PART 5. CONTEXT FREENESS OF NATURAL LANGUAGES

REDUPLICATION IN MOHAWK (Postal 1964, Langendoen 1977).
(Mohawk: An Americal Indian language spoken in the area around Niagara.)

1. Noun incorporation through reduplication:

(1) *kaksa*?a ka *nuhwe*?s ne-ka*nush*a
girl likes house
The girl likes the house.

We can incorporate *house* into *likes*:

(2) *kaksa*?a ka **nush** *ø nuhwe*?s kik→ ka*nush*a
girl house-likes this house
The girl likes this house.

Similarly:

(3) *kaksa*?a ka*nuhwe*?s ne-ka*sheht*
girl likes car
The girl likes the car.

(4) *kaksa*?a ka**sheht** *ø nuhwe*?s kik→ ka*sheht*
girl car-likes this car
The girl likes this car.

Ungrammatical are:

(5) **kaksa*?a ka **sheht** *ø nuhwe*?s kik→ ka*nush*a
girl car-likes this house

(6) **kaksa*?a ka**nush** *ø nuhwe*?s kik→ ka*sheht*
girl house-likes this car

2. Nomimalization: verbs can nominalize to produce noun stems. Crucially, incorporated verbs nominalize, and such complex nominals incorporate:

kanush-øts→ries ⇒ *kanush-øts-rihsra*
house find house finding

kanush-øts-rihsra-nuhwe ⇒ *kanush-øts-rihsra-nuhwe?tsra*
house finding likes house finding liking

So we can find grammatical sentences like (7):

(7) *t kanush- øts→rihsra -nuhwe?tsra- hs→hsra- karat-tsra- yeri*
house finding liking evil-being praising is good

kik→ *kanush- øts→rihsra -nuhwe?tsra- hs→hsra- karat-tsra*
this houde finding liking evil being praising

This praising of liking finding a house being evil is good.

In semi-transliterated English, we find the following situation:

(8) a. The girl *house*-likes this *house*.
b. The girl *house-liking* likes this *house-liking*.
c. The girl *house-liking-praising* likes this *house-liking-praising*
.....

This is unbounded copying. The fact that the Mohawk equivalences of the sentences in (8) are grammatical in Mohawk is not is by itself not enough to show that Mohawk is not context free: the fact that L has a non-contextfree subset shows nothing about L: For alphabet Σ, Σ^* has intractable subsets, but is perfectly regular.
Langendoen 1977 gave the formal argument to show that Mohawk is not contextfree:

Let R = the girl house {liking, praising}* admires this house {liking, praising}*

Let L = the girl house α admires this house α
where $\alpha \in$ {liking, praising}*

1. L is not context free, since you can make a homomorphism with $\alpha\alpha$.
2. R is regular.
3. Postal's empirical claim: R $\cap$ Mohawk = L
(transliterated, of course)
Hence Mohawk is not context free.

The argument from Mohawk, the first argument in the literature, was challanged by Pullum and Gazdar 1982.
They point out that Postal mentions that Mohawk also contains **possessed-incorporation**:

(9) i?i k-*nuhwe*?s ne-ka*nush*-a
I like house

(10) i?i k-*nuhwe*?s ne-*sawatis-hrao*-ka*nush*-a
I like John's house

The element *hrao* is masculin-human inflection, which is particular to human possessors and is absent if the possessor is an abstract noun.
Possesed-incorporation incorporates the head noun of the possessive noun phrase into the verb, and the head noun drops in the possessive noun phrase:

(11)	i?i	**hrai-nuhs**-	*nuhwe*?s	ne-*sawatis*	e
	I	house-	like	John's	empty

Pullum and Gazdar point out the following. It is crucial to Postal?Langendoen's argument about non-contextfreeness that (12) is ungrammatical:

(12) *the man *house-praising-liking* admires this *house-liking-praising*.

But the fact is that (the transliteration of) (12) is **not** ungrammatical in Mohawk, it just has an interpretation that is less expected: we can interpret it as a possessive construction with the head noun dropped:

> The man house-praising-liking admires this house-liking praising**'s** house-praising-liking.

This may be semantically weird, but you can make up a fairy-tale story to deal with that problem, the point is: (12) is **not ungrammatical**.

Grammatically, we see: the **string:**

> The girl house-liking likes this house-liking.

is **ambiguous** between *this house liking* being the object noun phrase, and *this house liking* being the possessor of the object noun phrase, with the head noun (*house liking*) dropped. This means that, as far as the string set of Mohawk is concerned, the Postal-Langendoen argument **collapses**: It has not been shown that the string set of Mohawk is not context free.

What does this tell us about Mohawk? Let us think about **strong generative capacity**.
One would think that the **structure** of the sentence with the possessor reading will be different from the structure of the object reading. If one assumes that this means that the **rules** building possessive structures are the **rules** building object structures **plus** rules for the possessors, then you can argue that the first set of rules cannot be context free.

The problem is that this argument is highly theory dependent. In the first place, there is the general problem that it goes as follows: I tell you: the grammar of Mohawk cannot generate α with structure T in a context free way. And you tell me: well, you wouldn't **want** to generate α with structure T anyway.
It is very difficult to make arguments here that will be accepted cross-theoretically.

Secondly, the argument that the rule set of the possesor structures is an extension of the rule set for the object structures is also theory dependent.
It is based on the idea that you can **restrict** the grammar of Mohawk to a **sub**grammar that generates a **fragment** of Mohawk.
But this is, again, a theoretical claim that not everybody would be willing to endorse. In grammar formats where the grammar is though of as a set of grammaticality constraints (like principles and parameters), the **general rules** may well be such that you cannot **avoid** having possesors. That means that in order to get a possessorless fragment of Mohawk you would have to **add constraints to Mohawk**, and that means adding rules that aren't actually part of the grammar of Mohawk.

But then, you can get a non-contextfree fragment of Mohawk that way, but it has no bearing on the grammar of Mohawk itself.

Given this, it will not be possible to prove that Mohawk is not (strongly) contextfree to the satisfaction of all linguists. Nevertheless, I think it **is** possible to give an argument that relies only **minimally** on particular details of the grammatical analysis, and ought to be acceptable to anybody who accepts those minimal details, which, on might hope, would include 'the ordinary working linguist'.

Such an argument can be based on the notion of **surface structure string**.
This is a **grammatical** notion, but not one that depends on a lot of particular details about the grammatical framwork or analysis, in particular, it doesn't depend on their being a **level** of surface structure.

The idea is as follows:
1. Most grammatical theories assume that **case** is assigned by the grammar, by the syntax. This is called **grammatical case**.
2. Most grammatical theories assume that the morphology has access to the case assigned by the grammatical theory and realizes it as case morphology (where present), sometimes regularly, sometines irregularly. This is called **morphological case**.
3. Most grammatical theories assume that the morphology operates on the output of the syntax, which minimally includes strings generated by the syntax, **plus** features like case, gender, number features, etc.

