

What to compare in historical linguistics?

Jinyuan Wu

April 23, 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 (“narrow”) 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 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,¹ you will know that it seems we cannot define wordhood in a completely intuitive way in *abstract* or *pure* syntax.² 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_D [two [-s [[ugly]_{AP} [blackbird]_N]_{FP}]_{Num'}]_{NumP}]_{DP}]*.³ 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 ugly blackbirds* are probably to be analyzed as something like (2), where DEFINITE and PLURAL are going to be replaced by the particle *the* and the suffix -s after phonological realization (§ 2.3). Here following the usual Distributed Morphology assumption that *blackbird* 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. The adjective phrase *ugly* has its own internal structure, but for simplicity it is not represented here.

¹Note that Distributed Morphology can be formulated into a form quite close to Word-and-Paradigm theories of morphology (Ermolaeva and Edmiston 2018).

²Lexicalists will push back – but I believe they are wrong (Bruening 2018).

³We are describing the phrase structure using *functional heads*; see the end of this section, and § 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) [DEFINITE [two [PLURAL [[ugly]_{AP} [√black √bird]_{NP}]_{FP}]_{Num'}]_{NumP}]_{DP}

2.1.2 Abstract syntax as hierarchies of grammatical categories around roots

There is nothing in abstract syntax besides roots and what are known as functional heads in generative syntax, more commonly called grammatical markers in descriptive linguistics. And the grammatical categories, together with “arguments” they introduce (including clausal arguments, adjective phrases, etc.) are wrapped around the core 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,⁴ 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 agentive 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 $X \rightarrow Y \rightarrow Z \rightarrow W$), and the polypersonal direct-inverse indexation on the main verb (with the form $X \rightarrow Y$) 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 $1 \rightarrow 3 \rightarrow 2 \rightarrow 3$ configuration is morphologically the same as $1 \rightarrow 2$ (Jacques 2021, p. 848, (67)). Similarly, both $2 \rightarrow 3 \rightarrow 1$ and $2 \rightarrow 1 \rightarrow 3$ are equivalent to $2 \rightarrow 1$ in argument indexation (Jacques 2021, p. 584), and both $3 \rightarrow 3 \rightarrow 1$ and $3 \rightarrow 1 \rightarrow 3$ are equivalent to $3 \rightarrow 1$ in argument indexation (Jacques 2021, p. 310).⁵ 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 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.⁶ Therefore *be able to* and *want* are not auxiliaries – yet.

⁴See standard Chomskyan generative syntax textbooks.

⁵On the other hand, $3 \rightarrow 1 \rightarrow 2$ becomes $3 \rightarrow 1$, and $3 \rightarrow 2 \rightarrow 1$ becomes $3 \rightarrow 2$. 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.

⁶You 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, which roughly corresponded the 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)⁷ complains about using constituency to represent the relation between roots 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**.⁸ 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 *core* of a construction, which is *not* a grammatical item and is *blackbird* here.

(3) [the_{DEFINITENESS} two_{PLURAL} [ugly [black-bird]_{nominal compound} ^{-S}PLURAL_{modification}]_{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. The clause doesn’t see finished materials clearly and can only see things like person, number, etc., but it sees tense, aspect, modality, etc. clearly. 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*

⁷ Although Dixon is strongly against formal linguistics, his Basic Linguistic Theory is largely in line with what we describe here.

⁸ We however do not endorse Construction Grammar, as *constructions* defined in this way are still subject to compositional analyses.

been being consulted, into something known as e.g. *verb phrase*,⁹ and then study the hierarchical relations between components of that verb phrase within the 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 head (the root at the center of a construction, like *blackbird* in *the two ugly blackbird*, not a functional 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 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 (we can also write it in a more compact way as in 3: see 4):

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.

⁹When the flatten-tree approach is not adopted, *verb phrase* often refers to the nucleus clause (i.e. 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.

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!

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 flat-tened tree), because they all belong 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. This is shown in (4): the already finished materials are labeled in gray, leaving only the verb compound and the tense marker in black.

