed: all arguments, adverbials, etc. 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, things like the tense, aspect and modality markers are built up *within one batch* when the clause is being built. This intuition is related to the cyclic nature of syntax, and in particular, the *phase* in generative syntax. Thus markers of grammatical categories of valency, tense-aspect-modality, and speech forces (imperative, interrogative etc.) are *closer* to the verb in this sense. We therefore find a way to flatten the deeply hierarchical syntax tree: we just package everything *newly introduced into the clause*, like *have been being consulted*, into something known as e.g. *verb phrase*,[8] and then study the hierarchical relations between components of that verb phrase within the

[6] Although Dixon is strongly against formal linguistics, his Basic Linguistic Theory is largely in line with what we describe here.

[7] We however do not endorse Construction Grammar, as *constructions* defined in this way are still subject to compositional analyses.

[8] When the flatten-tree approach is not adopted, *verb phrase* often refers to the nucleus clause (i.e.

verb phrase. Thus problem (c) is solved. The result will be comparable to how the clause structure is represented in Quirk (2010): a flat syntactic tree is given first (p. 45, no excessive hierarchies of grammatical categories mentioned first), and then the authors go into the details of the hierarchy and relative scopes of auxiliaries (p. 121).

As for (b), we can replace the notion of layered functional projections by the notion of a bunch of *dependency relations* with different closeness to the lexical head. Actually the remaining constituency relations posited in (3) can also be described in terms of dependency relations: the relation between *ugly* and *blackbird* is closer than that between *the* and *blackbird*, etc. Without functional heads, dependency and constituency are still equivalent. Which language to use depends on the features of the language, like whether there are multiple topicalization and focalization (which probably will make dependency-based analysis a good choice as it makes starting easier).

Therefore, in the notation of typical descriptive grammars, we may say that there is nothing in abstract syntax besides lexical roots and grammatical categories, relations, and constructions. A one-to-one transform between the more tradition description and the generative description that seems more exotic but involves less primitive concepts is sketched in this section.

2.2 Commonly understood wordhood cannot be defined based on abstract syntax

2.2.1 Wordhood as small constituency?

The abstract syntax defined above causes a problem. If we insist on defining a *syntactic* 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:

1. *sleep* and *walk* are first placed together to form a compound, meaning that someone is walking while sleeping, with a metaphoric meaning of 'taking action blindly'.

2. The compound takes *he* and *into this frustrative situation*, two already well-formed phrases, as its arguments. *he* is structurally higher in the sense that it binds the other when a reflexive appears (thus *he*$_i$ *dreamwalked into this problem caused by himself*$_i$). The argument structure is formed.

3. The clause is in the simple past TENSE.

4. The agent in the argument structure, by default, is promoted to the subject position, as the pivot of the whole clause (which can be tested in coordination, etc.).

So *sleepwalk* forms a constituent, and can be seen as a word. Yet the past tense marker *-ed*, being added into the clause much later, has a scope that covers the whole clause. *sleepwalked* is *not* recognized as a word based on constituency!

TP) minus the subject. See e.g. Huddleston and Pullum (2002). The flat-tree notion of verb phrase is related to how wordhood is defined in § 2.2.2.

2.2.2 Wordhood in flat-tree syntax

Now, another way to define wordhood is based on the flat-tree approach mentioned in § 2.1.3. This immediately solves the problem of *sleepwalked*: now the verb, everything related to the voice, tense-aspect-modality are considered to form one "constituent" (with the definition of constituency modified a little bit to be consistent with the flattened tree), because they all below to the new things added into the clause when it is formed (see the list in the last section), and hence *sleepwalk-ed* is a word.

But now it seems we have to recognize that *have been performing* is also a *word* under this definition, if we define wordhood based on the flattened version of abstract syntax. We can make it even radical by pointing out that *have been being annoyed* is also a word in this sense. However, usually people will just call it a *verb phrase*.

In certain languages, stacked auxiliaries do have a strong "word" vibe, and what is originally considered a verb phrase may really be eventually considered a word. A good example is Modern Japanese: so-called auxiliaries in the School Grammar system, once closely inspected, look more like suffixes and not true auxiliaries, as nothing can be inserted between them and the verb stem. The so-called verb phrase in the rGyalrongic language Jiaomuzu is now described in a way that is not quite phrase-like (Prins 2016). But abstract syntax does *not* guarantee this: in English, the verb phrase (in the flattened-tree meaning) may contain materials outside of the batch newly introduced into the clause structure as well: we have *he has recently discovered that …*

2.2.3 Comment: inflection and derivation

We note that a "word" defined in § 2.2.1 is always a part of another "word" defined in § 2.2.2. Roughly speaking, the so-called "wordhood" defined in § 2.2.1 can be described as the **derived stem**, while the so-called "words" in § 2.2.2 are forms to be found in a inflection table possibly containing periphrastic forms. Note, however, that certain operations commonly known as derivation fall under the category of the latter: nominalization (*his skillful playing of the nationan anthem*, cf. the non-finite gerund clause *his skillfully playing the national anthem*), involves alternation of the subcategorization frame of the root *play*, and [*√play*, n] or [*√play*, v] both do not form constituencies in the sense of § 2.2.1. Furthermore, in Jacques (2021), valency alternation is classified as derivation, probably because of the morphological structure of the verb (derivational affixes appear to be a part of the extended stem, which is then placed into a rigid template; see § 2.3). The derivation/inflection also seems to be based on a variety of not necessarily converging criteria, just like wordhood.