4. If we assume, and this seems plausible too, that the morphology reads the yields of the syntactic trees, then this means that the grammatical features that the morphology needs to acces are visible on the yields of the syntactic trees. This means that **the yields of the syntactic trees are a bit more abstract than the morphologically realized strings.**

The yields of the syntactic trees that are input for the morphology I will call the surface structure strings. This forms itself a **language** in a bit richer alphabet which, in the case of Mohawk, I will call **Surface Structure Mohawk**.

Surface Structure Mohawk, then, is a more abstract language than the string set of Mohawk.

The fact that the possessed noun is empty in the possessor construction, means that is it empty after morphological spell-out. But, if we assume morphological spell-out, we can unproblematically assume that in the alphabet of Surface Structure Mohawk (the theoretical alphabet) there is a **symbol** *e*, which the morphology spells out as the empty string.
But that means, and this is the crux, that it is not very controversial to assume that **not only** does the grammar of Mohawk generates different syntactic structures for the ambiguous string of Mohawk in question (and anybody would assume that), **but also** that this ambiguity is actually **created** by the morphology: the syntax generates two distinct syntactic structures, and the distinction still shows up in the yields of the syntactic trees **before** the morphology operates.

Again, not everybody would agree to this, that is why you cannot argue to everybody's heart's desire that Mohawk is not strongly context free. But many researchers would accept this without a problem. And you **can** give an argument that the latter ought to accept.

The argument now goes as follows: the grammar of Mohawk generates two surface structure Mohawk string (13a) and (13b) , that the morphology will map on the same string (12) of Mohawk:

(12) the man *house-praising-liking* admires this *house-liking-praising*.

(13) a. the[nom] man[nom] *house-praising-liking* admires
this [gen] *house-liking-praising* [obj].
b. the[nom] man[nom] *house-praising-liking* admires
this [gen] *house-liking-praising* [gen] e[obj].

Since Surface Structure Mohawk is a language in the technical sense of the word, we can unproblematically apply the Postal-Langendoen argument to **this** language, rather than to the stringset of Mohawk.
But in Surface Structure Mohawk the set of sentences of the form (13b) are **not** in the regular language (in the alphabet of Surface Structure Mohawk) that we intersect Surface Structure Mohawk with. As a consequence, the intersection of our regular language with the appropriate regular language, is indeed precisely the set of Surface Structure Mohawk strings of the form (13a), a non-context free language. Consequently, Surface Structure Mohawk is not context free.

From this it follows that:

> **Mohawk is not strongly context free**: **any linguist who accepts that Mohawk is generated through something like Surface Structure Mohawk must assume that the tree set of Mohawk is not context free**.

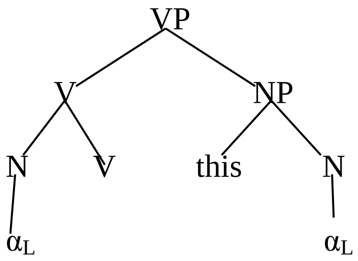
Of course, one need not accept the particular details of what I have assumed about Surface-Structure Mohawk. The point is that one can still drop some of my assumptions, as long as the yields of the syntactic trees make enough distinctions that the strings of Mohawk don't the argument will go through.
The general point is that you need to make only very few and not very controversial assumptions for the argument about strong generative capacity to go through.
And I think that this is a good thing, because it is the arguments about strong generative capacity that are interesting, not about weak generative capacity (we see a strong argument for this later, when we discuss Dutch).

A FOOTNOTE:
A general problem needs to be adressed briefly.
Suppose we make the following assumptions:
1. The rules forming complex nouns are **lexical** rules and context free.
2. Nominalization takes a VP and forms a **lexical noun**, a new basic item, formally a terminal symbol.
(Technically, this gives us an infinite lexicon, such languages are called infinite cardinality languages.)
3. Incorporation tells us that a lexical noun can incorporate into a verb.
We get the following rules:

$$\mathrm{VP} \Rightarrow_{\text{NOMINALIZATION}} \begin{array}{c} \mathrm{N} \\ | \\ [\text{yield}(\mathrm{VP})]_L \end{array}$$

where $[\text{yield}(\mathrm{VP})]_L$ is a terminal symbol.

INCORPORATION:



CONSTRAINT: both occurrences of α_L are the same

This looks perfectly context free: instead of checking whether two arbitrarily long strings are the same, we only require that two lexical items are the same lexical item.

This looks context free, but it only moves the problem to the lexicon.
You need to guarantee that if the first occurrence is $[_N\ \alpha_L]$, the second N can only expand to α_L. That is easy to guarantee by subcategorization with features if the lexicon is **finite,** but it requires infinitely many categories if the lexicon is not finite.
The problem, then, is that the innocent looking constraint that the two occurrences be the same **cannot be enforced** with context free means.
So the problem stays.

THE LEXICON ON BAMBARA (Culy 1985).
(Bambara: a languages spoken in Mali.)

Bambara has a construction:

NOUN o NOUN with meaning:
whatever NOUN

(1) wulu o wulu
dog dog
whichever dog

And the following is ungrammatical:

(2) *wulu o malo
dog rice

Bambara has nominalization:

(3) wulu + nyini + la ⇒ wulu-nyini-na (l→n, phonology)
dog search dog searcher

(4) wulu + filè + la ⇒ wulu-filè-la
dog watch dog watcher

Nominalization is recursive:

(5) wulu-nyini-na + nyini + la ⇒ wulu-nyini-na-nyini-na
dog searcher search dog searcher searcher

And these nouns can occur in the NOUN o NOUN construction:

(6) wulu-nyini-na-nyini-na o wulu-nyini-na-nyini-na
dog searcher searcher dog searcher searcher
whatever dog searcher searcher

And, in that construction they have to have the same form.

Now we can construct an argument:

Let:
R = wulu-(filè-la)^n(nyini-na)^m o wulu-(filè-la)^k(nyini-na)^p $(n,m,k,p \geq 1)$

R is a regular language.

Let:
N = wulu-(filè-la)^n(nyini-na)^m o wulu-(filè-la)^n(nyini-na)^m $(n,m \geq 1)$

N is not context free.

The empirical claim is: Bambara $\cap$ R = N
Hence, Bambara is not contextfree.

While Culy notices as a special aspect of the construction that it shows that in Bambara the **lexicon** is not contextfree, this depends on the status of the rule forming NOUN o NOUN. If we delegate this rule to the lexicon, then indeed the whole process take place in the lexicon. If we regard this rule as a syntactic rule of forming complex nouns, then the situation is exactly the same as what Postal found in Mohawk: a syntactic construction has a co-occurrence constraint, which must be satisfied by strings of unbound length, and the latter you can generate with nominalization, which (presumably) maps from the syntax to the lexicon.