- (4) [[*he*]_{subject,i} [*—*]_i [*sleep-walk*]_{verbal compound} [*into this frustrating situation*]_{location}]_{argument structure} ^{-PAST}]_{declarative clause}

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 root. 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 ...*

In the same way, we may even want to say that *the* and *blackbirds* in *the two ugly blackbirds* form a syntactic word, if we accept the definition of wordhood developed above: in the same way we gray out the already finished syntactic objects in (4), we should gray out *two* and *ugly* as they are well-formed phrases on their own, while *the* and *blackbirds* are newly introduced materials when constructing the noun phrase. Arabic and Hebrew’s nouns actually fit well in this definition of noun wordhood, but English nouns certainly do not.

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 an 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 national anthem*, cf. the non-finite gerund clause *his skillfully playing the national anthem*), involves alternation of the subcategorization frame of the root *play*, and [$\sqrt{\text{play}}$, n] or [$\sqrt{\text{play}}$, 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; § 2.3.2). The derivation/inflection also seems to be based on a variety of not necessarily converging criteria, just like wordhood. And given the inherent ambiguity of the derivation/inflection distinction, the concept of *step* is not always useful.

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 Phonological realization guided by lexicon

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 phonological realization of the roots and grammatical items, and List C that records idiomized meaning of everything. We should note that what is discussed here is about ideal competence of a person already fluent in a language. We can expect that the human brain always wants to find shortcuts and does not start the whole structural building process from scratch all the time (Matchin and Hickok 2020), and tends to store finished trees in the mental lexicon, and that Lists A, B, and C are actually kind of mixed in the actual brain.

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:¹⁰ what is being lexicalized does not have

¹⁰Semantic 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). Therefore, it can be a hotbed for *syntactic* fossilization, i.e. gradual erosion of the internal structure of a stored structure in the lexicon.

Note that syntactic fossilization is not always accompanied by loss of compositional semantics: syn-

to be a word (Harley and Noyer 1999). Thus, we should place our hope in 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 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 [\sqrt{eat} , v, T[PAST]]: root *eat*-, verbalized, in past tense – basically a *verb phrase* in the sense of § 2.2.2. Then **vo-cabulary insertion** happens, which gives all pieces in the syntactic structure (roots or functional items) phonological forms, turning them into substantive **exponents** (5).¹¹ 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).

- (5) a. $\sqrt{\text{love}}$, v \rightarrow love-
- b. T[PAST] \rightarrow -ed
- c. $\sqrt{\text{eat}}$, v, T[PAST] \rightarrow 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 \sqrt{eat} , n cannot be phonologically realized, simply because there is no such thing in the mental dictionary of English speakers. Thus **eat* (n.)

2.3.2 Bundles in phonological realization as morphological words

Now we see something that looks like a good definition of wordhood. Post-syntactic reordering of exponents treats different parts of the *verb phrase*, or the noun phrase minus adjectives, or whatever considered to be a word in the sense of § 2.2.2, 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 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

tactic fossilization of a prototypical clause structure into a verb morphological template involves nothing non-compositional. But as is said above, as a shortcut in the brain, a completed tree – like a simple clause – therefore can also be stored and undergoes mild semantic fossilization, which starts its syntactic fossilization.

¹¹Some terms with meanings similar to *exponent* should be noted here. The term **formative** is sometimes used to refer to a piece in morphology from the parsing side, not necessarily with a clearly understood grammatical function – it may be even totally historical and mark no synchronic grammatical category. A *exponent* on the other hand always has a clearly specified grammatical function. Thus we may say “zero exponent” – but rarely we say “zero formative”.

Sometimes, in certain Distributed Morphology papers, it means things in List A, i.e. abstract primitive syntactic objects. This meaning is completely opposite to its usual surface-oriented meaning.

Finally we have the good old term *morpheme*, which unfortunately is theoretically loaded, implying a transparent, one-to-one relation between form and meaning, and indistinguishability between roots and grammatical items.

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.

