Universal languages

Keny Chatain

October 15, 2018

INCOMPLETE DRAFT

(Most proofs and definitions can be found in the appendix ; clickable links allow easy back-and-forth between main text and appendix)

Introduction

Background. In the linguistic community, different people will associate the claim of existence of UG to different contents. One common interpretation is the following:

> **Claim A: Universal Grammar-s.** The language faculty consists in a description of a class of languages. The task of learning consists in correctly selecting a language from that class that corresponds to the input.

Under Claim A, children may, for instance, be born with the ability to describe context-free grammars. They then start to construct a context-free grammar that corresponds to their input. Different languages correspond to different elements of that set. Another claim, which Chomsky seems to endorse, is the following:

> **Claim B: Universal Grammar-$\emptyset$** The language faculty consists in a description of one abstract language. The task of learning consists in correctly finding a mapping from this language to phonological forms and vice-versa, so that the produced forms correspond the input.

Under claim B, children may, for instance, be born with the language $\Omega = \{a^n b^n \mid n \in \mathbb{N}\}$. They soon learn to map this language to $\{(\text{the cat the dog})^n(\text{chased chased})^n \text{ arrived} \mid n\}$. (This is only provided as an example ; it stands to reason that UG will have to be more complicated than Ω to capture all possible constructions of English). Different languages correspond to different mappings of the structures of UG.

In both claim A and claim B, the complexity of observed[1] natural languages is a direct correlate of the complexity of the underlying representations. The observed center-embeddings of the English language point for instance to the underlying context-freeness of UG (claim B), or to the possibility of writing context-free languages with UG (claim A). Therefore, the study of the "*surface complexity*" of languages provides an interesting tool to probe the underlying complexity of UG under both claims.

It would however seem that this kind of study does not directly reveal which of claim A and claim B is true. Indeed, any surface complexity will either be accommodated as a complexity of the underlying description of languages (claim A) or as a complexity of the underlying language (claim B).

Results The goal of this article is to show formally (under maybe unrealistic assumptions) that this intuition is incorrect and that the predictions of the two claims regarding surface complexity can sometimes be teased apart. My main result is that if the complexity of observed human languages is that of *mildly context-sensitive* grammars, as described by 2-multiple context-free grammar, tree-adjoining grammars, Stabler's minimalist grammars, combinatory categorial grammars, then claim B is untenable.

For the problem to be amenable to formal analysis, I make some simplifying assumptions, which may all be rejected. If they are, the point of this article can be seen as one of principle: claim A and claim B are distinguishable on the basis of external data only. One such simplifying assumption, and probably the most objectionable, concerns the nature of the mapping to the phonological interface posited by claim B. I posit that such mappings are rational transducers, and therefore complexity-wise simple. This corresponds to the assumption that any complexity beyond finite-state complexity observed in the phonological output is directly imputable to UG.

Another simplifying assumption is that the language described by UG under claim B is a set of strings. This certainly goes against the standard view that UG should describe set of syntactic structures or trees. I take this approximation to be of minor impact since low-complexity encoding schemes (e.g. Polish notation) can translate trees to strings and vice-versa.

This article's main result will be prefaced with other results which do not carry theoretical value but helps us understand the formal underpinnings of claim B. I will for instance show a criterion that establish whether a given class of observed languages can be seen as the image of one single language under a number of mappings. I will also prove that the class of context-free languages can be described as the image of one single language under a number of mappings, thus providing a stronger version of the celebrated Chomsky-Schützenberger theorem. The same result will be offered for recursively enumerable languages.

[1] I will call *observed languages*, the languages understood as sets of phonological strings. This is the kind of languages that linguists interact with in order to get to the *abstract language*, i.e. the underlying mental representation. Claim A correspond to the claim that there are multiple abstract languages for different languages, claim B to the claim that there is only one.

0.1 Setting up the stage

Under claim B, the set of possible human languages may formally be described as $\{f(L) \mid f \in \mathscr{F}\}$, where $\mathscr{F}$ is a fixed set of mappings and L the underlying language described by UG. As justified in the introduction, I take $\mathscr{F}$ to be $\mathscr{R}$ the set of rational transductions.