Query: *o* in Bambara is also the demonstrative determiner *that*. In English, we have noun phrases with appositives, as in: ***Buck, that idiot****, asked a silly question*. It needs to be checked that *o* in the position relevant for Culy's argument cannot be similarly a demonstrative, otherwise the argument for weak generative capacity may be in trouble after al.

So both Mohawk and Bambara involve phenomena on the border of the syntax and the lexicon. The next case only involves the syntax. If you find it a problem that I only discuss funny 'fringe' languages, I will now turn to the language that maybe most of us speak natively:

SYNTACTIC COPYING IN MANDARIN CHINESE. (Radzinsky 1990)
(Mandarin Chinese: a language spoken in the China's.)

Mandarine Chinese has two kinds of yes-no questions. The kind which is important for us (brought to the linguistic attention by Jim Huang) is A not A questions.

(1)	ta	zai	jia	ta	bu	zai	jia
	she	at	home	she	not	at	home

Is she at home?

The construction can involve different constituents:

(2)	ta	[zai	jia	bu	zai	jia]
	she	at	home	not	at	home

(3)	ta	zai	[jia	bu	jia]
	she	at	home	not	home

But you cannot delete just any part. Li and Thompson have observed that in general elements that form a semantic unit must be deleted together.
In transliterated Mandarin Chinese:

(4) a. You like her shirt not like her shirt.
b. *You like her not like her shirt.
c. *You like her shirt not like her .
d. *You like shirt not like her shirt.

This means that if we look at A not A questions with NPs and restrict the recursion to adjectives, such deletions are not possible:

(5) ni xihuan geng-da de pingguo bu xihuan geng-da de pinguo
you like bigger GEN apple not like bigger GEN apple

(6) ni xihuan geng-da geng-hao de pingguo
you like bigger nicer GEN apple
bu xihuan geng-da geng-hao de pinguo
not like bigger nicer GEN apple

But not:

(6) *ni xihuan geng-da geng-hao de pingguo
you like bigger nicer GEN apple
bu xihuan geng-hao geng-da de pinguo
not like nicer bigger GEN apple

The adjectives have to be in the same order, and the number has to be the same.

We can construct an argument:

Let:
R = ni xihuan geng-da α de pingguo bu xihuan geng-hao β de pinguo
where $\alpha,\beta \in$ {geng-da, geng-hao}*

R is a regular language.

Let:
L = ni xihuan geng-da α de pingguo bu xihuan geng-hao α de pinguo
where $\alpha \in$ {geng-da, geng-hao}*

L is not context free.

The empirical claim is: Mandarin Chinese $\cap$ R = L.
Hence Mandarin Chinese is not context free.

However, once again, there is a problem with the argument, similar to the problem we had in Mohawk. (6) may be ungrammatical as a yes-no question, but it is **grammatical** as a conjunction:

You like a bigger nicer apple and not an nicer bigger apple.

This means that the empirical claim about Mandarin Chinese is incorrect, and the argument collapses.

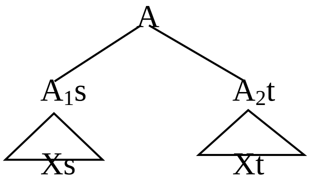
In this case it is not so clear to be that there is an easy way around the problem (i.e. an easy way to show that Mandarin Chinese is strongly non-context free).

A not A questions and conjunctions have the same surface form. While there is a constraint on the A not A questions, and Radzinsky argues that the claim is syntactic in nature (not lexical: full NPs occur in the construction; not discourse based: the construction is only sentence internal; not semantic: synonyms are not allowed), it depends **heavily** on the details of the grammar, whether we can separate the conjunction structures from the A not A structures **in the syntax**, let alone, in a way that is visible on Surface-Structure-Mandarin-Chinese.

Suppose we assume that these constructions have actually the very same syntax, but different semantic interpretations, with the constraint that the yes-no interpretation is only possible if the parts are syntactically identical. Then the sentences that are claimed to be ungrammatical are in fact generated without any problem, they just cannot have the required interpretation. But then the syntax may well be context free. Since it seems to me rather plausible that that is the correct interpretation of the facts, the argument cannot be rescued.

SHOULD COPYING WORRY US?

We have seen some cases of evidence for a non-contextfree copying rule.
What does this do to the grammar, besides making it not context free?
Not much, I think, because in the cases discussed, we are not dealing with a construction that interacts with other constructions a lot.
The syntax can be very simple:
Suppose we have two features s (for source) and t (for target) which need to match on both daughters of a node. The matching itself is perfectly context free:



The syntactic rule, which is clearly not context free is: copy the surface structure tree dominated by X_s under X_t.

But note that this is not a restructering of deletion rule: it doesn't change anything in the structure, it doesn't delete anything, it just copies. Think of it in terms of parsing. If there is no interaction with other constructions, parsing such copying structures is not very expensive. Assume that the parser has already analyzed the string up to the yield of A_1s, and it comes to the first symbol of the yield of X_t. It needs to backtrack a little to match the yield of X_s and the yield of X_t , but this is a linear amount of backtracking, and allowing a little bit of linear backtracking is not a problem for either the parsing time or the complexity of parsing.

If we speculate about the human parser, the situation is probably even simpler. It is very plausible to assume that the human parser is actually sensitive to repetitions of this sort: such repetitions are highly salient, and we know that the human parser **is** sensitive to contextual and semantic information. Hence, the human parser probably

just uses a non- linguistic substrategy of **pattern-recognition** here: we are good at pattern-recognition, and we can assume that the human parser **recognizes** the identity of pattern directly (not as part of the grammatical rules, but by using, contextually, the human capacity of pattern recognition). This means that the human parser would recognize the repetition right away, and copy the tree that it constructed for the first copy **without losing any time or space**.
Hence, such copying, though requiring an operation that goes beyond context free would not seriously add to the complexity of the grammar.
(Also, as we will see, this copying does not go far beyond context free. We have already seen that it can be done in index languages, but in fact, we will see that it can be done in languages that go only mildly beyond context free.)