In a sentence: *what are moved together in phonological realization form a morphological word*.¹² Or to be more concise: *morphological wordhood is about morphological selection*. Note that this definition also defines function words, which contain no roots, but may have behaviors comparable to content words. We have just seen how the auxiliary *has* appears in the English *present perfect* in third person: the “stem” *have* is purely a PERFECT marker here, but when it is combined with the third person singular *present* marker *-s*, it inflects just like an ordinary verb. This solves the problem in § 2.2.4.

We note that the process sketched here generally tends to make exponents of grammatical categories closer to the root also closer to the root in the linear order after phonological realization. In reality, this is not always the case: we have both *layered* morphology which represents the hierarchical structure and *slot-filler* or *template* morphology, which is linear and flat and defined in terms of slots, which is however still within the formal complexity class of Distributed Morphology, as local dislocation rules, some of which may have phonological motivations, can reorder the exponents; syntactic ordering of grammatical categories is also a possible explanation (Bye 2020).¹³ By this logic, almost arbitrary reordering of affixes is also not impossible, which is indeed what is observed in Chintang; we nevertheless acknowledge that the affixes are indeed affixes, because they pass tests for grammatical wordhood (i.e. morphological wordhood in this note), like obligatoriness, selectivity of hosts, and interaffix dependence, which clearly indicate that they are collected together by post-syntactic morphological operations (Bickel et al. 2007).

2.3.3 Accidentally fixed linear orders

What is sketched above (§ 2.3.2) is the most natural definition of morphological wordhood from the perspective of *production*. We may also want to define morphological wordhood from the perspective of *parsing*: whatever seems to have a fixed morphological pattern is often recognized as a morphological word. This leads to an interesting corner case of morphological wordhood definition, although it is not as drastic as § 2.2.1 and § 2.2.2.

Certain “morphological templates” identified in a surface-oriented analysis do not seem to need any post-syntactic relocations. Consider an imaginary dialect of English, where there were far less tense-aspect-modality adverbs. This would result in a lot of *have been being asked* sequences not interrupted by inserted adverbs. Further let us suppose that this dialect of English had lost the subject-auxiliary inversion rule (not uncommon in contemporary casual speech: *you know what?*). A linguist analyzing this would face the dilemma of whether to consider the whole sequence as a mor-

¹²In a lexicalist theory they may be known as X^0 nodes, as in e.g. Bickel et al. (2007). However, recent studies in Minimalist syntax are becoming more suspicious of the syntactic status of head movements, and here we follow Distributed Morphology and consider them to be formed by post-syntactic operations.

¹³What seems more problematic to syntax-oriented theories of morphology like Distributed Morphology is the existence of extended exponents (e.g. Bickel et al. 2007). This however can still be captured by assuming fission of a morphological feature (Bobaljik 2017).

phological word. It definitely looks like a morphological word, but no post-syntactic relocation of abstract or concrete morphological pieces is necessary to generate it.¹⁴ This is indeed the case of Modern Japanese discussed in § 2.2.2. This is probably more clearly seen in the evolution of Turkish: certain “suffixes” are written as auxiliaries in Ottoman Turkish and are written as suffixes in modern Turkish, but the linear forms are exactly the same, so we see two competing analyses of the same linear surface form.

We may want to comfort ourselves by focusing on the fact that a sequence like this will likely soon grammaticalized into an authentic morphological template. We will go back to this issue when discussing clitics – discussed immediately below.

2.3.4 Clitics and morphological wordhood

Probably a more important phenomenon that may disrupt the definition of morphological wordhood in § 2.3.2 is **clitics**. A clitic, just like ordinary affixes, needs to be attached to something, but it is less picky when choosing the host and deciding where to land. A good example is the Latin conjunction *=que*, which can be attached to any inflected head noun, sometimes an adjective, in a noun phrase coordination. The fact that they need to be attached to something seems to suggest that they pass the morphological wordhood test in § 2.3.2, as they are dislocated by morphological rules as well (Harley and Noyer 1999), and should we consider them to be a part of the morphological word they attach to as well?

