What to compare in historical linguistics?

Jinyuan Wu

February 11, 2026

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*$_{\text{D}}$ *[two [-s [[ugly]*$_{\text{AP}}$ *[blackbird]*$_{\text{N}}$*]*$_{\text{FP}}$*]*$_{\text{Num'}}$*]*$_{\text{NumP}}$*]*$_{\text{DP}}$.[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 ungly 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 *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. The adjective phrase *ugly* has its own internal structure, but for simplicity it is not represented here.

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

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

[3] 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) [DEFINITE [two [PLURAL [[ugly]$_{AP}$ [√black √bird]$_{nP}$]$_{FP}$]$_{Num'}$]$_{NumP}$]$_{DP}$

How (2) is turned into its surface form involves post-syntactic operations that bridge syntax and phonology, commonly known as (realizational) morphology. This is discussed in § 2.3.

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 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).[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. The arrangement of tense-aspect-modality adverbs

[4] See standard Chomskyan generative syntax textbooks.

[5] 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.

appears to be a part of the same hierarchy (Cinque 1999). Therefore we are *not* observing an ordinary complement clause construction. 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.

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. The resulting theory framework is what is used in Huddleston and Pullum (2002) and many more "structuralist" grammars; for a summary, see Reynolds, Arora, and Schneider (2023). We note that Huddleston and Pullum (2002) still mix functional heads and content word heads: in the way they deal with what are known as prepositions, for instance, they treat prepositions almost as a part of speech of *content words*, and then goes on to talk about syntactic prepositions. They however should draw a clear line between prototypical peripheral case markers i.e. syntactic preposition and prototypical complement-taking adverbs i.e. prepositions that are content words, and then discuss gradience (i.e. multiple analyses) between the two. Unfortunately they do not. If this partial recognition of functional heads is to be continued, they probably should recognize coordinators as heads as well, which they however are reluctant to do for obvious reasons (Reynolds, Arora, and Schneider 2023).[7]

Dixon (2009, p. 49)[8] 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

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

[7] Another inconsistency in Huddleston and Pullum (2002) is that sometimes they give up explicit constituency analyses and focus on linear orders. This happens when they analyze different adjuncts: some of them are actually peripheral arguments, while some belong to the tense-aspect-modality system (Cinque 1999). This probably explains why some find the older CGEL, i.e. Quirk (2010), to be more self-consistent in the theoretical framework, although the latter has rather ill-defined terminology (Huddleston 1988).

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

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**.[9] Thus (2) may be replaced by something like (3), which avoids problem (a) by replacing concepts like DP or NumP by "a definite noun phrase with a numeral", and also contains the final phonological realization of grammatical categories (DEFINITE →*the*) for convenience. Now since functional heads are eliminated, the term *head* can be kept to the *core* of a construction, and not grammatical markers. We intend to use the term *head* loosely here: any reasonable constituency containing the center root of a construction can be called a head, and thus we can say that *blackbird* heads (3), although within *blackbird* we may still say that *-bird* is the head. In cases like *the two ugly blackbirds*, we call *blackbird*, a head of the construction that is deeply idiomized and may contain more than one root, the *head stem* of the construction or the stem at the center of the construction, to avoid meaningless discussions on what is the head *root* of the construction, although the definition of the term *stem* is inherently vague (§ 2.2.3). It is hard to see what should be seen as the core of a construction (e.g. when analyzing coordination) and in this case we simply stop using the term *head*.[10]

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

(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. 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 been being consulted*, into something known as e.g. *verb phrase*,[11] 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 stem 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*

[9] We however do not endorse Construction Grammar, as *constructions* defined in this way are still subject to compositional analyses. See § 2.3.1.

[10] If we go back to generative syntax, the (functional) head is the coordinator.

[11] When 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.

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.

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 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. 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]$_{\text{subject},i}$ [—$_i$ [sleep-walk]$_{\text{verbal compound}}$ [into this frustrative situation]$_{\text{location}}$]$_{\text{argument structure}}$-PAST]$_{\text{declarative clause}}$

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 certain do not.

A *word* defined in this section, when it contains a root,[12] probably should be considered as a *form* in an inflection table i.e. a **paradigm**, and not literally a word, as an inflection paradigm maps a root with all newly added materials in the construction headed by the root, i.e. the grammatical categories, into a concrete form. And in an inflection paradigm, *wordhood* defined in this section starts to make sense. The English *has …been …reading* "word", which contains all *newly added* materials in a PRESENT PROGRESSIVE clause with a third person subject, is indeed a *inflected form* in the English verbal inflection paradigm, although there is no guarantee that all parts of the inflected form are held together by post-syntactic morphological operations (§ 2.3.1): the auxiliaries are subject to movements, and adverbs may interrupt the linear continuity.

Whether the syntactic "wordhood" defined in this section sounds absurd is strongly language-specific. We have argued above that *the* and *blackbirds* in the noun phrase *the extremely ugly and annoying blackgirds* contain all the materials newly added into the noun phrase besides the already finished *extremely ugly and annoying*, and therefore form a syntactic word. in Modern German, considering the article and the noun as one word – despite them being separable – is beneficial for descriptive purposes, as

[12]The standard of being finished varies. In languages where the distinction between TP and CP is clear, we may want to collect materials in CP into one "phrase" or a syntactic "word" in the flattened-tree sense of this section, and collect materials in TP into another "phrase" or "word". In this case the CP "word" contains no root – but in this case, the CP "word" quite likely consists of purely grammatical items, and therefore is irrelevant to the discussions here.

case inflection appears on the article, and the article-plus-noun combo looks quite like the English *have ...been ...reading* and fills a cell in the nominal inflection paradigm. Besides definiteness, the nominal inflection paradigm should also include the category of case, which marks the grammatical relation of the noun phrase. In Japhug, case is marked by independent words (Jacques 2021, p. 8.2.1), and although considering the case marker and the noun stem as one word sounds reasonable for speakers of German, Latin, or Sanskrit, it sounds absurd for Japhug speakers. What is language-specific here is actually the definition of *morphological* wordhood, i.e. what formatives in a syntactic "word" in this section are actually brought together by post-syntactic operations (§ 2.3.2).