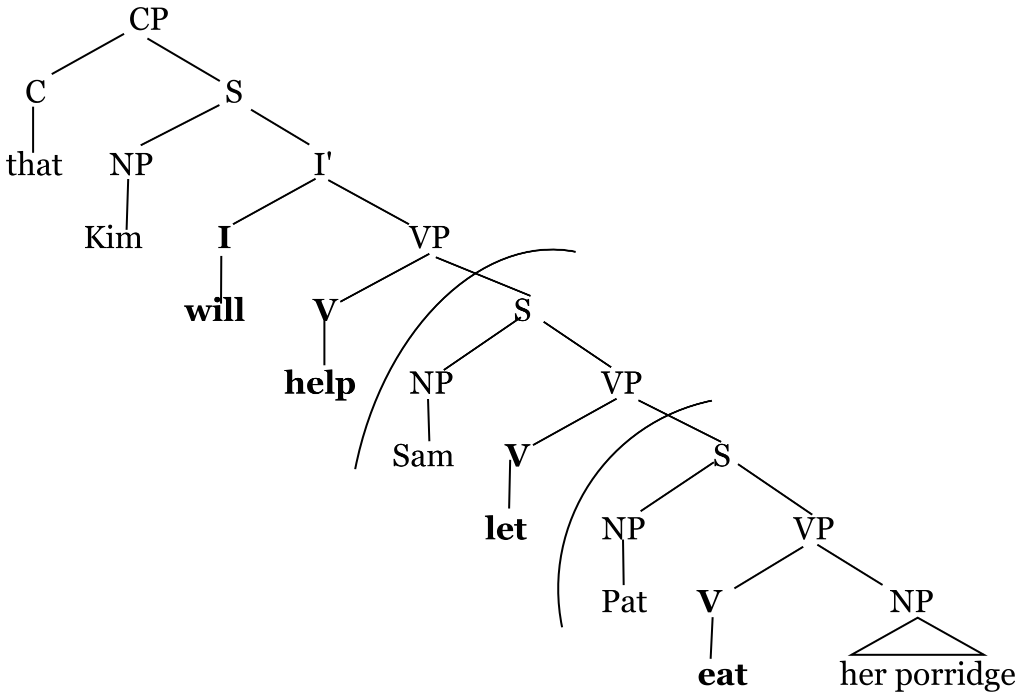
This means that, in a way, the arguments discussed so far, do not pose a very deep grammatical challange. This is not true for the last grammatical phenomenon that I will discuss here: **cross serial dependencies**. This phenomenon differs from the previous ones in that it **does** seem to interact with everything grammatical under the sun. This is clear from the literature as well: as against the previous phenomena, there is a huge literature on the grammar of cross serial dependencies.

CROSS SERIAL DEPENDENCIES

We are now concerned with verbs that take bare infinitives (infinitives without *to*) as complements, verbs like *let, make, help, see, hear***,** as in (1). (I use *that*-complements instead of sentences systematically here, because we will be concerned with verb second languages, and at the moment I am not interested in the verb second situation.)

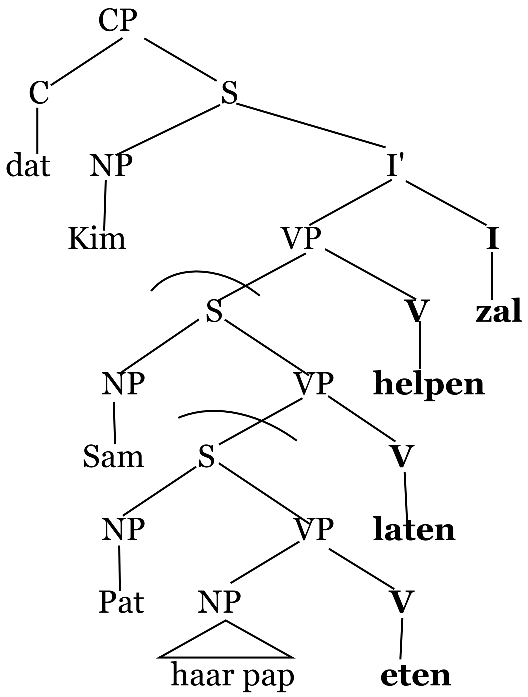
(1) That Kim will *help* Sam *let* Pat eat her porridge.

A standard assumption (for English, and not in every framework) is that cases like (1) get a **small clause analysis**. I will follow that assumption here, but only because the discussion doesn't really depend on it (and I got to choose something).:



In Dutch, German, and Swiss-German, V and I, are assumed to be on the right. The same class of verbs (well, their cognates) take bare infinitives (except that there isn't something corresponding to *make*).

This means that, following the English analysis, we would expect to find for Dutch the following:



(2) *Dat Kim Sam Pat haar pap eten laten helpen zal.

But this is **not** what we find in Dutch.
What we find is that the infinitives behave like a **cluster**, in fact, a constituent *helpen laten eten:*

FACT 1: The order of the infinitives is inversed from the above structure.

(3) Dat Kim Sam Pat haar pap zal helpen laten eten.

FACT 2: While the order of the infinitives is strictly fixed, the tensed auxiliary in Dutch can occur on either side of the cluster, and - in embedded clauses - only there, and there can't be something else in between, suggesting that we are indeed dealing with a constituent:

(4) Dat Kim Sam Pat haar pap **zal** [*helpen laten eten*]
(5) Dat Kim Sam Pat haar pap [*helpen laten eten*] **zal.**

But not, for instance, (6), nor adverbials, cf. (7), with adverbial *morgen* (tomorrow).

(6) *Dat Kim Sam Pat **zal** haar pap [*helpen laten eten*]
(7) a. Dat Kim Sam Pat haar pap morgen **zal** [*helpen laten eten*]
b. *Dat Kim Sam Pat haar pap **zal** morgen [*helpen laten eten*]

EXCURSUS ON VERB SECOND
I will assume a version of the standard account of the verb second phenomenon for the Germanic languages (though I don't think that much hinges on which account you assume). The standard account (going back to Hans den Besten in the seventies) takes its starting point in the observation that in Dutch and German, lexically filled complementizers and the tensed verb are in complementary distribution: if there is a lexically filled complementizer, the tensed verb is at the end of the sentence, if there isn't a lexically filled complementizer, the main verb is where the complementizer would be, if there were one.

We will only consider here what happens in declarative CPs.
Dutch does not allow adjunction to CP or to C'. This means that inside CP there are exactly two positions available: SPEC-of-CP and C:

[$_{CP}$ SPEC-of-CP [$_{C'}$ C IP]]

The verb second effect consists of the following two facts:

> **FACT 3**: In declaratives, the tensed verb(stem) is in C iff C is not filled by a lexically realized complementizer.
>
> **FACT 4**: In declaratives, where the tensed verb(stem) is in C, SPEC-of-CP must be lexically filled.