We are interested in the reverse problem: given a posited class C of possible observed human languages, can it be seen as the image by all possible mappings of a single language L? If it is, we will say that L *generates* C or that C is generative[2]. In other words, $C = \{R(L) \mid R \in \mathscr{R}\}$, which we will abbreviate as $C = \langle L \rangle$. For convenience, we will also say that L generates L' just in case there is a rational transducer T such that $L' = T(L)$. We note $L \prec L'$.

If claim B is true, the class of all possible human languages ought to be *generative*. We can now ask for particular well-established classes of languages whether they are generative or not. If they are not, then claim A and claim B make divergent predictions: claim A predicts that so long as a formal description of the class can be provided, that class could be the class of human languages, claim B predicts it isn't.

First on our list of class is the class of *context-free languages* ; this class is known to be inadequate to capture human languages so any result about this class is of little theoretical impact.

1 CFL and strong Chomsky-Schutzenberger theorem

1.1 Proving that context-free languages are generative

Consider the enumerable class of context-free languages CFL. Half of the work of proving that CFL is generative is done by the following theorem

Theorem 1 (Chomsky-Schützenberger)**.**

$$CFL = \{L \mid \exists n, \exists T, T \textit{ is a rational transducer} \wedge L = T(D_n)\}$$

where D_n is the n-th Dyck language

In our parlance, this theorem states $CFL = \bigcup_n \langle D_n \rangle$. This does not show that CFL is generative. However, this theorem, as we will now see, is very suggestive. Suppose we dropped the standard assumption that languages need be over a finite alphabet (cf 5.1). Then we could consider the language $D_\infty = \bigcup_n D_n$, the language of well-parenthesized expressions when the number of parenthesis is infinite. Extending the notion of rational transduction to languages over infinite alphabets, it would seem that the theorem of Chomsky-Schützenberger entails $CFL = \langle D_\infty \rangle$. This of course does not make sense since the notion of rational transduction has not been defined over infinite alphabets.

[2]This corresponds exactly to the notion of *principal ideal* used elsewhere

To prove that CFL is generative, one would need to find a language over a finite alphabet that can fulfill the role of D_∞. As it turns out, there is a way to put in correspondence words over an infinite alphabet with words over binary alphabet. This is provided by the following alphabetic morphism:

Definition 1 (Infinite-to-finite conversion)**.** *ϕ is a morphism defined by $\phi(a_i) = a_0^i a_1$ for all i*

Since this morphism is injective, as can easily be checked, it does not lose any information about the original words. Through this morphism, we can now define a version of D_∞ over a binary alphabet. This new language, call it Ω, is simply $\phi(D_\infty)$. This language can now be shown to generate CFL, just as D_∞ was. Furthermore, it is context-free.

Theorem 2 (Strong Chomsky-Schützenberger)**.**

$$CFL = \langle \Omega \rangle$$

Proof. See appendix □

This shows the class of context-free languages is generative. Claim A and Claim B are not distinguishable on this class.

2 A criterion for generativity

In this section, I study the notion of generativity in more details. At the end, I provide a necessary and sufficient criterion for generativity on *enumerable* class of languages closed under finite union and rational transduction. We start with the following easy properties, which help motivate the restriction on the criterion below:

Theorem 3. *If C is generative, then:*

- *C is closed under rational transduction.*
- *C is closed under finite union.*

Proof.
- Straightforward.
- If $C = \langle L \rangle$, and $R(L)$ and $R'(L)$ are in C then $R(L) \cup R(L') = (R \cup R')(L)$ is in C.

□

The following criterion yields a intensional characterization of generativity and in that is useful. In practice, it is only really useful for showing non-generativity (see discussion of mildly-context-sensitive languages for an example of use).

Theorem 4 (Criterion for generativity)**.** *If C is an enumerable class of languages closed under finite union and rational transduction, only one of the following statements hold:*

- *C is generative*
- *$C = \bigcup_i C_i$ where (C_i) is a strictly increasing sequence of generative classes.*

Proof. See appendix for proof that one of these statements has to hold.

Let's just prove that both statements cannot be true at the same time. Indeed if $C = \langle L \rangle$ for some L and $C = \bigcup_i C_i$, then there must be some j such that $L \in C_j$. Since C_j is closed under transduction, $\langle L \rangle \subset C_j$. But since the (C_i) are strictly increasing, $C_j \subsetneq C$, a contradiction.