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*, which are however trees *with holes*: [$\sqrt{play}$, n] or [$\sqrt{play}$, 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).

Therefore we have already seen two criteria for the derivation/inflection distinction above: the difference between § 2.2.1 and § 2.2.2, and the distinction between the inner and outer parts of morphology. We can propose yet another criterion: what is subject to stronger idiomization should be considered derivation.[13] There is certainly correlation between the three criteria. Idiomization works in a bottom-up manner, so more external grammatical categories are less likely to be idiomized, and hence we expect to see non-idiomized grammatical categories (tense, aspect, etc.) surrounding idiomized grammatical categories (valency, etc.), just like how any word defined according to § 2.2.1 is a part of another word defined according to § 2.2.2. We expect hierarchical syntax to at least be partially reflected by the morphology, and hence both the difference between § 2.2.1 and § 2.2.2 and the difference between degree of idiomization should lead to the morphology being divided into inner part and an outer

[13]There are all kinds of other criteria proposed to distinguish derivation from inflection. Some say that derivation is less productive. This is a direct conclusion of our idiomization-based definition. Others say that derivation is more messy. But inflectional can be messy too, involving several categories, some of whose values are prohibited for various reasons. The Japhug tense-aspect-modality-evidentiality-speech force complex is a good example: the values of the grammatical categories cannot be put together compositionally; rather, we have 11 allowed configurations (Jacques 2021, p. 1082). Emphasizing that derivation is messy is another way to say that there are so many fossilized forms – which is equivalent to our idiomization-based definition.

Yet another definition says derivation is more recursive than inflection. True, but this seems to originate from the structure of the syntactic tree, whose lower parts allow more low-level recursions (like putting a vP into another vP), and grammatical markers in these recursive constructions can be put together into a morphological word. So this distinction is not the same as all the three criteria proposed in the main text, but can be explained in similar structural terms.

part.[14] But the correlations are clearly not strong enough for us to expect that they should draw the same line between derivation and inflection.

The derivation/inflection distinction therefore is not primitive and should be reduced to a set of not necessarily converging criteria – in the same way wordhood is deconstructed. Since the derivation/inflection distinction is relevant to the definition of the **stem**, and equivalently, the definition of the **lexeme** under the assumption "derivation creates new lexemes while inflection turns a lexeme into a word", the two concepts cannot be defined cross-linguistically.

Note that the term *lexeme* does have a universal definition if we give up the expectation to see a clear-cut derivation/inflection: whatever in List B (see § 2.3.1) is a lexeme, that's to say, whatever form with its morphosyntactic profile – including the morphological realization and the syntactic environment, like valency – specified in the lexicon is a lexeme. This definition however does not assume that there exists a "lexeme" stage in the grammatical machine of a language. For instance, it is possible for a language to have "half-nouns", which participate in nominal derivation and have particular preferences of being either the first or the second constituent in a compound (Di Sciullo 2005; Scher and Nobrega 2014). Neo-classical roots like *geo-* ('earth') in English clearly satisfy these conditions, and given that they have morphosyntactic properties that are not inferrable from general principle of linguistics and can only be stored in the lexicon, they're clearly lexemes. But they are distinct from other lexemes in that "they can't be used independently" – another way to say that they can't be noun phrases on their own.

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,

[14] Note, however, that this is not universal: in Athabaskan languages, for instance, in the prefixal chain, derivation prefixes come *before* inflectional prefixes. This can be explained by assuming that the verb stem undergoes certain movements (Rice 2000, p. 78).

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 sketch 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 contains all abstract grammatical atoms: roots and grammatical categories. They are the input to the machine of grammar.

List B guides the morphological realization of a construction – which provides food for the following phonological operations. The two steps are therefore collectively known in generative syntax as the phonological component of grammar. 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. Note that the term *syntactic* here means *abstract syntax*. It is not until post-syntactic realization that some syntactic phenomenon 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!), which is how we capture English subject-auxiliary inversion and similar phenomena. 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 pieces in the syntactic structure (roots or functional items) phonological forms, turning them into substantive **exponents** (5).[15] 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).[16]

(5) 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

[15] Some 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.

[16] Non-concatenative morphology is captured by stipulating an abstract morpheme, like *template* in Tucker (2011). This is kind of like incorporating diachronic analysis into synchronic morphological analysis: historically there might have been affixes that cause umlaut in the root, and now we stipulate a ghost morpheme to reproduce the reflex of its effects. Some may argue that this is artificial and human brain merely does pattern recognition and builds schemas of word structures. True – but *what kind of* schemas are allowed? If the answer is "possible reanalyses of previously concatenative ones", then the approach we advocate for here has nothing substantially different from the word schema approach.

in the mental dictionary of English speakers. Thus **eat (n.)* In the same way, subcategorization i.e. valency is specified by List B as well: that a verb is transitive can be captured by stipulating that the root in that verb only pairs with a "transitive" feature TRANS in List B, which introduces the object into the syntactic tree, and the root appearing without that transitive feature is not an entry in List B.

List C is about meaning. It contains constructions varying 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: what is being lexicalized does not have to be a word (Harley and Noyer 1999). 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). Therefore, it can be a hotbed for *syntactic* fossilization, i.e. graduate 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: syntactic 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.

It is possible that List C contains some half-finished trees with holes in them:[17] this is how we capture English prepositional verbs and verb-particle constructions. Note that certain inconsistencies between entries List B and List C are to be expected. It is possible that an abstract syntactic tree is realized in an irregular way but is interpreted in a compositional way. Example: *went* is the collective realization of [*√go*, PAST], but the only idiomization in its interpretation is the collective interpretation of *[v √go]* as the action of going: there is no idiomization regarding how PAST is interpreted. On the other hand, an abstract syntactic tree can be realized in a perfectly regular way but undergoes idiomization, as in *kick the bucket*. We do not consider everything stored in the lexicon to be perfect, transparent form-meaning pairs.

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 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

[17] Including, but not limited to, syntactic "words" in § 2.2.2: half-finished trees may contain "finished materials" in § 2.2.2 as well, like *the bucket* in *kick the bucket*.

are left on their own. A "word" defined in § 2.2.2 is therefore realized as one or more morphological words.