Since something must occur in the one and only position higher than the tensed verb, the tensed verb is in second position. It is also in second position because of the following constraint:

> **FACT 5**: What can occur in first position in the verb second construction in declaratives is **anything**, as long as:
> **1. It is a constituent.**
> **2. Putting it there does not violate independent syntactic constraints.**

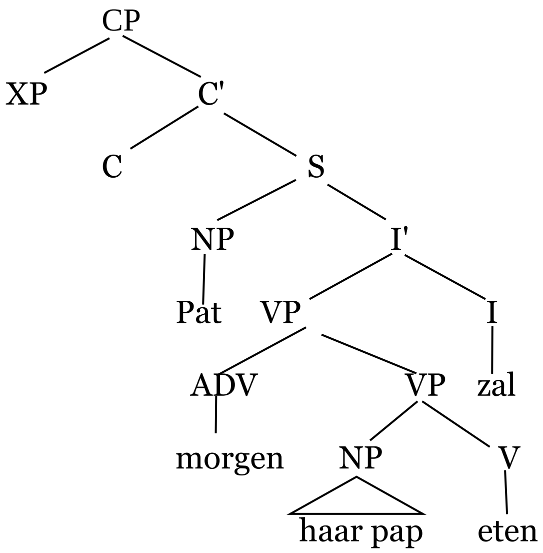
What this means is the following.
If there is a complementizer, the finite verb(stem) is inside the IP (as it is in English, except I is sitting on the other side):

Dat Pat morgen pap **zal** eten.
That Pat tomorrow porridge **will** eat

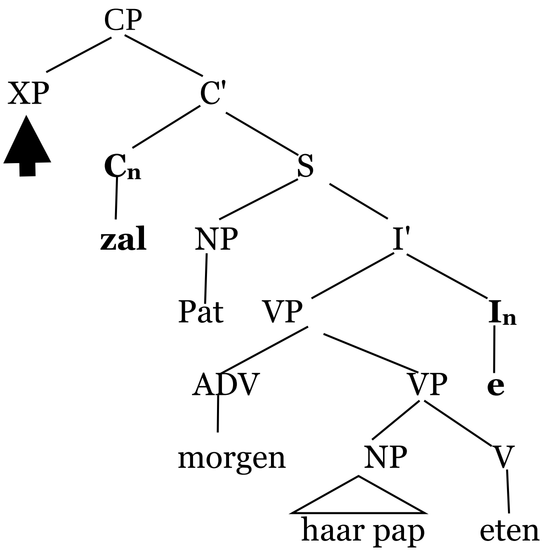
When the complementizer is empty, we do not have a structure (8a) but (8b):

(8) a. ----- Ø Pat morgen pap **zal** eten.
b. ----- **zal$_n$** Pat morgen pap e$_n$ eten.

I.e., we work from the structures:



and for verb second:



The constraints of the verb second construction say that the position XP must be filled by a constituent. Which constituent?
Answer: any of the constituents in this structure (except for the tensed verb in second position) as long as putting it in position XP doesn't violate syntactic constraints.

So, all of the cases in (9) are felicitous:

(9) a. **Pat** *zal* morgen pap eten.
b. **Morgen** *zal* Pat pap eten.
c. **Pap** *zal* Pat morgen eten.
d. **Eten** *zal* Pat morgen pap.

But the cases in (10) are **also** felicitous:

(10) a. **Pap eten** *zal* Pat morgen.
b. **Morgen pap eten** *zal* Pat.

What is not felicitious is to move something that is **not** a constituent:

(11) a. *__Morgen pap__ *zal* Pat eten.
b. *__Morgen eten__ *zal* Pat pap.

What is not felicitous is to move something that violates syntax. For instance, (12a) is felicitous, but (12b) and (12c) are not:

(12) a. **Haar pap** *zal* Pat morgen eten.
b. *__Haar__ *zal* Pat morgen pap eten.
c. *__Pap__ *zal* Pat morgen haar eten.

And, interesting enough, (13) is **not** felicitous either:

(13) *__Pat morgen pap eten__ *zal*.

(13) is not felicitous, even though **Pat morgen pap eten** is a constituent.
But this constituent in C contains the trace of the moved tensed verb, which is moved over the tensed verb, and this violates syntax.

The facts about what can occur in first position is of crucial importance for our purposes because it gives us crucial information about the verb cluster.

Take the verb cluster example we gave before:

dat Kim Sam Pat haar pap zal helpen laten eten.

Now make the complementizer empty, and put the tensed verb in the C position:

----- zal_n Kim Sam Pat haar pap helpen laten eten e_n

FACT 6: The verb cluster is a constituent, a complex verb (category V).

This is shown by the fact that the verb cluster itself can occur in first position, but its parts cannot:

(14) a. **Helpen laten eten** *zal* Kim Sam Pat haar pap.
b. *__Helpen__ *zal* Kim Sam Pat haar pap laten eten.
c. *__Laten__ *zal* Kim Sam Pat haar pap helpen eten.
d. *__Eten__ *zal* Kim Sam Pat haar pap helpen laten.
e. *__Helpen laten__ *zal* Kim Sam Pat haar pap eten.
f. *__Helpen eten__ *zal* Kim Sam Pat haar pap laten.
g. *__Laten eten__ *zal* Kim Sam Pat haar pap helpen.

This is explained if the verb cluster is a complex verb, and this means that the cluster is formed by **incorporation**.

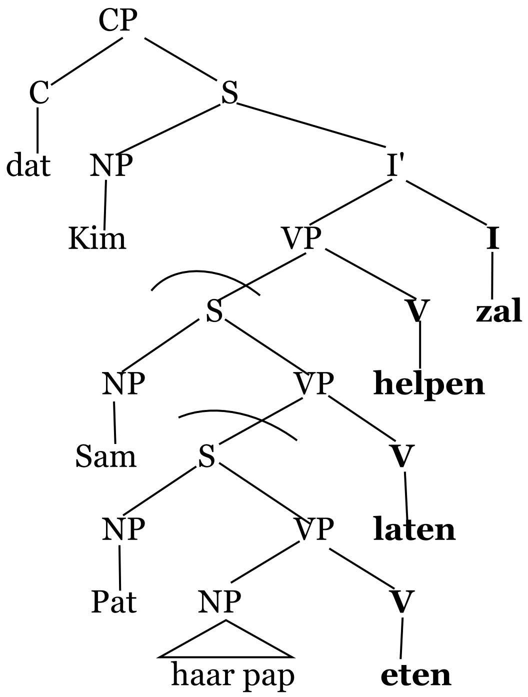
FACT 7: The verb cluster is sitting in the position of the **lowest V node** in the tree (and not, for instance, extraposed higher up on the right side of the tree, as earlier analyses have it).