□

The astute reader will notice that there is an interesting clash with the Chomsky-Schützenberger theorem here. According to this theorem, $CFL = \bigcup_n \langle D_n \rangle$. The sequence $(\langle D_n \rangle)$ is clearly increasing. But since CFL is generative, this sequence must be ultimately constant. So there must be some n such that $\langle D_n \rangle = CFL$. Which n? I venture the conjecture that D_1 generates CFL and that D_0 does not. Basically, it seems that with encoding of the like of ϕ, one can show that two types of parenthesis is enough to encode all possible parenthesized expressions. In other words, D_1 generates all the D_n. Showing that one type of parentheses won't be enough and that D_0 does not generate CFL seems more difficult so I leave the question open here.

We conclude this section with a more spectacular result, which suggests generative classes may be quite big and complex. This suggests that claim B is less restricted of an hypothesis than one may think.

Theorem 5 (Wild generation theorem)**.** *Let C be an enumerable class. Then C is included in a generative class.*

Proof. See appendix. □

3 Mildly-context-sensitive languages

In this section, we prove that the class of mildly context-sensitive languages is not generative. This class allows serial dependencies that context-free grammars don't, but only in limited amount, so that its complexity remains close to that of context-free languages, to which languages for the most part comply. If this class of languages genuinely constitute the class of possible human languages, then our result show that claim B cannot be correct, i.e. that human language cannot be reduced to mapping of a single language to the interfaces.

Many mildly context-sensitive formalisms have been proposed. While each of them comes with its own structures, they have been shown to be all weakly equivalent to a

restriction of multiple context-free languages, called 2-multiple context-free languages (hereafter MCFL). So for the purpose of proving, I will define and adopt 2-MCFL.

Definition 2. *A 2-**multiple context-free grammar** (hereafter 2-MCFG) is constituted of:*

1. *a set of symbols $\mathscr{S}$, which can be unary or binary, called the non-terminals*
2. *A distinguished unary non-terminal S, the end symbol*
3. *an alphabet Γ, subset of Σ*
4. *a set of rules P of the form:*

$$A_0(s_1, s_2) \leftarrow A_1(x_1, y_1), A_2(x_2, y_2) \dots A_n(x_n)$$

or

$$A_0(s_1) \leftarrow A_1(x_1, y_1), A_2(x_2, y_2) \dots A_n(x_n)$$

where :

- *A_0, A_1,... A_n are unary or binary symbols (choose the form above which corresponds to the arity of A_0)*
- *the s_i are strings made of the variables x_i and letters of the alphabet Γ ; the x_i's have at most one occurrence in the s_i's*

The rules in 2-MCFGs should be understood as rules of deduction. The language defined by a 2-MCFG is the set of strings w such that $S(w)$ can be deduced. Here follows an example of a 2-MCFG and examples of deduction:

(1) a. Language to be described: $L = \{www w \mid w \in \Sigma_1\}$, the language of powers of four.

b. 2-MCFG that defines L:

- Non-terminals: A, B, S
- Rules:

$$\begin{array}{rcl} A(\epsilon, \epsilon) & \leftarrow & \\ A(ax, ay) & \leftarrow & A(x, y) \\ A(bx, by) & \leftarrow & A(x, y) \\ B(x, y) & \leftarrow & A(x, y) \\ S(xyx'y') & \leftarrow & A(x, y), B(x', y') \end{array}$$

- Derivation of the word $abababab$: TO BE COMPLETED

To conclude the proof, we need to invoke another refinement of the notion of 2-MCFLs. In the current formalism, the rules of the grammars may have an arbitrary large number of premises. From that, one can define a more restricted set of grammars:

Definition 3. *A 2-MCFG is a (2,i)-MCFG, if the right-handside of all rules has at most i symbols. The class of languages defined by (2,i)-MCFGs is called (2,i)-MCFLs.*

This restriction creates a hierarchy of class, each encompassing the previous one.