In a sentence: *what are moved together in phonological realization form a morphological word.*[18] 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 (think about bubble sort), some of which may have phonological motivations, can reorder the exponents; syntactic movements of functional heads is also a possible explanation (Bye 2020).[19] The emergence of domains, whose boundaries are marked by (morpho-)phonological rules, in the morphological template can be explained by some "slots" being auxiliaries with affixes attached to them (McDonough 2008). 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

[18] In 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.

[19] What seems more problematic to syntax-oriented theories of morphology like Distributed Morphology is the existence of extended exponents: a syntactic feature being realized by several pieces, the position and function of each having no clear syntactic motivation. This however can still be captured by assuming fission of a morphological feature (Bobaljik 2017).

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 morphological word. It definitely looks like a morphological word, but no post-syntactic relocation of abstract or concrete morphological pieces is necessary to generate it.[20] 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 be reanalyzed as an authentic morphological template, just like how personal pronouns evolve into agreement markers (§ 3.1.3). The possibility of accidental fixed linear orders is also relevant to the problem whether a sequence of clitics with a fixed order should be considered as a part of the morphological word they attach to – 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 *lately* incorporated into the verb. Latin *=que* is also attached to things when everything else is formed, so it is attached to nouns or adjectives *lately*.

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

[20] To be fair, if a content word ahs 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.

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 **endoclitic-cization**), and their positions of clitics in the verb morphological template are subject to the control of other formatives (Harris 2000).

Once the boundary of pre-formed morphological words is not respected by 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 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. Even within one language we can have several standards of phonological wordhood, longer ones of which may be named *phonological domain/phrase*. Below we are using the term *phonological word* casually most of the time: a phonological word means a phonological unit that is not too long.

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]$_{\text{temporal}}$ [– biǎn [Cháoyáng]$_{\text{locative}}$]$_{\text{verbal predication}}$ [lù bā qiān]$_{\text{non-verbal predication}}$

evening relegate PLACE road eight 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 stem. But it is also possible that the morphological word is a compound and the two phonological words from this morphological word contain the two branches of the compound (Dixon 2010, p. 23). And nothing prevents the split of a morphological word into two phonological words, each containing a half of the stem, which gets split in the process. We discuss this issue in more details in § 2.4.3.

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. These phonological words lack clear morphosyntactic statuses.

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 due to sandhi. In Fiji, the form *a isele* is phonologically adjusted to *ai sele*, making the first phonological word contain a morphological word and a segment of the second morphological word (Dixon 2010, p. 25). If *nuncle* was the older form, and the English article *a* is usually stressed, then *an uncle* from *[a nuncle]* would illustrate exactly the same process.

Actually, (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. At the end of § 2.4.1, we also mention the possibility of a long morphological word breaks into multiple phonological words, which do not necessarily have morphosyntactic significance.

A root getting cut in halves by phonology is much less common. However, with some morphology, this is not impossible. I've heard some rather interesting forms from Mandarin Chinese speakers living in the US. The form *de-bu-chūlái-bug* 'can't get debugging done' is probably formed by first placing the English verb *debug* into the Mandarin verb form *V*-不出来, where 不 *bù* is a negative suffix and 出来 *chūlái* is originally a directional verbal complement but later grammaticalizes into a completive marker. In casual speech, when *V* is disyllabic, the suffix chain can be infixed between the two syllables of *V*. This is probably due to the reanalysis of a verb phrase like *[V-bù-chūlái Object]* and its simpler version *[V Object]* into two morphological words where the object is simple in its linear size:[21] the suffix chain is then recognize as an *infix* chain. Now, *debug* does not have a verb-object internal structure, but it is a disyllabic verb, so infixation happens, and we we get the rather unusual form *de-bu-chūlái-bug*, in which *de-bu*, *chūlái* and *bug* form three prosodic 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 like valency alternation also known as derivation in some contexts (§ 2.2.3).

[21] It is possible that *[V-bù-chūlái Object]* contains more than one heavy-light prosodic words (§ 2.4), but we can still say that it forms a prosodic domain, and hence a certain type of phonological word when its size is small, so this doesn't violate our generalization in § 3.2.1. This is an instance of how more than one type of small phonological units can be defined in a language.

- 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 (§ 2.3.1).

- 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 the underlying syntactic constituency, without nay post-syntactic morphological operations (§ 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, reanalysis of the sequence into a morphological word 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 and evolution (that is to say, units that are slightly distorted within their boarders by language change; cf. genes in biology) 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 complicated clause is likely decomposed into pieces when transmitted.[22] 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.

Moreover, language change involves both phonological and morphosyntactic changes. Laws governing phonological changes have been listed in § 1. Laws governing morphosyntactic changes 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.

Most, if not all, language changes can be explained by a combo of these mechanisms.[23] Besides regular Neogrammarian sound change laws and borrowing, all the

[22] Some 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.

[23] An unfortunate but prevalent tendency in linguistics is people using all sorts of buzzwords without clearly defining what they are talking about. Dixon (2009) criticizes the problem within the generative circle, but the problem seems to be much worse in certain "usage-based" communities, for a formalist, with ample time, is usually able to tell explicitly and clearly their theory of how a certain form is produced, while this is less true in certain schools of usage-based theories. For instance, all kinds of allegedly special properties have been attributed to grammaticalization, sometimes to support the idea that the grammar of a language consists of constructions *without analyzable internal structures*. On the other hand, Campbell (2013, pp. 284-285) notes that most, if not all, instances of grammaticalization can be reduced to combinations of phonological evolution and reanalysis. It seems as if some researchers first have a rather vague idea of grammaticalization, and then say "this is grammaticalization, that is grammaticalization", without any attempt to write down a list of possible mechanisms of language change. Another buzzword is *frequency effects*, which, to our best knowledge, does not go beyond the synchronic grammatical framework we sketch in this note: it's surely possible that a speaker's mental lexicon contains two competing analyses of the same syntactic object, one idiomized (but still with analyzable internal structures), another completely compositional. Showing that frequency effects exist is theoretically neutral and says nothing about how language works in brain.