This is shown by the fact that **haar pap helpen laten eten** forms a constituent for the verb second construction: crucially, the cases in (15a-c) are felicitous:

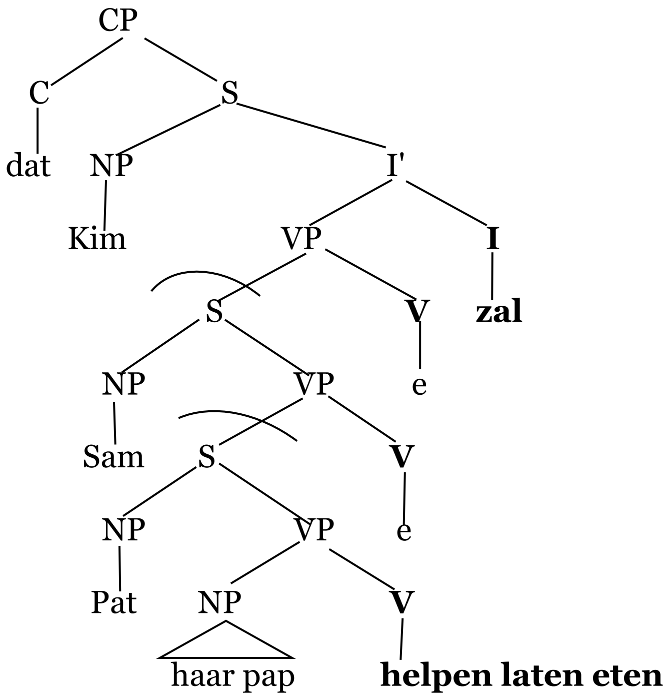
(15) a. **Haar pap helpen laten eten** *zal* Kim Sam Pat.
b. **Pat haar pap helpen laten eten** *zal* Kim Sam.
c. **Sam Pat haar pap helpen laten eten** *zal* Kim
d. ***Kim Sam Pat haar pap helpen laten eten** *zal*.

(15d) is out for the same reason as (13).

If we go back to the structure we gave for Dutch before:



Then we see that the facts given above suggest that incorporation **raids** the non-tensed right side of the tree:



From this structure we derive (16) and verb second structure (17):

(16) Dat Kim Sam Pat haar pap helpen laten eten *zal*.

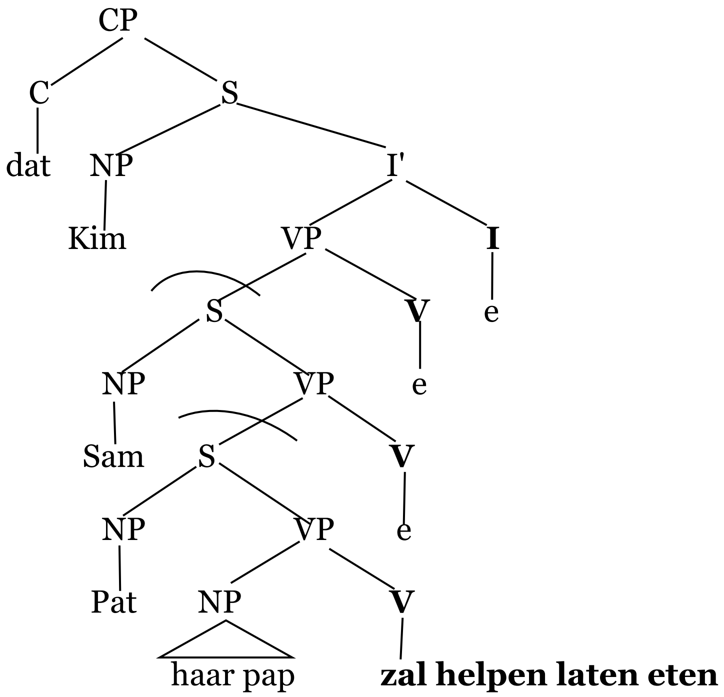
(17) ----- *zal*$_n$ Kim Sam Pat haar pap helpen laten eten e$_n$.

Here (17) forms the basis for all the verb second cases.

The only case we do not derive is (18):

(18) Dat Kim Sam Pat haar pap *zal* helpen laten eten.

I will assume that we derive (18) by allowing incorporation to optionally incorporate the tensed verb stem as well:

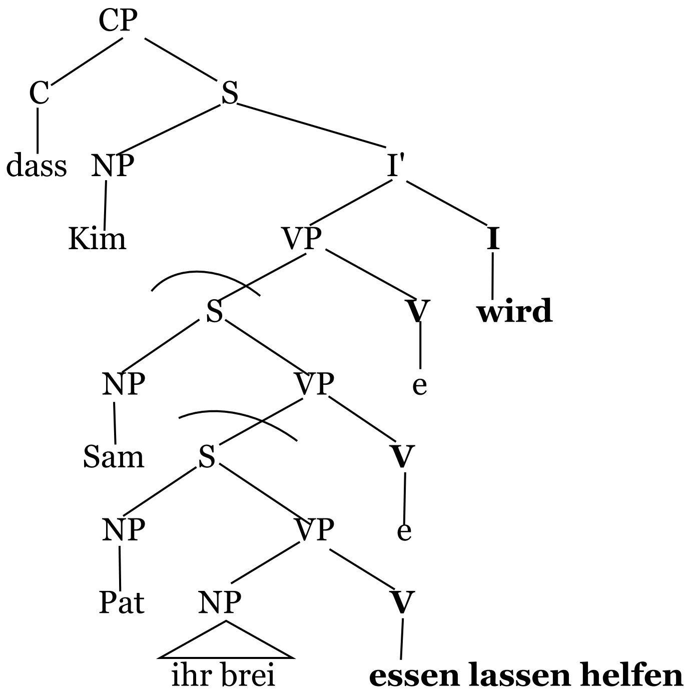


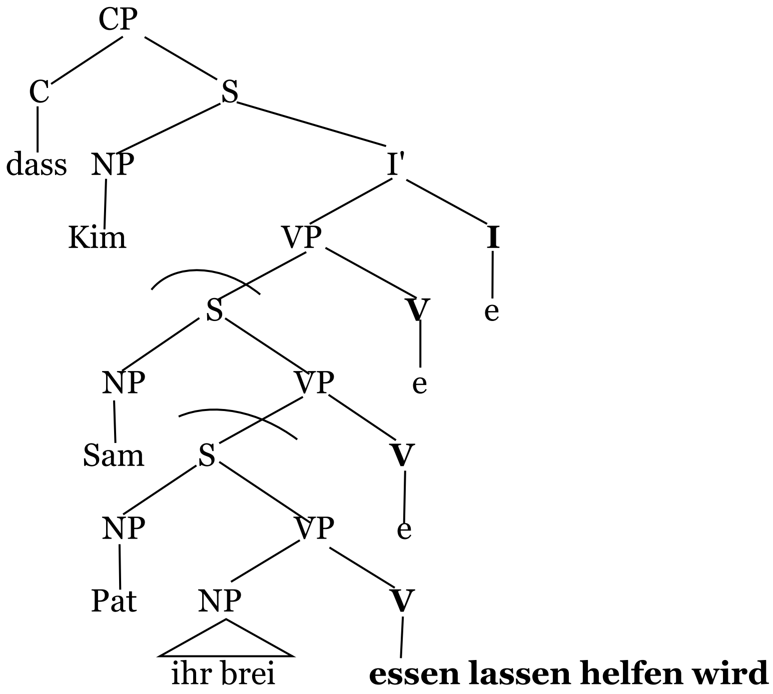
Since in Dutch it is only the tensed verb stem that moves, ***zal* helpen laten eten** cannot itself move to C. Making the obvious assumption that *zal* cannot move out of the incorparation V, it follows that incorporated *zal* cannot form the basis for verb second constructions at all. So the only thing that this assumption adds is that we get cases like (18) as well.