Lemma 1. $2MCS = \bigcup_i (2,i)\text{-}MCFL$

Proof. Straightforwardly, since the number of rules in a MCFG is finite □

Furthermore, it is known that this sequence of class is strictly increasing:

Lemma 2. $(2,1)\text{-}MCFL \subsetneq (2,2)\text{-}MCFL \subsetneq \ldots$

Proof. See [Rambow and Satta, 1999]. □

Lemma 3. *For all i, $(2,i)MCFL$ are closed under rational transductions.*

Proof. See appendix (TODO: write proof) □

With these lemmas, we fall squarely within the case of application of our criterion. We can now seamlessly deduce our main result

Theorem 6. *2-MCFL is not generative.*

Proof. The sequence of (2,i)-MCFL is a strictly increasing sequence closed under rational transduction, whose union is 2-MCFL. By the criterion, 2-MCFL is not generative. □

4 Recursively enumerable languages

5 Conclusion and discussion

References

[Rambow and Satta, 1999] Rambow, O. and Satta, G. (1999). Independent parallelism in finite copying parallel rewriting systems. *Theoretical Computer Science*, 223(1-2):87–120.

Appendix

5.1 Enumerable class of languages

Throughout, we'll make use of the following enumerable alphabet $\Sigma = \{a_0, a_1, a_2, \ldots\}$. We say the support of a language L to be the set $\{a \in \Sigma \mid \exists w \in L, a \prec w\}$, the set of letters used by the language. Unless otherwise specified, all our languages have finite support.

Definition 4. *A **class of languages** is any set of languages with finite support.*

We will note Σ_n the restricted alphabet $\{a_0, \ldots, a_n\}$. When talking about Σ_1, it will be convenient to rename the letters a_0, a_1 with a and b respectively.

Important remark. Whenever I will be talking about a rational or a context-free language, the reader should understand a rational or a context-free language over a *finite alphabet* $\Gamma \subset \Sigma$. In our parlance, all languages considered will have *finite support.*

5.2 Rational Transducers

In this note, we will often make use of rational transduction.

While rational transduction is often described in terms of a machine that effects the transduction, it will be useful for us to define it in terms of its closure properties.

Definition 5. *The set of rational tranducers $\mathscr{R}$ over Σ is the smallest set such that:*

- *for any rational languages L_1, L_2, $L_1 \times L_2$ is in $\mathscr{R}$*
- *for any rational transducer $R_1, R_2 \in \mathscr{R}$, $R_1 \cup R_2 \in \mathscr{R}$, $R_1 \circ R_2 \in \mathscr{R}$ and $R_1^* \in \mathscr{R}$.*

Since rational languages have *finite support*, it will follow that rational tranducers too have *finite support.* In other words, any rational tranducer will be a subset of $\Sigma_n * \times \Sigma_n *$ for all n. Useful for what comes next is the following remark:

Theorem 7. *$\mathscr{R}$ is enumerable*

Proof. This follows from the machine characterization of rational transduction. Any rational transducer may be defined by a machine $(Q, \Gamma, \Gamma', \delta, \omega, F)$, where:

- Q is a finite subset of **N**: **the set of states**
- Γ and Γ' are finite subsets of Σ: **the input and output alphabets**
- $\delta : Q \times \Gamma + \epsilon \to \mathscr{P}(Q)$: **the transition function**

- $\omega : Q \times \Gamma + \epsilon \times Q \rightarrow \Gamma' *$: **the output function**
- $F \subset Q$

So the set of machines is itself a enumerable union of finite sets, therefore enumerable. Since there exists a surjective mapping from machines to rational transducers, it follows that $\mathscr{R}$ is enumerable. □

5.3 Dyck languages

The n-th Dyck language D_n is the language of well-parenthesized expressions with n-parenthesis. Given the alphabet Σ, we'll consider that the a_{2i} is the set of opening parenthesis and a_{2i+1} is the corresponding set of closing parenthesis. For convenience we can rename these symbols accodingly:

- $[_i \stackrel{def}{=} a_{2i}$
- $]_i \stackrel{def}{=} a_{2i+1}$

An example is worth a thousand words:

(2) a. $[_0]_0[_0[_0]_0]_0 \in D_0$
b. $[_0[_1]_1]_0[_1[_0[_0]_0]_0]_1 \in D_1$
c. $[_1[_2[_1]_1]_2]_1[_0[_0]_0]_0 \in D_1$

Formally:

Definition 6. *D_n is the equivalence class under the equivalence relation $\equiv_n$ generated by the following:*

- $[_0]_0 \equiv_n \epsilon$
- $[_1]_1 \equiv_n \epsilon$

  ...