The same problem appears in language acquisition, where the fact that children seem to first learn chunks of frozen phrases without clearly internal syntactic composed structures. We agree that chunks are often learned as a whole – which we will also discuss in the discussions below – but chunking says nothing about whether analyses on the internal structure are being done. Children may be trying to build their internal grammars soon after they learn some chunks from grown-ups. A statistic analysis of their language outputs actually *is* consistent with this hypothesis. Note that child language *is* supposed

procedures look like one type of analogy or another. We have laws for analogy as well, but they are more sporadic: a sound change law applies to all words that satisfy its conditions, but an analogy law does not need to apply to all constructions that satisfy its conditions.

3.1 Types of reanalysis

3.1.1 Purely syntactic reanalysis

Reanalysis can work on quite large structures. Therefore the unit of transmission can be much larger than what we commonly conceive as words, or does it? Before continuing our discussions on units of transmission in historical linguistics, we need to first have a look at examples of *syntactic* reanalysis, and see whether they truly challenge our assumption above that the basic unit of transmission is something in the list in § 2.5.

Here is a good demonstration of the idea of reanalysis running on the level of embedded clauses. In the history of Finnish, a sound change *-m* > *-n* makes accusative singular and the genitive singular endings homophonous, and this eventually leads to reanalysis of something like *I saw [him [doing sth.]*$_{\textit{relative clause}}$*]*$_{\textit{object: noun phrase}}$ into *I saw [his doing sth.]*$_{\textit{object: gerundive clause}}$, creating a new complement clause construction (Campbell 2013, pp. 275-276).

It turns out that syntactic reanalysis in this example happens *silently*: it creates no new forms on the surface. This should be *expected*, as we do not expect syntactic reanalysis to create forms that are not acceptable at all in the parent language: it is *re*-analysis, and the two analyses pertaining to the same surface form must both be valid when the reanalysis happens. The only surface form change happens because of a regular sound change applied to word endings, which *triggers* the reanalysis, but is conceptually independent from it.

In this example, we may say that the whole clause structure – with all the hierarchical syntax – is the unit being transmitted. But the change has a single locus: the genitive ending, which now has the function of labeling the subject of a gerundive clause.

In grammaticalization of auxiliaries, the starting point is a complement clause construction, and the destination, i.e. an auxiliary construction, has an identical surface form. We can say that the unit being transmitted is the whole clause – which is true. But again we can identify a single locus of change: the complement-taking verb loses its status as a content word. Apart from that, the starting point and the destination even have comparable constituency structure.[24]

If we inspect the two examples give about, we will notice that what are being transmitted are actually *idioms* in the broader sense (§ 2.3.1). A complement clause construction *is* an idiom in the broader sense because it is lexicalized and is interpreted as a whole with a coherent meaning, which undergoes gradual shift and may one day be analyzable as having purely a tense or aspect or modal meaning. And then reanalysis of the main verb as an auxiliary happens. Compositionality, contrary to idiomization,

to seem to be impoverished and lack diversity due to the limited vocabulary – but its statistic distribution is still far from what is expected from a simple memory-based model (Yang 2013).

We aim to avoid buzzwords without theorization in this section.

[24]Huddleston and Pullum (2002, p. 65) indeed calls complement-taking verbs *catenative verbs*, and auxiliaries are just a special case of catenative verbs.

only hinders reanalysis: the rigid hierarchy of auxiliaries, the change of the positions of adverbs, etc. tell an auxiliary construction and a complement clause construction apart (§ 2.1.2). The same line of argument works for the Finnish complement clause construction.

Let's go for more examples. The Latin particle *etiam* probably was formed along the line of $[et\ [iam]_{focused}\ [\ldots]_{nucleus\ clause}]_{coordinate\ clause} > [etiam\ \ldots]_{clause}$. Again, the surface form is never lost in the reanalysis. What was being reanalyzed – the sequence *et iam* – is again an idiom: the structure *[et [iam ...]]* 'and now, ...' probably gradually gained the meaning of 'and also, ...' or 'besides, ...' and then reanalysis happens. Note that strictly speaking, this example involves morphological reanalysis as well, because a new formative without clear internal structures, i.e. *etiam*, is created (§ 3.1.3).

Similarly, when a verb root and a voice marker attached to it are being reanalyzed into a synchronic root, we notice that the surface form is not changed, and what are being reanalyzed is an idiom. Numerous synchronic verb roots in Modern Mandarin with verb-object internal structures (which however is irrelevant in argument structure alternations, proving that they have *syntactically* fossilized), like 关心, can further prove the point. Or consider *whodunit*, from *who's done it*, an idiomized question representing the topics of many detective novels, which then is recategorized into a noun, and soon is used to refer to a variety of detective novels.

We can summarize our observations above into the follows. First, the unit of transmission in syntactic reanalysis, strictly speaking, is an idiom: the whole construction can be arbitrarily large, but the active region of reanalysis is always an idiom. Second, usually the surface form of the idiom is not changed during reanalysis. Third, if we want to identify a single locus of reanalysis, i.e. summarize the reanalysis into the function change of a single item or the creation of a new form, then the locus is usually quite small in its phonological size: it can be a case suffix, a morphological word, or the emergence of a small particle like *etiam*.

Therefore, the unit of transmission in syntactic reanalysis falls within the definitions of wordhood in § 3: the syntactic construction being alternated is always an idiom, i.e. a lexicalized syntactic object, which can be big or small in the size of its abstract syntactic tree, while the *locus* of change is usually something small in its phonological size, possibly a phonological word. We will go back to the problem of the phonological size of the locus of reanalysis in § 3.2.1.

3.1.2 Morphological reanalysis

In § 3.1.1, we have discussed syntactic reanalysis which keeps the surface form of a construction while changing its underlying syntax. In this section, we consider morphological reanalysis, which does not change the underlying syntactic structure but alternates the forms of formatives in morphological realization of a construction due to *rebracketing*. Morphological reanalysis sometimes combine several formatives into one, or more rarely leads to backformation, which, strictly speaking, already involves syntactic reanalysis (§ 3.1.3). Pure morphological reanalysis is mostly about moving the boundaries of formatives in a given form, or turning a particle or a clitic into an affix, or vice versa (possibly, although rare; Campbell 2013, pp. 251).