If the answer is “no”, then a clear-cut distinction between clitics and affixes has to be made. An observation is that the attachment of clitics to other words is “late”, while the post-syntactic relocation of typical affixes is “early”: typical affixes form their own units first, and then clitics are attached to these units (e.g. Jacques 2021, p. 485). The “late incorporation” definition of clitics seems to be surprisingly stable cross-linguistically. In Romance languages, personal clitics can only attach to the verb: but they are nonetheless clitics because they are not compatible with noun phrase arguments, and therefore they have to be originally pronouns, only *late* incorporated into the verb. Latin *=que* is also attached to things when everything else is formed, so it is attached to nouns or adjectives *late*.

However, in these cases, the clitics cannot see the internal structure of a morphological word, and can never cross the boundary of a morphological word: affixes reordering is possible, but a clitic cannot be incorporated into the units they have already formed (Embick et al. 2007). Whether this statement is universally true or not is not clear, and we have evidence against it. It seems that in Udi, we have personal agreement formatives that can be attached to both focalized elements and verbs. So, these formatives are more picky than the Latin *=que*, but still much more flexible than typical affixes do, and it seems they should be analyzed as clitics. It is then observed that these clitics can invade the verb morphological template (known as **endoclitization**), and their positions of clitics in the verb morphological template are subject to the control of other formatives (Harris 2000).

¹⁴To be fair, if a content word has layered morphology, then the rigid order of formatives in it may also transparently reflect syntax. But post-syntactic morphology is doing something here: it *selects* grammatical categories it wants to put into the inflectional paradigm, and leaves the rest to auxiliaries, particles, etc. In *have been being asked*, post-syntactic morphology is still doing things, but it’s mostly constructing the auxiliaries one by one, without any non-trivial relation *between* them.

Once the boundary of pre-formed morphological words is no longer a problem for cliticization, distinguishing clitics from affixes becomes less easy. Chintang, a rather unusual language in which prefixes take arbitrary orders, also have focus clitics that can be incorporated into the verb and they occasionally appear between prefixes, making a focus clitic just like another prefix that can relocate arbitrarily. Besides verbs, pronouns and adverbs can also be focalized by the same clitic, but this doesn't say much because there are affixes that apply to both nouns and verbs: what *can* show that a formative is a clitic is that it can attach to multiple hosts *in the same construction*, but what we see in Chintang is that we need to shift the focus for the clitic to jump to another word.

Tests on the linear order of formatives are usually not completely convincing in these cases. In theory, if language-specific morphological factors can reorder exponents to create a flat template with a rigid linear order regardless of syntax, the same can be done to clitics, and indeed in Udi, just as is shown above, personal agreement clitics, when incorporated into verbs, have their pre-specified slots in the morphological template. Or we are allowed to reorder both affixes and clitics in a quite wild way, as is the case in Chintang. Or no reordering is done: if auxiliaries can receive a rigid linear order without the possibility of intervention of any other materials, then clitics can, too.

Often, we start by noticing that a sequence of formatives follows a rigid order which doesn't seem to be completely transparently come from syntax (having a slot-and-filler structure that doesn't reflect syntactic scopes of the grammatical categories, or having a layered structure not covering all grammatical categories attested in the grammar of the language, showing morphological selection), while another set of formatives are strictly attached behind them and sometimes after other particles (e.g. Jacques 2021, § 11.6.2). But in this case, what actually does the heavy-lifting job is the morphological coherence of the formatives with the rigid order: we let the rigid order to guide ourselves to *possible* morphological words, as many languages do not madly alter the order of formatives, but we always need further evidence to support morphological wordhood there. On the other hand, without a rigid order, morphological wordhood can still be determined, as in the case of Chintang, by tests like obligatoriness, selectivity of hosts, and interaffix dependence (Bickel et al. 2007).

In conclusion, the only thing that sets clitics apart is evidence supporting them being incorporated into a morphological word lately. What counts as valid evidence includes the formative in question being in conflict with some other components in the construction (thus a formative in conflict with explicit arguments is a pronoun in disguise, hence likely a clitic), or the formative being able to attach to multiple hosts *without any alternation of the construction containing it*, or the formative seeming to interact with a morphological unit already well-formed. Other criteria, like rigidity of linear order, are generally not decisive in the most puzzling cases but can provide candidates for wordhood.