- $[_n]_n \equiv_n \epsilon$

5.4 Transducing classes

Definition 7. *A class C is closed under rational transduction if for all L in C, for rational transducers T, $T(L) \in C$*

5.5 Proof of strong Chomsky-Schützenberger

We first need to show that Ω generates all D_n for all n. From weak Chomsky-Schützenberger, it will then follow that $CFL \subset \langle\Omega\rangle$.

For that, we construct a rational transducer from Ω to D_n. Consider the following relations between words of Σ_1 to words of Σ_{2n+1}: $R_i = \left\{a_0^i a_1\right\} \times \{a_i\}$ for $i \leq 2n+1$ and $R_{2n+2} = a_0^{2n+2} a_0^* a_1 \times \epsilon$. All these relations are rational transducers because they are the Cartesian product of two rational sets. So the relation $R = (R_0 \cup \ldots \cup R_{2n+2})^*$ is a rational transducer.

The relation R is defined so that for every i from 0 to $2n+1$, $R(\phi(a_i)) = \{a_i\}$ and for every $i > 2n+1$, $R(\phi(a_i)) = \epsilon$. From this, it is easy to show that if $w \in \Omega$ and can be written as $\phi(w')$ where $w' \in D_\infty$, $R(w) = \left\{w''\right\}$ where w'' is the word obtained from w' by removing all letters a_i for $i > 2n+1$. Consequently, $R(\Omega) = D_n$.

So $CFL \subset \langle\Omega\rangle$. To show the converse, we just need to prove that Ω is a context-free language. Since context-free languages are closed under rational transductions, it will follow that any language generated by Ω is context-free, thus $CFL \supset \langle\Omega\rangle$. To do that, I provide a grammar that generates Ω:

$$S \to SS \tag{1}$$
$$S \to \epsilon \tag{2}$$
$$S \to P a_1 \tag{3}$$
$$P \to a_0 a_0 P a_0 a_0 \tag{4}$$
$$P \to a_1 S \tag{5}$$

We just need to show that this a grammar of Ω indeed. The following lemma do the trick.

Lemma 4.

$$S \to^* \phi([_n)\ S\ \phi(]_n)$$

Proof. Use rule 3 and n times rule 4 and one time rule 5 to get: $S \to^* a_0^{2n} a_1\ S\ a_0^{2n+1} a_1$ □

Lemma 5. *For every word w in D_n, $S \to^* \phi(w)$*

Proof. D_n can be generated by the following grammar:

$$S \underset{D_n}{\to} SS \tag{1'}$$
$$S \underset{D_n}{\to} \epsilon \tag{2'}$$
$$S \underset{D_n}{\to} [_n S]_n \tag{3'}$$

If w is in D_n, then $S \underset{D_n}{\rightarrow} \ldots \underset{D_n}{\rightarrow} w$. From this derivation of w, one can construct a derivation of $\phi(w)$ by replacing every use of rule $1'$, $2'$ by the corresponding 1 and 2 and every use of $3'$ by use of the derivation in lemma 4. □

From this it follows that the language generated by the grammar contains Ω. The converse is obtained by the following lemma.

Lemma 6. *If $S \rightarrow^* w$, then $w \in \Omega$*

Proof. By induction over the length of the derivation $S \rightarrow^* w$. The only derivation one-step long derivation is $S \rightarrow \epsilon$. $\epsilon = \phi(\epsilon)$ is clearly in Ω.

For longer derivations, if the first step is $S \rightarrow SS(\rightarrow^* w)$, then one can find two words w_1 and w_2 such that $w = w_1 w_2$ and $S \rightarrow^* w_1$ and $S \rightarrow^* w_2$ in less steps. By the induction property, $w_1, w_2 \in \Omega$. So $w = w_1 w_2 = \phi(w_1')\phi(w_2') = \phi(w_1' w_2')$ for some w_1' and $w_2' \in D_\infty$. Since all the D_n are closed under concatenation, $w_1' w_2' \in D_\infty$ and consequently, $w \in \Omega$