In English we occasionally see rebracketed noun phrases. Thus we see *a nuncle*, which is a result of re-bracketing *an=uncle*, the phonological realization of a simple noun phrase with the underlying representation $[a_{SINGULAR,INDEFINITE}\ [uncle]_{noun\ stem}]_{noun\ phrase}$,

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*. Another example is the emergence of *nickname*, which was formed by dropping the initial *a* in *an eke name*.

The main difference between formation of *nuncle* or *nickname* and processes in (§ 3.1.1) is that in the former, the underlying syntax is kept the same: the only thing that changes is the morphological exponent of the root meaning 'uncle'.

The exact same process appears in other languages as well. French *licorne* is formed by double rebracketing: *unicorne* was probably resegmented as *un icorne*, giving Old French *icorne* $\overset{\text{article mod.}}{\longrightarrow}$ *l'icorne* > *licorne* 'unicorn' (Alkire and Rosen 2010, p. 305).

Turning to the question of the unit of transmission. What is being transmitted and undergoes re-bracketing here is the whole sequence *an=uncle*. Note that here boundaries of morphological words do not matter: *an* is not a morphological 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. Therefore, morphological words are not the only units of transmission and evolution.

Further, it is highly unlikely that a construction with an arbitrary size can be morphologically reanalyzed: we have never seen any English nominal idiom starting with a vowel acquiring an initial *n*-. Like the case in § 3.1.1, there is likely a locus of morphological reanalysis, which is either some sort of phonological wordhood or syntactic "wordhood". The latter, intuitively, is not feasible, as is discussed in § 3.4.1. Thus the size limit of the construction being morphological reanalyzed seems to be related to phonological wordhood (§ 3.2.1).

3.1.3 Two types of reanalysis happening together

Besides the emergence of *etiam* in § 3.1.1, a good example of morphological and syntactic reanalysis happening together is how pronouns become agreement markers. A minimal full clause, consisting only the main verb and its arguments in the pronoun forms, may be idiomized to a certain extent (§ 2.3.1), which is later reanalyzed as a morphological word and the personal pronouns are reanalyzed as agreement markers, meaning that the personal markers stop being clitics and start to be affixes, which can appear even when noun phrase arguments are present. In this case, we see syntactic reanalysis (which often turns a topic-comment construction into a subject-predicate construction: $[Mary]_{topic}$, $[she\ likes\ it]_{comment:nucleus\ clause}$ > $[[Mary]_{subject}$ *she-likes-it*$]_{nucleus\ clause}$), and we also see morphological reanalysis which gives us new personal agreement affixes. Note that additional phonological changes happening to the new verb may alter the form of the personal agreement affixes.

3.2 Processes based on the phonological word

3.2.1 Phonological word as the origin of size limit in reanalysis

In § 3.1, we have noticed that in syntactic reanalysis and in morphological reanalysis, the locus of reanalysis all seems to have a size limit, and that size limit cannot be easily determined in morphosyntactic terms: *et iam* is the phonological realization of the fixed part of an idiom, *whodunit* originates from a whole question, the precursor

of an auxiliary is a morphological word, *nuncle* emerges from reanalysis of *an uncle*, which is a full noun phrase. The syntactic statuses of these reanalyzed forms strongly vary, and the only choice left to us is to assume that the size limit comes from the requirement that a reanalyzed unit should be a phonological word of some kind.

3.2.2 Reanalysis of phonological words without clear morphosyntactic structure?

The next question is, can all phonological words be morphologically reanalyzed? In all examples given about, the phonological word in question has a certain morphosyntactic status, which means they contain either one or multiple complete morphological words. But we have phonological words that contain multiple morphological words but do not have any reasonable morphosyntactic status on their own: consider *to=the=fat* in § 2.4.2, which hardly is an idiom. Furthermore, the phonological word *to=the=fat*, although having no morphosyntactic relevance, at least contains three morphological words, no more, no less, just like *an=uncle* or *who's=done=it* which do have morphosyntactic relevance, but in principle, a phonological word can contain a morphological word and a half, or half a morphological word (§ 2.4.3). Can these phonological words without clear morphosyntactic statuses be reanalyzed?

Deciding whether a phonological word without morphosyntactic significance being reanalyzed is hard when the phonological word contains the head stem of a construction (§ 2.1.3), because of the same reasons mentioned in § 3.4.1: if we have materials after *uncle* in *an uncle* (as in, say, French), then the phonological word *an=uncle* has *no* morphosyntactic status. If a phonological word like this is never reanalyzed, then we have refuted the possibility that a phonological word without a clear morphological status can be reanalyzed. However, a phonological word like this being successfully grammaticalized doesn't mean historical linguistic laws work on phonological words without morphosyntactic statuses: *[an uncle]* appears more frequently than *[an uncle who drives poorly]* or similar forms, so we can't be sure that *an uncle* is truly being reanalyzed in *an uncle who drives poorly* or similar forms.

On the other hand, a phonological word like *to=the=fat*, which has no status in morphosyntax and does not contain the head stem of the construction it is in, usually does not get reanalyzed. In English, rebracketing of a noun phrase always produces new nouns, and so is rebracketing in Romance languages. The most likely explanation seems to be that a phonological word has to have some sort of morphosyntactic significance in order to undergo reanalysis: thus *an=uncle* gets reanalyzed, and *to=the=fat* does not.[25]

Let's turn to phonological words that contain segments of morphological words, and start with a phonological word containing a segment from a morphological word nearby (§ 2.4.3). It has been proposed that such a phonological word may be passed to daughter languages as a morphological word, which may alternate the form of the root. Some have argued that the mysterious *s*-mobile in Indo-European languages, i.e. an initial **s*- appearing sometimes in Proto-Indo-European roots and sometimes doesn't, is due to the final consonant of a nominal ending drifting to the start of the main verb following it because of phonological sandhi (Sihler 1995, p. 169). The phonological word consisting of the main verb and the initial **s*-, strictly speaking, has no syn-

[25]The linear sequence *to=the=fat* is not the fixed part of any idiom; on the other hand, *et iam* is a fixed part of the idiom *[et iam ...]*. Thus the latter can be reanalyzed into a single word (§ 3.1.1).

chronic morphosyntactic significance, but gets reanalyzed as a morphological word anyway. If this is true, we probably need to slightly modify the condition of phonological words being reanalyzed: a phonological word that can be syntactically and/or morphologically reanalyzed should look like a morphosyntactic unit *in parsing*.