2.4 Phonological wordhood

A final type of wordhood is defined not via morphosyntax, but via phonology: what forms a unit in phonology is considered a word. Dixon (2010, § 10.3) overviews some of common criteria used to define the phonological word. Cross-linguistically, different phonological processes may happen in different domains (Schackow 2015, p. 62),

making a cross-linguistic definition of phonological wordhood impossible. What is cross-linguistic is the relevant phonological phenomena, not the exact definition of phonological wordhood.

Moreover, phonological wordhood can be inconsistent with all types of wordhood defined above: its relation with morphological words can be quite arbitrary, and it does not always respect syntactic “wordhood” defined in § 2.2.1 or § 2.2.2. Formation of phonological words can ignore a lot of morphosyntactic information. To illustrate the point, let us start with an easy example (6). Prosody plays an important role in modern Mandarin: an utterance is divided into a series of disyllabic prosodic words from left to right, with the first syllable being heavy and the second light. Thus in the reading convention in modern Mandarin Chinese of Classical poetry, a line like (6a) is read as (6b). Note that the first prosodic word is inconsistent with the constituency relations in (6a).

- (6) a. 夕贬潮阳路八千
 [xī]_{temporal} [- biǎn [Cháoyáng]_{locative}]_{verbal predication} [lù bā
 evening relegate PLACE road eight
 qiān]_{non-verbal predication}
 thousand
 ‘In the evening, [I] was relegated to Chaoyang, the road being eight thousand miles.’
 b. 夕贬 | 潮阳 | 路八千

In (6b), a phonological word – 夕贬 – contains two morphological words which do not form a syntactic unit in the sense of § 2.2.1 or § 2.2.2, a phonological word is identical to a morphological word (潮阳, a place name), and a phonological word is a clause on its own (路八千). Thus, there is no guarantee that a simple relation exists between morphological and phonological wordhood.

Now, cross-linguistically, a review of possible relations between morphological and phonological wordhood can be found in Dixon (2010, § 10.6). Not many generalizations can be made on what is found.

2.4.1 A morphological word consisting of more than one phonological word

First, it is possible that a morphological word consists of more than one phonological word. Basically this means the morphological word breaks into parts in the middle.

A complex verb (a morphological word) in Mandarin containing a disyllabic root and a directional complement, like 支楞起来, has to be divided into two phonological words, one containing the root, another containing the directional complement. A similar phenomenon is observed in Moloko, where a grammatical word (itself containing a root and affixes) attracts several clitics to form a verbal complex, which then is reorganized into two phonological words (Friesen 2017, p. 202).

In Bickel et al. (2007), it is proposed that all prefixes themselves are phonological words and attach to phonological words, which leads to the nearly free order of formatives before the stem.

What are shown above all have a pattern: a morphological word breaks into phonological pieces, with one piece containing the root. But it is also possible that the two phonological words from one morphological word both contain roots, when the morphological word is a compound (Dixon 2010, p. 23).

2.4.2 A phonological word consisting of more than one morphological word

Second, it is possible for a phonological word to consist of more than one morphological word. This usually means the morphological words are too lightweight to appear alone as phonological words.

We may immediately think of cliticization. If a clitic is not considered to be a part of its host, then cliticization that involves local dislocation not motivated by pure phonological reasons can be seen as producing phonological words containing multiple morphological words (§ 2.3.4). It should be noted that some clitics may not be completely fused into the host they attach to in phonology.

We also have examples of *phonological* cliticization, which are clearer for demonstrating the idea of a phonological word consisting of multiple morphological words. Dixon (2009, p. 49) calls *to* and *the* in *to the fat man* clitics, because they are not stressed, and almost form a phonological word with *fat*. What makes them different from clitics in § 2.3.4 is that the positions of *to* and *the* are never shifted by any post-syntactic morphological adjustments: they are attached to *fat* only because they are linearly next to it, and when another word, like *famous*, is inserted between them and *fat*, they have no reason to be attached to *fat*. On the other hand, clitics in § 2.3.4 need to undergo readjustments of their positions. Thus *to* and *the* in *to the fat man* are purely *phonological* clitics, and when they are attached to content words following them, a phonological word containing several morphological words (e.g. *to=the=fat*) forms.