If the first step of the derivation is $S \rightarrow Pa_1$, then the following steps must be a sequence of application of 4 followed by one application of 5. This is the sequence of lemma 4. So for some k, the derivation is as follows $S \rightarrow^* \phi([_n)\ S\ \phi(]_n) \rightarrow^* w$. This means that $w = \phi([_n)\ w_0\ \phi(]_n)$ and that $S \rightarrow^* w_0$ in less steps. So by the induction property, $w_0 = \phi(w_0')$ for some $w_0' \in D^\infty$. Thus, $w = \phi([_n w_0']_n)$. So $w \in \Omega$. □

5.6 Proof of the criterion

The following tool will be needed in the proof

Lemma 7. *Let C be an enumerable class of languages closed under finite union and rational transduction. If L_1 and L_2 are in C, then there exists a language L in C such that $\langle L_1 \rangle \cup \langle L_2 \rangle \subset \langle L \rangle$*

Proof. First, let's find in C a copy of L_2 that has a support disjoint from that of L_1. Since L_1 has finite support, we can find an n such that the support of L_1 is included in Σ_n. One can then define a alphabetic morphism ϕ defined on the support of L_2 such that for all l, $\phi(a_l) = a_{l+n+1}$. Let $L_2' = \phi(L_2)$; since no letter is mapped to an element of Σ_n by ϕ, the supports of L_1 and L_2' are disjoint.

Next, consider $L = L_1 \cup L_2'$. Let S_1 and S_2 be the supports of L_1 and L_2 respectively. By construction, $S_1 \cap S_2 = \emptyset$. Let w_1 and w_2 be two words of L_1 and L_2. We can now define two rational transducers: $R_1 = Id_{S_1 *} \cup \left(S_2^* \times \{w_1\}\right)$ and $R_2 = Id_{S_2 *} \cup \left(S_1^* \times \{w_2\}\right)$. I claim that $L_1 = R_1(L)$ and $L_2 = R_2(L)$. Here is the computation that shows it:

$$\begin{aligned} R_1(L) &= \left[Id_{S_1*} \cup \left(S_2^* \times \{w_1\}\right)\right](L_1 \cup L_2) \\ &= Id_{S_1*}(L_1) \cup Id_{S_1*}(L_2) \cup \left[S_2^* \times \{w_1\}\right](L_1) \cup \left[S_2^* \times \{w_1\}\right](L_2) \\ &= L_1 \cup \emptyset \cup \emptyset \cup \{w_1\} \\ &= L_1 \end{aligned}$$

And *mutatis mutandis* for $R_2(L)$. So L generates both L_1 and L_2 and consequently all the languages they generate ; so $\langle L_1 \rangle \cup \langle L_2 \rangle \subset \langle L \rangle$. Furthermore, since L was obtained from L_1 and L_2 by union and rational transduction only, $L \in C$ □

Most of the criterion's content is included in the following lemma:

Lemma 8. *Let C be an enumerable class of languages closed under finite union and rational transduction. Then $C = \bigcup_i \langle L_i \rangle$ where the sequence of classes $(\langle L_i \rangle)$ is increasing.*

Proof. Let (L'_i) be an enumeration of the languages in C. Then $C = \bigcup_i \langle L'_i \rangle$. We can construct the L_i from the L'_i by induction. Our induction will guarantee that the following holds:

- $(\langle L_i \rangle)$ is increasing
- $\bigcup_{i \leq n} \langle L'_i \rangle \subset \langle L_n \rangle$

One can take L'_0 to be L_0. Now assume all L_i up till n have been constructed. L_n and L'_{n+1} are in C. By lemma 7, there exists a L_{n+1} such that $\langle L_n \rangle \cup \langle L'_{n+1} \rangle \subset \langle L_{n+1} \rangle$. From this, it follows that a) $\langle L_n \rangle \subset \langle L_{n+1} \rangle$ and b) $\bigcup_{i \leq n+1} \langle L'_i \rangle \subset \langle L_n \rangle \cup \langle L'_{n+1} \rangle \subset \langle L_{n+1} \rangle$, concluding the induction.