It is indeed possible that a phonological word containing only a segment of a split stem is considered a full morphosyntactic unit. This is known as *clipping*. Note, however, that clipping is sometimes regular morphology: a large number of clipped forms can be analyzed using template morphology, which takes the first two syllables of a lexicalized form and possibly make sure the second syllable has vowel *i*, (Bat-El 2019), disproportionally creating nouns with a diminutive meaning, like *Iggy* < *Ignatius* or *veggie* < *vegetable* (Jamet 2009), and sometimes even adds phonological materials into the word (*duckie* < *duck*). This type of clipping indeed works on a type of phonological words (namely, disyllabic prosodic words) without considering the internal morphosyntactic structure of a word, but it is not reanalysis or any other diachronic change: it is *prosodic morphology*, i.e. morphology that invokes non-trivial phonological processes. Another diminutive template allows only one syllable (*math* < *mathematics*).

It seems that it is generally hard to find instances where a root was clipped because of synchronic phonology (not impossible: § 2.4.3), and later in a daughter language, only one clipped half appears. There seems to be an tendency to transmit the root as a whole. The reason probably is that a phonological word containing only half of the root does not look like something with a morphosyntactic status as all, and will likely not be "misread" as such. So the reanalysis process will never start.

A final question is whether a phonological word made solely of affixes, clitics, etc. that gets detached from a long morphological word (see the Moloko example in § 2.4.1) can be transmitted to later stages of the language. This is also related to how a long morphological word is transmitted (§ 3.3.2). TODO

3.2.3 Sound change laws

It seems Neogrammarian sound change laws have inputs from both phonological and morphological wordhood. French liaison is a perfect example. Liaison means a linking of the final consonant of a morphological word to the initial vowel of the next morphological word, and it happens obligatorily between determiners and the following word and between clitic pronouns and the verb. We note that in both cases, liaison happens within one phonological word. In environments prohibiting liaison, the final consonant of a morphological word is dropped. Evolution of liaison therefore involves both morphological word boundary ("drop the final consonant of a morphological word, unless instructed otherwise"), and phonological wordhood (which also explains why liaison between a polysyllabic preposition and the word after it is rarer).

Now we ask the question again: are all phonological words eligible for sound change laws? This is a hard question to answer for liaison, as liaison today does not happen in all syntactic environments: thus no liaison is possible between *et* and the word after it. This can be explained by stipulating that *et* actually did *not* form a phonological word with the word following it in some historical periods (by assuming that *et* had a special underlying phonological form or something like that), or by assuming that today's liaison has been re-morphologized and no longer transparently reflect historical sound change. Both hypotheses seem to suggest that unlike the case

of morphological reanalysis, regular sound laws are purely phonological and impose no constraints on the morphosyntactic status of phonological words they apply to. But this question probably isn't particularly meaningful in determining the unit of transmission: it the sound laws apply blindly to all phonological words, then they don't get to decide what gets transmitted and what doesn't.

3.3 The role of morphological words

§ 3.2 makes one feel as if *morphological* reanalysis solely targets *phonological* words now. In this section, we discuss what is the role of morphological words without phonological status.

Latin *=que* doesn't leave any traces on the evolution of nominal declension of Romance languages. The reason is probably because *=que* is too volatile: it may appear or it may not, causing it to be forgotten in the evolution of Latin into Romance languages. In this case, what is transmitted is a morphological word, which is not necessarily a phonological word. This fact alone however cannot prove that some historical linguistic laws targeting morphological words and not phonological words, because the *=que*-less form of a noun is also a phonological word, which, by virtue of being identical to a morphological word, was transmitted. To demonstrate that morphological wordhood has a role *independent* of phonological wordhood, we need to look for processes targeting morphological words that can't be explained by processes in § 3.2.

3.3.1 Reordering of formatives

It seems the shift of morpheme positions can only be explained in terms of morphological wordhood, because there is no conceivable processes targeting the phonological word, like morphological reanalysis (§ 3.1.2) that results in a reordering of formatives. On the other hand, formative reordering seems natural from the *production* side: for instance it may happen because of analogy to positions of other affixes with similar functions (Campbell 2013, p. 252).

We should however always remember that it's possible that a formative that seems to have different positions in related languages because of different paths of grammaticalization: it's possible that the formative's original form in the parent language was not an affix at all.

3.3.2 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? These morphological words usually break down into several phonological words in actual utterances 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. TODO

3.4 The paradigm and syntactic "wordhood"

3.4.1 Morphological change without linear order?

We want to first argue that morphological changes do not specifically target generic syntactic "words" as defined in § 2.2.2 whose formatives do not necessarily form a linear sequence.

When two formatives in one syntactic "word" are not adjacent to each other in the surface form, we have no reason to expect that the two influences each other according to processes in § 3.1.2 or according to conditional sound change laws. Therefore, the only possible diachronic alternation that directly targets a flattened-tree syntactic word without the necessity for the formatives to be in one place in the surface form is something like a particle becoming an affix, something in the form of *[A … Σ] > […A-Σ]*, where *A* is a grammatical marker of a construction headed by *Σ*. But this can also be explained from the perspective of linear order: modifications in the construction headed by *Σ*, by definition, do not obligatorily appear, and it is possible that *[A Σ]*, being the most frequent surface form now, possibly forming a phonological word *[A-Σ]* (§ 3.2.1), is the unit reanalyzed.