The examples given here look like the inverse of what is given in § 2.4.1. There, a morphological word (i.e. the realization of a flattened-tree syntactic “word” defined in § 2.2.2) breaks into pieces, and one piece contains the root. Here, morphological pieces – one of which contains the root – are put together to form a larger unit in phonology. But we should note that this does not always have to be the case: 夕贬 in (6) is also a phonological word containing two morphological words, and both morphological words are made of purely roots. Moreover, 夕贬 and *to=the=fat* are both *not* flattened-tree syntactic “words”: the morphological words in them are there purely out of accident, without any implications on syntactic structures.

2.4.3 Irregular correspondences

Third, a particularly interesting fact is that it is possible for a part of a morphological word to be attracted to a nearby morphological word to form a phonological word (Dixon 2010, p. 24). (6) already suggests that this is possible, although 贬 there does not form a morphological word with its locative argument, and 夕贬 therefore (6) should still be described as a phonological word containing two morphological words.

2.5 Interim summary

In synchronic description of morphosyntax, the concept of wordhood is in principle not a must: in morphosyntax, we can talk about small constituents, lexicalization (as idiomization), formatives appearing together because of morphological operations. We do have word-like units in phonology, but they are sometimes inconsistent with whatever morphosyntax wordhood we propose.

Anyway, we have done a thorough survey of everything in morphosyntax and in phonology that looks kind of like a word. Our findings are summarized in the list below.

- Syntactic wordhood based on syntactic constituency (§ 2.2.1). This definition is too narrow and essentially is a over-narrow definition of the *stem* (i.e. one or more roots plus derivations), which excludes certain processes also commonly known as derivation (§ 2.2.3).
- Syntactic wordhood based on flattened-tree constituency (§ 2.2.2). Things considered to be a part of a single word in § 2.2.1 are also considered to be a part of a single word in § 2.2.2. This definition of wordhood is too broad: essentially throws almost every part of grammar into an inflection table, as now we have to acknowledge that *have*, *been* and *exercising* in *have recently been actively exercising* form a single word. Interestingly, criteria (e-f) in Dixon (2010, pp. 15-16) are all satisfied in “words” defined according to the standard of § 2.2.2: you do not see multiple occurrences of an auxiliary in a verb phrase (flattened-tree version) either!
- Criterion (b) in Dixon (2010, p. 13) states that a grammatical word (i.e. a morphological word, as we have no well-defined syntactic wordhood) has a conventionalized coherence and meaning. We note that this is true for words defined by intuition, but also true for some roots, and certain phrases and even clauses.
- Criteria (a, c) in Dixon (2010, pp. 13-16) state that a grammatical word is one or more lexical roots to which morphological processes have applied, whose formatives always occur together. These criteria are essentially syntactic wordhood defined in § 2.2.2 narrowed down by purely morphological i.e. realizational considerations: formatives that pass the test of § 2.2.2 and are collected into one location by post-syntactic morphological operations form a word (§ 2.3.2). The main problem of this definition is the existence of clitics, which also satisfy the “togetherness” condition. What sets affixes and clitics apart is that cliticization happens *later* than formation of morphological words consisting of affixes and roots because of reduced morphological selectivity. Still a clear distinction is not always possible (§ 2.3.4). Another problem is that in certain languages, a list of formative appearing together not allowing intervention of any other materials may simply be a transparent reflection of syntax (§ 2.3.3).
- Criterion (d) in Dixon (2010, p. 14) states that formatives in a morphological word generally occur in a fixed order. This is not always true, as is discussed in § 2.3.3: a fixed order may be a direct reflection of syntax and what we see may just be a sequence of particles and auxiliaries (although in this case, grammaticalization will soon happen), and certain uncontroversial morphological words allow reordering of affixes. Fixed linear orders often accompany wordhood in many languages, but a double check to check if morphological selection is always needed.
- Phonological wordhood can be defined according to multiple standards, and a cross-linguistic definition is impossible. We should note that cliticization is also relevant in forming a phonological word. Further, phonological words and morphological words do not necessarily have clear relations (§ 2.4). Criterion (g) in Dixon (2010, p. 18) uses pause between words to define wordhood, which may be seen as one type of phonological wordhood.