2.2.4 Wordhood of function words

In abstract syntax, roots (and structures formed around them) and grammatical markers are different. Therefore even if we can define something like wordhood of function words, it will be different from the wordhood definition we desire for content words. And we do not have a clear definition of wordhood of function words, either. We may say that a grammatical marker belonging to a larger construction is a function word. Thus *the* in (2) is a function word. But in Latin, we have the *=que* clitic which works just like a conjunction, and yet it has to be attached to something else and usually

is not considered a word. Therefore for grammatical markers, wordhood is still not something definable in abstract syntax.

2.3 Lexicon, phonological realization, and morphological wordhood

2.3.1 Vocabulary insertion and morphological wordhood

Now we go to the second half of the story in § 2.1. The abstract syntax (e.g. 2 or 3) has to be linearized into the phonological i.e. surface representation. The whole process of course is guided by the lexicon, which may give us a proper definition of wordhood. In Distributed Morphology the lexicon contains List A containing roots and grammatical items, List B that guides vocabulary insertion, and List C that records idiomized meaning of everything. List A is useless in defining wordhood. List C is also hopeless, because the constructions it contain vary wildly in size: we have meanings of (category-less) roots, meanings of roots plus categorizers (thus *buffalo* in a verbal environment means 'to intimidate'), and even meanings of a whole sentence. Lexicalization is simply idiomization or in other words semantic fossilization:[9] what is being lexicalized does not have to be a word (Harley and Noyer 1999).

Thus, we should place our hope to List B.

In Distributed Morphology, phonological realization of an utterance is done by so-called post-syntactic operations: post-syntactic rules adjust the positions of lexical roots and grammatical items. It is not until this step that some cross-linguistic syntactic variance appear: for instance, where the main verb eventually appears (a syntactic property) may be determined by whether functional heads along the TP-CP hierarchy are "strong" and have to attract something to them for correct surface realization (sounds morphological). A bundle for example may look like [√*eat*, v, T[PAST]]: root *eat-*, verbalized, in past tense – basically a *verb phrase* in the sense of § 2.2.2. Then **vocabulary insertion** happens, which gives all formative phonological forms (4). This is not the final step of the syntactic operations, because the concrete phonological forms still need to undergo certain phonological reconstructions (a most radical example is Semitic template morphology; Tucker 2011).

(4) a. √love, v →love-
b. T[PAST] →-ed
c. √eat, v, T[PAST] →ate

The fact that certain roots are only used as verbs or nouns can be simply explained by stipulating that the other configurations do not have corresponding List B entries: thus √*eat*, n cannot be phonologically realized, simply because there is no such thing in the mental dictionary of English speakers. Thus **eat (n.)*

Now *this* looks like a good definition of wordhood. Post-syntactic reordering of formatives treats different parts of the *verb phrase*, or the noun, in different ways. T[PAST] in English does not want to stay alone, and wants to get attracted to something bigger: once it gets attracted near the verb root, it can no longer go away from it, besides some possible local dislocations. This is why we call *loved* or *ate* a word. on the

[9] Semantic fossilization does have syntactic effects: they may block certain movements, like topicalization of a prepositional phrase after a verb, to avoid disrupting interpretation (Nediger 2017).

other hand, things like *has been considering* are considered multi-word: T[PRESENT] still has to be attached to something else, but this time, we have a PERFECT aspect (or secondary tense, depending on terminology) feature in the clause as well, which is realized as *have-* and the two combine into *has*. The Asp[PROGRESSIVE] feature, having no tense marker to combine with it, takes the *been* form, while the main verb is in the *-ing* form surrounded by the progressive aspect. Basically, post-syntactic operations never collect T[PRESENT], T[PERFECT], Asp[PROGRESSIVE] and the the root into one bundle: the first two are placed into one bundle, the third and fourth are left on their own.

2.4 Phonological wordhood

夕贬潮阳路八千
xī biǎn cháo yáng lù bā qiān

2.5 Interim summary

- Lexicalized items: roots, words in the usual sense, phrases, clauses.
-

3 The unit of transmission

What is discussed in § 2 is what is in the mind of someone already speaking a language. Yet when they talk to their kids, their kids know nothing about roots or abstract grammatical items: the kids analyze what they hear and build their own grammars and lexicons. Also, the mental lexicons of adults undergo graduate change because of social factors as well. In theory, any historical law proposed on language evolution should be based on psycholinguistics of language transmission. Such a microscopic foundation however is currently lacking, and the only thing we can do is to go over § 2.5 and check whether they are the unit of language transmission.

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).

Dixon, Robert MW. *Basic linguistic theory volume 1: Methodology*. Vol. 1. OUP Oxford, 2009.

Harley, Heidi and Noyer, Rolf. "Distributed morphology". In: *Glot international* 4.4 (1999), pp. 3–9.

Huddleston, Rodney and Pullum, Geoffrey K. *The Cambridge Grammar of the English Language*. Cambridge University Press, 2002.

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

Nediger, Will. “Unifying structure-building in human language: The minimalist syntax of idioms”. PhD thesis. UMich, 2017.

Prins, Marielle. *A grammar of rGyalrong, Jiǎomùzú (Kyom-kyo) dialects: A web of relations*. Vol. 16. Brill, 2016.

Quirk, Randolph. *A comprehensive grammar of the English language*. Pearson Education India, 2010.

Tucker, Matthew A. “The morphosyntax of the Arabic verb: Toward a unified syntax-prosody”. In: (2011).

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