So we see there is a lot of noise here. If *A*'s transformation into a prefix is always accompanied by *A* and *Σ* being a phonological word, then what historical linguistic laws work on is phonological words with syntactic "wordhood", not all syntactic "words". If we want to demonstrate that the flattened-tree syntactic "word" defined in § 2.2.2 is a unit that historical linguistic laws work on, we need to demonstrate that *A* and *Σ*, being one flattened-tree syntactic "word", can evolve into a morphological word in one step, even when the latter is never attested in the parent language. For instance, a change in the form of *[A … Σ] > […Σ-A]* (note the reversed order) is a convincing piece of evidence suggesting that historical linguistic laws work on flattened-tree syntactic "words", because no explanation on surface linear orders is possible. Yet *[A … Σ] > […Σ-A]* seems to be rare, if not impossible. An instance of this is the imaginary evolution of a *preposition* (which may be separated from the head noun by adjectives, etc.) into a case *suffix*. This trajectory seems to be never attested.

This is probably to be expected, as we do not expect a form unacceptable in the parent language to emerge overnight in language evolution, even when the syntactic derivation of the form is quite similar to the underlying morphosyntax of the parent language.

3.4.2 Changes to cells in the inflection paradigm

Some language change processes do seem to be about the syntactic "word" defined in § 2.2.2, processes where a cell in an inflection paradigm is filled by something else or where a new inflectional form is added. The emergence of a periphrastic form can certainly be seen as a process targeting a syntactic "word", especially considering that the final state is not a phonological or morphological word. Often, this process is the same as the grammaticalization of the auxiliary (§ 3.1.1), where the synchronic "word" plays the role of the background of the grammaticalization, and the locus of language change is the auxiliary, but extension following the appearance of the auxiliary creates linear forms that were originally not possible, and this process is no longer the same as the grammaticalization process. For instance, after the emergence of the English PROGRESSIVE aspect, *it is raining* is added into the paradigm of the zero-place verb

rain. Previously the whole form was not possible, and the locus of language change here is the emergence of the whole form, not just the auxiliary.

Note, however, that nothing is *transmitted* in extension: extension makes use of transmitted and reanalyzed materials to build new forms. So if we ask what the transmitted units are here, the only answer is the auxiliary, the nonfinite forms of the verb, etc. Not the syntactic word.

On the other hand, it's reasonable to analyze suppletion as a morphological reanalysis. There, what is changed is how an underlying representation of a syntactic "word" is realized. What is being transmitted is an inflected word.

3.5 Dropping grammatical markers

TODO: under which conditions are grammatical markers dropped?

3.6 Summary: patterns of change observed

A list of possible language changes is given in Campbell (2013, Chs. 10-11). Based on the discussions above, we classify these operations into the following categories:

- In *syntactic* reanalysis (e.g. grammaticalization of auxiliaries, function change of an affix, combining two particles into one), the basic unit of transmission is an idiom, and the locus of change (with our without changes to the surface form) is usually a phonological word (or a "prosodic domain" with usually a small size) with a certain morphosyntactic status, often the fixed part of the idiom (§ 3.1.1, § 3.1.3).

- In *morphological* reanalysis (e.g. shifting formative boundaries, combining two particles into one, backformation), the basic unit of transmission is usually a phonological word with a certain morphosyntactic status, which can be a fixed part of an idiom, a morphological word, or a phrase or even a clause (§ 3.1.2, § 3.1.3). The locus can be as narrow as an affix. Morphological reanalysis typically makes parts of the unit being transmitted more tightly bound, but the opposite process, like reanalysis of an affix as a particle, is possible.

- Morphological reanalysis may occasionally happen to phonological words that actually have no morphosyntactic statuses on their own, but appear as if they are morphosyntactic units. This is a possible origin of the Indo-European *s*-mobile (§ 3.2.2).

- Regular sound change laws have access to both phonological wordhood and morphological wordhood (§ 3.2.3).

- Certain processes, like morpheme reordering, are not based on phonological words with morphosyntactic statuses (§ 3.3.1), and are likely based on morphological words, which may or may not be phonological words.

- Creation of periphrastic forms, whose surface forms were not previously attested, should be attributed to processes targeting syntactic word. This however is extension and makes use of earlier transmitted materials (§ 3.4.2). Extension can also happen at the level of morphological word.

Therefore, among the definitions of wordhood in § 2.5, idioms and syntactic "words" are either backgrounds of language change, or playgrounds for reused materials to form new forms, but are typically not the units being passed as a whole. Instead, the direct units of transmission are phonological words with morphosyntactic statuses (morphological words, idioms, phrases, clauses, etc.) and morphological words possibly not necessarily with phonological statuses.

What are *not* the unit of transmission are abstract entities, i.e. roots and grammatical formatives, which are to be inferred by each generation of learners from the transmitted units, and are not directly available in discourses. As an instance, in the reconstruction of Proto-Afro-Asiatic morphology, the unit being considered is the morphological word (Wilson 2020, p. 21).

The unit of transmission sketched above may prompt people to argue that it is the only natural unit in a synchronic theory of grammar (as diachronic laws should eventually find their origin in synchronic psycholinguistic mechanisms), and hence, it is probably a good idea to adopt, say, a flavor of Construction Grammar, as a better synchronic linguist theory. It is however worth noting that generative syntax does not deny the role of surface forms, rather, it assumes that as soon as a surface form is coined and received, it *immediately* gets reanalyzed according to the root-plus-functional head framework above. The theoretical difference between generative syntax theories and Construction Grammar theories therefore lies in *how analogy happens*. Given the often sporadic and limited materials we have at hand, verifying which of the theory families is more reliable is typically not within the ability of historical linguistics.[26]

3.7 Synchronic and diachronic rules on words

From a methodologically minimalist perspective, it is therefore safest to assume no relation between diachronic and synchronic rules in word formation. That said, certain commonalities can clearly be found: they're for instance typically formed in the same rewriting rule formalism and certain operations – like those depending on arithmetics – are equally impossible in both types of rules. Some correlations between diachronic and synchronic rules therefore should exist.

Some may stipulate that the set of possible morphophonological processes allowed by human language faculty can be obtained by applying attested historic linguistic rules to the simplest agglutinative morphology to explain the overlap between the two types of rules. This explains the overlap between the two types of rules, but itself is not clearly motivated. On the other hand, one may assume that historic rules come from learners only acquiring the products of certain synchronic rules while forgetting the existence of the rules themselves. In this way many, if not most, diachronic rules are just fossilized synchronic rules: only their products are transmitted from the older generation to the younger generation, which then get reanalyzed in a different way.