German shows the **inverse word order** for the cluster from Dutch. That is, German shows the word order that you would get in English by putting V and I on the right:

(19) Daβ Kim Sam Pat ihr Brei essen lassen helfen wird.
eat let help will

However, German has incorporation and verb second, like Dutch: it is **not** inverted English, because the constituent tests show the verb cluster to be a constituent in German as well. Thus, the difference between Dutch and German lies in **the direction of incorporation** (this means too that in German incorporating the tensed verb won't make a difference on the surface):


CP
C
S
dass
NP
I'
Kim
VP
I
S
V
wird
NP
VP
e
Sam
S
V
NP
VP
e
Pat
NP
V
ihr brei
essen lassen helfen


CP
C
S
dass
NP
I'
Kim
VP
I
S
V
e
NP
VP
e
Sam
S
V
NP
VP
e
Pat
NP
V
ihr brei
essen lassen helfen wird

A NOTE ON SEPARABLE PREFIXES

Many verbs have separable prefixes:

eten	*op*eten	helpen	*mee*helpen
eat	finish-eat	help	with-help

The prefix-verb compex is a normal verbal constituent. That is, apart from the separable behaviour discussed below, the prefix-versions of the verbs above behave just like the non-prefix versions.

So, they occur in the verb cluster:

(1) dat Kim Sam Pat haar pap zal *meehelpen* laten *opeten*.

They occur as a unit in first position, and cannot be broken up:

(2) a. *Opeten* zal Pat haar pap.
b. **Op* zal Pat haar pap eten.
c. **Eten* zal Pat haar pap *op*.

But the prefixes are separable in the following contexts.

A. In *te* (*to*)-infinitives the *te* separates the prefix from the verb:

(3) a. *te opeten b. op te eten
to finish-eat
c. *te meehelpen d. mee te helpen
to with-help

B. In verb-second it is only the tensed verb-**stem** that occurs in second position: the prefix stays where it is at the end:

(4) a. *Pat *opeet* haar pap -.
b. Pat *eet* haar pap *op* –.
c. *Kim *meehelpt* Sam -.
d. Kim *helpt* Sam *mee* -.

C. I already showed that in forming the verb cluster, the prefixes can stay attached to the verb. But they don't have to:

Incorporation optionally separates the prefixes and the verbs, and then the prefixes incorporate in the same order as the verbs do, before the verbs:

__[P_1 V_1]

__[P_2,V_2]

__[P_3,V_3]

__[P_4,V_4]

__[$P_1P_2P_3P_4$, $V_1V_2V_3V_4$]

That is, the the verb-cluster with prefixes behaves like a complex prefix-verb:

(5) dat Kim Sam Pat haar pap ***mee op*** *helpen laten eten* zal.
*dat Kim Sam Pat haar pap *op mee laten eten* zal

And this structure behaves like a verbal constituent: the *mee op* cannot be separated into first position, but the whole thing can oocur there:

(6) a. **mee op* zal Kim Sam Pat haar pap helpen laten eten.
b. *mee op helpen laten eten* zal Kim Sam Pat haar pap.

And the prefixes are really part of the cluster: arguments and adjuncts cannot occur to their right:

(7) a. *dat Kim Sam Pat *mee op* **haar pap** *helpen laten eten* zal.
b. *dat Kim Sam Pat haar pap *mee op* **morgen** *helpen laten eten* zal

Now we come to the tensed verb *zal*. We have seen it already at the end, as usual. We know it can also occur on the other side of the verbal cluster. Where is it with respect to the prefixes? Answer: at the beginning of the verb sequence, **after** the prefixes:

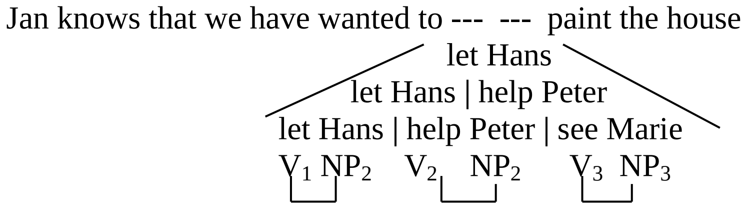
(8) a. *dat Kim Sam Pat haar pap **zal** *mee op* helpen laten eten..
b. dat Kim Sam Pat haar pap *mee op* **zal** helpen laten eten.

This suggests that indeed the tensed verbstem can incorporate, and that's how you get this order.

(It seems then that in Dutch a verb-stem can raise to I and get tense, and stay there (if C is filled), or move on to C. Or, the verb-stem can get tense by incorporating tense itself.)

THE NON-CONTEXTFREENESS OF SWISS GERMAN, DUTCH AND GERMAN.

English:


Jan knows that we have wanted to --- --- paint the house

let Hans

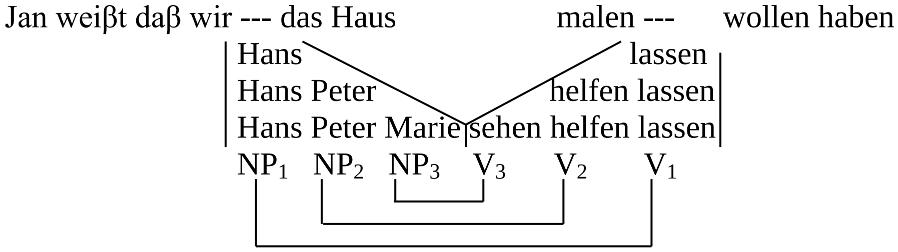
let Hans | help Peter

let Hans | help Peter | see Marie

V_1 NP_2 V_2 NP_2 V_3 NP_3

Adjacent, non-overlapping dependencies.

German:


Jan weiβt daβ wir --- das Haus malen --- wollen haben

Hans lassen

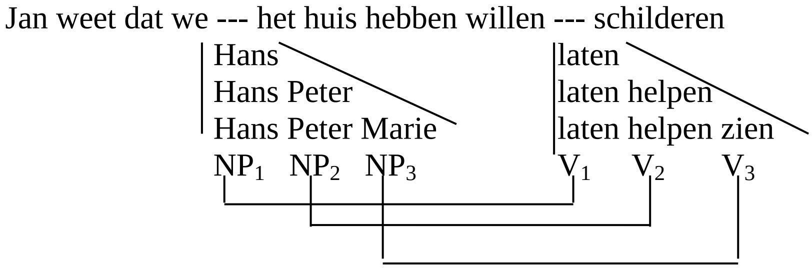
Hans Peter helfen lassen

Hans Peter Marie | sehen helfen lassen

NP_1 NP_2 NP_3 V_3 V_2 V_1

Center embedded dependencies.