Given this wild diversity of wordhood criteria, what is striking is not the absence of clear wordhood in certain languages, but that these criteria still roughly converge in many languages. An actual human learner, despite being able to acquire a language with high irregular correspondences between morphological and phonological wordhood, may still feel the necessity of relatively regular correspondences between the two, which enables more shortcuts in language processing in the brain (§ 2.3.1). We are going to tentatively touch this topic below.

3 The unit of transmission and evolution

We expect the unit of transmission in historical evolution of languages to be neither too big (clause-like) nor too small (root-like). It is frequent that roots does not appear in any natural utterance: Latin is a quite good example. A big unit like a clause is likely decomposed when transmitted.¹⁵ So the primary locus of transmission probably will be something – probably more than one – in § 2.5. 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.

Language change involves both phonological and morphosyntactic changes. The latter involves reanalysis (alternation of underlying structure with the surface form staying the same), extension (the surface form varies, often after reanalysis introducing new forms, like auxiliaries, while the underlying structure does not undergo major changes), and borrowing. As extension creates new forms which are absent in the parent language, reanalysis should be our main focus when studying the unit of language transmission from the morphosyntactic perspective. Below, we consider certain situations where the definition of *word* is unclear, and discuss the units historical linguistic laws work on in these situations.

3.1 Phonological word containing multiple morphological words

When a phonological word consists of multiple morphological words, will the phonological word be the transmitted unit, or will the morphological words be the transmitted unit?

It is clear that some instances of reanalysis are due to attempts to parse *phonological* words. Thus we see *a nuncle*, which is a result of re-bracketing *an=uncle*, which is the phonological realization of a simple noun phrase with the underlying representation $[a_{\text{SINGULAR,INDEFINITE}} [\text{uncle}]_{\text{noun stem}}]_{\text{noun phrase}}$, where the indefinite, singular categories are realized as *a*, which, by virtue of appearing before a vowel, becomes its allomorph *an*, and then is attached to the noun stem *uncle*. What is being transmitted and undergoes re-bracketing is the whole sequence *an=uncle*. In this instance, the abstract syntactic structure is *not* at work, or otherwise the appearance of *n-* cannot be explained, and boundaries of morphological words do not matter: *an* is not a morpho-

¹⁵Some large units, like stories or legal principles, can be transmitted quite stably, as is seen in e.g. Indo-European languages (Fortson IV 2011, Ch. 2). But what are transmitted here are *abstract ideas*, and quite different sentences can be used to describe these cultural traits. They are valuable in providing candidates for cognates, but provide little materials for clause-level or phrase-level comparison.

logical clitic but a phonological one (§ 2.4.2), and therefore it always has a boundary with *uncle*, and if this boundary is respected here, the re-bracketing can never appear.

In other cases, morphological words seem to be the unit of transmission. Latin *=que* doesn't leave any traces on the evolution of nominal declension of Romance languages.

3.2 The role of abstract syntactic structures

A question is whether historical linguistic law work on *syntactic* “wordhood”. We may want to argue that flattened-tree syntactic “wordhood” (§ 2.2.2) is important in historical syntax, because this is how we get auxiliaries. But we can alternatively explain the grammaticalization of auxiliaries by stipulating that the *linear* sequence *have performed* first frequently appeared in the complement clause construction, and then was reanalyzed as an auxiliary construction. To prove that a certain reanalysis is based on hierarchical syntax and not just linear order and phonological wordhood, we need to show that even when what is attested is always in the form of [A B C], if A and C are somehow “close” to each other according to a certain standard (like those in § 2.2.2), in daughter languages, there is a possibility for A and C to be moved together by morphophonological rules or a possibility for A to grammaticalize.