The synchronic rules developed by later generations of speakers to synchronically generate certain forms may (partially) rediscover now defunct earlier synchronic

[26] A related claim commonly seen in the usage-based literature is that language changes sometimes happen without language acquisition, and happen when adult speakers regard frequently appearing sequences as "chunks". We have criticized lack of theorization in certain works in this school in § 3. Plus, chunking in adult speakers can be easily analyzed as dephrasal derivation or similar processes, in which a complex structure "wears" e.g. new vP projections around it, and becomes a lexeme.

rules. Rules proposed to capture the templatic morphology of Semitic languages without stipulating the existence of non-continuous consonantal skeletons, for instance, may coincide with diachronic rules that actually happened historically that eventually derived the prototypical templatic morphology from a mostly concatenative starting point. But there is no guarantee that this necessarily happens.

4 Reconstruction based on content words

We rely on information from content words to identify regular sound correspondences in related languages. Usually we do so by identifying the stems and finally the roots first. Things like *an=uncle* > *a nuncle* that happen when a phonological word with a morphosyntactic status is transmitted and reanalyzed mess up the form of stems. When we have a brief idea of particles in the parent language we're trying to reconstruct, the increase of the burden of our workload is limited: when you know that the English indefinite singular article has two allomorphs *a* and *an*, and that we have *nuncle* and *nickname* but only *eke* and *name*, recovering *uncle* from *nuncle* and *eke-name* from *nickname* is easy. What is done here has nothing different from identifying a stable thematic vowel at the end of the stem which sometimes vanishes or is absorbed into the ending because of sound change laws – and this is to be expected, because the phonological word *an=uncle*, which is a minimal complete noun phrase, can be seen as a periphrastic cell of the English noun paradigm.

Formation of *nuncle* and absorption of the thematic vowel into the ending or the stem both largely preserve the structure of the root, although some phonological materials are appended to the root in the process. We also expect similar processes to cut off a part of the root. We however should not worry too much about irregular clipping of roots because once upon a time, a phonological rule broke a morphological word into segments and only one segment survives. Such incidents are not likely because the segments cannot be reasonably parsed individually, and hence are unlikely reanalyzed as full-fledged morphological words. Clipping happens more frequently because of (semi-)regular derivation devices (§ 3.2.2), or because of regular sound change rules (e.g. CVCV > CCV > CV – and half of the root is gone).

When we have insufficient knowledge about the syntax, identifying boundary shift of formatives can be hard. Indo-European *s*-mobile is probably the result of such a boundary shift, but its exact origin remains not completely clear even today. The irregular forms created by reanalysis however usually *do not* hinder the enterprise of determining phylogenetic relations between languages. Under the assumption that the language family in question can be described by the tree model, as long as we have *a* list of cognates from languages in the family in which there is no borrowed terms (which can be kicked out using standard techniques of Neogrammarian historical linguistics; Campbell 2013, p. 3.5), the comparative method can be applied. The tree model means we do not have to go over all cognates: a tentative evolution tree can be built by applying the comparative method to a sample of cognates that is large enough, and if adding more cognates into the comparison doesn't break the tree structure, we're all good. Thus, a root containing a mysterious *s*- at its initial can be temporarily removed from the list of cognates used in comparison, and this does not significantly weaken the confidence we have in the tree we get. The origin of *s*- or *n*- or clipped roots is to be discussed *after* the evolution tree is built. This is similar to what is done in

evolutionary biology: genes that are *not* particularly interesting are used to determine how genetically close different creatures are, and we can get to know how close they are *without* knowing more about how the creatures move, eat, digest, etc.

And this inevitably causes doubts on whether the tree model is correct all the time. Some may argue that a language have

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. Or maybe *B* hadn't been an affix until fairly recently.

References

Alkire, Ti and Rosen, Carol. *Romance languages: A historical introduction*. Cambridge University Press, 2010.

Bat-El, Outi. "Templatic morphology (Clippings, word-and-pattern)". In: *Oxford Research Encyclopedia of Linguistics*. 2019.

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.

Campbell, Lyle. *Historical linguistics*. Edinburgh University Press, 2013.

Cinque, Guglielmo. *Adverbs and functional heads: A cross-linguistic perspective*. Oxford University Press, 1999.

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

Di Sciullo, Anna Maria. “Decomposing compounds”. In: *Skase Journal of theoretical linguistics* 2.3 (2005), pp. 14–33.

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: *Glot 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. “Review Article: A comprehensive grammar of the English language”. In: *Language* 64.2 (1988).

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.

Jamet, Denis. “A morphophonological approach to clipping in English. Can the study of clipping be formalized?” In: *Lexis. Journal in English Lexicology* Special issue 1 (2009).

Matchin, William and Hickok, Gregory. “The cortical organization of syntax”. In: *Cerebral Cortex* 30.3 (2020), pp. 1481–1498.

McDonough, Joyce. “Athabaskan redux: Against the position class as a morphological category”. In: *Morphological analysis in comparison* (2008), pp. 155–178.

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

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

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

Reynolds, Brett, Arora, Aryaman, and Schneider, Nathan. "Unified syntactic annotation of English in the CGEL framework". In: *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*. 2023, pp. 220–234.

Rice, Keren. *Morpheme Order and Semantic Scope: Word Formation in the Athapaskan Verb*. Cambridge Studies in Linguistics. Cambridge University Press, 2000.

Schackow, Diana. *A grammar of Yakkha*. Language Science Press, 2015.

Scher, Ana Paula and Nobrega, Vitor Augusto. "Unifying neoclassical and stem-based compounds: a non-lexicalis approach". In: *Revista Linguíſtica* 10.1 (2014), pp. 74–98.

Sihler, Andrew L. *New comparative grammar of Greek and Latin*. Oxford University Press, 1995.

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

Wilson, David. "A concatenative analysis of diachronic Afro-Asiatic morphology". PhD thesis. University of Pennsylvania, 2020.

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

Yang, Charles. "Ontogeny and phylogeny of language". In: *Proceedings of the National Academy of Sciences* 110.16 (2013), pp. 6324–6327.