Dutch:


Jan weet dat we --- het huis hebben willen --- schilderen

Hans laten

Hans Peter laten helpen

Hans Peter Marie laten helpen zien

NP_1 NP_2 NP_3 V_1 V_2 V_3

Cross serial dependencies.

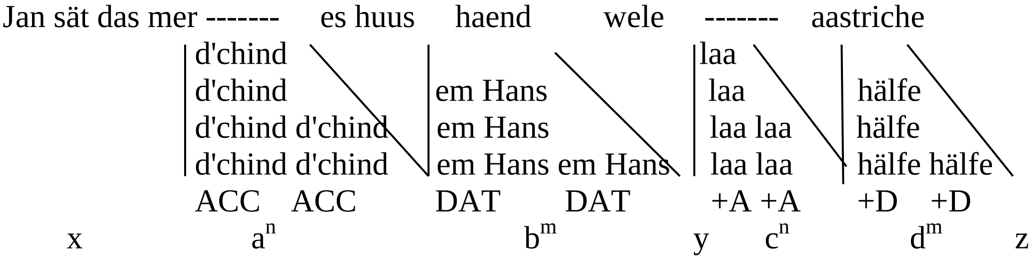
(So far, this is only of the form x a^nyb^nz, which is, of course, perfectly context free.)

Swiss German (Shieber 1985)

Shieber constructs an argument for the non-contextfreeness of Swiss German, which is based on the following facts.

-Swiss German has cross serial dependencies, like Dutch.
-*laa* [let] assigns **accussative case** to the subject of its small clause complement.
hälfe [help] assigns **dative case** to the subject of its small clause complement.
-Case is visible on the definite article:
d'chind [the child, accusative]
em Hans [Hans, dative]

This means that we find in Swiss German:



The relevant facts about Swiss German:
1. **If** all ACC-assigning verbs precede all DAT-assigning verbs, **then** all ACC-objects must precede all DAT-objects.
2. Swiss German allows argument drop, which means that there can be more verbs than NPs, but there **cannot** be more NPs than verbs, and in particular, there cannot be more ACC-objects than ACC-assigning verbs, and there cannot be more DAT-objects than DAT-assigning verbs.

Let:
R = Jan sät das mer (d'chind)^n (em Hans)^m es huus haend wele (laa)^k (hälfe)^p aastriche
$n,m,k,p \geq 0$

R is a regular language.

Let
L = Jan sät das mer (d'chind)^n (em Hans)^m es huus haend wele (laa)^k (hälfe)^p aastriche
$n,m,k,p \geq 0$ and $n \leq k$ and $m \leq p$

L is not contextfree.

Empirical claim: Swiss German $\cap$ R = L
Hence Swiss German is not context free.

The very same argument **can** actually be made for Dutch as well. Like in Swiss German, we find in Dutch:

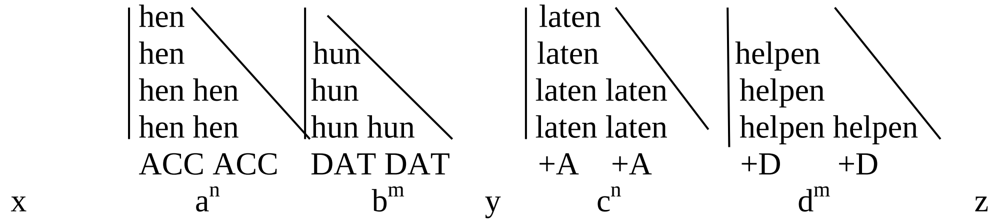
laten assigns the accusative.
helpen assigns the dative.

And the accusative/dative distinction is visible at one point in the pronominal system:

Dutch pronouns:	**Nominative**	**Dative**	**Accusative**
third person plural:	ze	hun	hen
		ze	ze

This means that we find in Dutch beauties like:

Jan weet dat we ------- het huis hebben willen -------schilderen



And we can make the same argument to prove that:
Dutch is not context free.

There is a problem with this argument for Dutch, which - I think - illuminates the weakness of arguments concerning **weak** generative capacity.

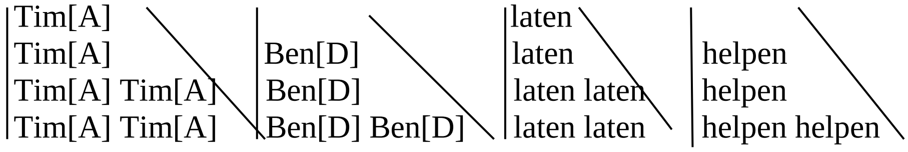
The problem is that the identification of the *hen/hun* distinction with the accusative/dative distinction is part of prescriptive Dutch, and has as such been part of grammar books since the 17th/18th century. But, it seems that **also then** the accusative/dative **roughly** fitted the usage, **but not quite**, and that is, for most speakers I know **still** the case.
This *de facto* means that we can prove that **Prescriptive Dutch** is not contextfree, and hence, for any speaker who has internalized the principles of prescriptive Dutch in this respect, we can prove that their language is not context free. But for speakers that have not internalized this part of prescritive Dutch, we cannot prove that their language is context free.

But this is stupid! Obviously my language and the prescriptive language don't differ in any essential way in complexity: if **their** grammar is non-context free then so is **mine**! We just can't show it. But that shows that arguments from generative capacity of the **string set** are a **frustratingly bad tool** for showing what complexity is hidden in **my** language, the one we can't proof to be non-context free.

For the working linguist there isn't really a problem.
The working linguist will show that there are reasons to assume the same accussative-assigning versus dative-assigning distinction for *let*/*help* in Swiss German, Dutch and German. And the working linguist will assume that where my language and prescriptive Dutch differ is in the **morphological spell out** of accussative case and dative case (roughly - I fear - for me both ACC and DAT pronouns can be spelled out as *hun* or *hen*, simply because I would use *ze* most of the times anyway, which itself is to the annoyance of my extremely prescriptive (Portuguese) brother in law.)

But, for those linguists who think it plausible that features ACC and DAT would be visible in the yields of the trees generated by the syntax, the argument can just be made **straightforwardly** for **Surface Structure Dutch**:

Jan weet dat we ------- het huis hebben willen -------schilderen



Tim[A] | | laten | |
Tim[A] | Ben[D] | laten | helpen
Tim[A] Tim[A] | Ben[D] | laten laten | helpen
Tim[A] Tim[A] | Ben[D] Ben[D] | laten laten | helpen helpen

And we show straightforwardly:

Surface structure Dutch is not context free.

And from this it follows:

The tree set of Dutch is not context free, and hence the grammar of Dutch is not context free.

But remember that the only relevant difference between the Dutch and German grammar of the verb cluster lies in the **direction of incorporation**. But that means that if the tree set of Dutch is not context free, the tree set of German isn't either. Hence it follows:

The tree set of German is not context free, and hence the grammar of German is not context free.