TODO: 且夫

3.3 Extremely big morphological words

Now we want to know if *all* morphological words are units of transmission. Small ones definitely do, but what about ultra-complicated ones with multiple slots? The question “are all morphological word units of transmission” therefore is equivalent to “is complex morphology stable in language evolution.” If long prefix chains and suffix chains are not stable at all

Still there is the possibility that what is being transmitted is actually a phonological segment of the morphological word in question.

4 Reconstruction based on content words

5 Reconstruction based on grammatical markers

6 Possible artifacts in a reconstructed proto-language

It is unlikely that a reconstruction proto-language of a language family looks incredibly close to the real common ancestor of that family. A reconstructed proto-language is a *model* that explains the similarities and differences we now observe in the daughter languages, which (a) does not necessarily have the historically correct phonetic values, and (b) likely conflates phenomena from different dialects and periods into one single system. Some of the problems can be avoided by more careful reconstructions, but reviving the historical common ancestor is generally speaking beyond the capacity of historical linguistics.

This means artifacts originating from techniques used in reconstruction, not real historical information, appear in the reconstructed language. Here we discuss some possible artifacts relevant to the problem of wordhood.

6.1 Lack of function words

Reconstructed languages generally do not have a ton of function words. This is because function words are almost always results of grammaticalization of content words. Take a look at function words in contemporary languages and you see grammaticalization. Take a look at classical texts of dead languages – and you still see grammaticalization. Therefore, when a

6.2 Complex morphology

It is possible that two affixes of a reconstructed word were never both highly productive historically. Thus in a form like Σ -*A-B*, it is possible that *A* was once a productive derivational suffix, but had fallen out of use when *B*, maybe originally a clitic, was finally grammaticalized.

Another problem is clitics being reconstructed as affixes?

References

- Bickel, Balthasar et al. “Free prefix ordering in Chintang”. In: *Language* 83.1 (2007), pp. 43–73.
- Bobaljik, Jonathan David. “Distributed morphology”. In: *Oxford research encyclopedia of linguistics*. 2017.
- Bruening, Benjamin. “The lexicalist hypothesis: Both wrong and superfluous”. In: *Language* 94.1 (2018), pp. 1–42.
- Bye, Patrik. “Morpheme ordering”. In: *Oxford Research Encyclopedia of Linguistics*. 2020.
- 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.
- Dixon, Robert MW. *Basic linguistic theory volume 2: Grammatical topics*. Vol. 2. Oxford University Press on Demand, 2010.
- Embick, David et al. “Linearization and local dislocation: Derivational mechanics and interactions”. In: *Linguistic analysis* 33.3-4 (2007), pp. 303–336.
- Ermolaeva, Marina and Edmiston, Daniel. “Distributed morphology as a regular relation”. In: *Society for Computation in Linguistics* 1.1 (2018).
- Fortson IV, Benjamin W. *Indo-European language and culture: An introduction*. John Wiley & Sons, 2011.
- Friesen, Dianne. *A grammar of Moloko*. African Language Grammars and Dictionaries 3. Berlin: Language Science Press, 2017.

- Harley, Heidi and Noyer, Rolf. “Distributed morphology”. In: *Glott international* 4.4 (1999), pp. 3–9.
- Harris, Alice C. “Where in the Word is the Udi Clitic?” In: *Language* (2000), pp. 593–616.
- 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.
- Matchin, William and Hickok, Gregory. “The cortical organization of syntax”. In: *Cerebral Cortex* 30.3 (2020), pp. 1481–1498.
- Nediger, Will. “Unifying structure-building in human language: The minimalist syntax of idioms”. PhD thesis. UMich, 2017.
- Prins, Marielle. *A grammar of rGyalrong, Jiāomùzú (Kyom-kyo) dialects: A web of relations*. Vol. 16. Brill, 2016.
- Quirk, Randolph. *A comprehensive grammar of the English language*. Pearson Education India, 2010.
- Schackow, Diana. *A grammar of Yakkha*. Language Science Press, 2015.
- 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.