The sequence of $(\langle L_i \rangle)$ satisfies the requirements of the lemma: it is increasing and since $\bigcup_{i \leq n} \langle L'_i \rangle \subset \langle L_n \rangle$ for all n, $C = \bigcup_i \langle L'_i \rangle \subset \bigcup_i \langle L_i \rangle \subset C$ so $L = \bigcup_i \langle L_i \rangle$. □

To conclude the proof of the criterion, we notice that two cases may happen:

- **The sequence $(\langle L_i \rangle)$ is ultimately constant.** Let's call k the rank after which it is constant. Then $\langle L_k \rangle = \langle L_{k+1} \rangle = \ldots = C$. C is therefore generative.
- **The sequence $(\langle L_i \rangle)$ is not ultimately constant.** This means that for all k, one may find a $k' > k$ such that $\langle L_k \rangle \subsetneq \langle L_{k'} \rangle$. From that fact, it follows that one can construct a subsequence $\left(\langle L_{\psi(i)} \rangle\right)$ of $(\langle L_i \rangle)$ that is strictly increasing. Since $\bigcup_{i \leq \psi(n)} L_i \subset L_{\psi(n)}$, $\bigcup_i L_{\psi(i)} = C$

This corresponds to the two cases from the theorem.

5.7 Wild generation theorem

I claim that without loss of generality, we may assume that C is closed under rational transduction. Indeed, consider C_0 the closure of C under finite union. Consider C_1 the closure of C_0 under rational transductions. I leave it to the reader to show that C_1 is still closed under union[3]. If C_1 is contained in a generative class then it will follow that C, which is smaller, will be contained in a generative class. We just need to show that C_1 is enumerable. This follows from the following two lemmas:

Lemma 9. *If D is enumerable, its closure under finite union is too*

Proof. Consider $D_n = \{L_1 \cup \ldots \cup L_n \mid L_1, \ldots, L_n \in D\}$. D_n is the image of D^n under the following mapping:

$$\begin{cases} D^n & \to & D \\ (L_1, \ldots, L_n) & \mapsto & L_1 \cup \ldots \cup L_n \end{cases}$$

Since D^n is enumerable, so is D_n its image. The closure of D under finite union is the enumerable union of D_n. It is therefore enumerable. □

Lemma 10. *If D is enumerable, its closure under rational transduction is too.*

Proof. The closure of D under rational transduction D' is the image of $D \times \mathscr{R}$ under the following mapping:

$$\begin{cases} D \times \mathscr{R} & \to & D' \\ (L, R) & \mapsto & R(L) \end{cases}$$

Since $\mathscr{R}$ is enumerable by theorem 7, so is D' □

So we can assume C to be enumerable, closed under rational transduction and finite union. By the criterion and its proof, we know that two cases may occur: a) C is generative, in which case the proof is complete, b) $C = \bigcup_i \langle L_i \rangle$ where $(\langle L_i \rangle)$ is strictly increasing. In fact, the proof of the criterion yields a stronger statement ; we can assume the L_n to be the union of languages M_i for $i \subset n$ where the the languages M_i have disjoint supports.

Next, consider the language $M = \bigcup_i M_i$. M has infinite support so it can't be the input to a rational transducer, but we can use the same infinite-to-finite encoding we used in the proof of the strong Chomsky-Schützenberger. Consider $K = \phi(M)$ where ϕ is the alphabetic morphism defined in the section **??**. To conlude the proof we just need to show that K generates all of the L_n's.

Consider a particular n. Let Γ be the finite support of L_n. The set of encodings by ϕ of letters that are *not* in Γ (i.e. $E = \left\{a_0^n a_1 \;\middle|\; a_n \notin \Gamma\right\}$) is a rational set, since Γ is finite. So $T_0 = E \times \{w\}$, where w is any word of L_n is a rational transducer. And so is $T = T_0 \cup \bigcup_{a_i \in \Gamma} \left\{a_0^i a_1\right\} \times \{a_i\}$. It is easy to show that for all $i > n$, $T(\phi(M_i)) = \{w\}$ and

[3] The reader can be inspired by lemma 7.

that for all $i \leq n$, $T(\phi(M_i)) = M_i$. Consequently, $T(K) = T(\phi(\bigcup_{i\leq n} M_i \cup \bigcup_{i>n} M_i)) = \bigcup_{i\leq n} T(\phi(M_i)) \cup \bigcup_{i>n} T(\phi(M_i)) = \bigcup_{i\leq n} M_i \cup \{w\} = L_n$

To conclude, K generates all the L_n, so it generates all languages in C. In short $C \subset \langle K \rangle$.