形式文法と言語学：簡単なまとめ（資料）

福井直樹
上智大学
(2017 年 9 月 6 日、於 慶應義塾大学)

Grammar: A *generative grammar* G is an ordered quadruple G=(V_N, V_T, S, F) where V_N and V_T are finite alphabets with $V_N \cap V_T = \varnothing$, S is a distinguished symbol of V_N, and F is a finite set of ordered pairs (P, Q) such that P and Q are in $(V_N \cup V_T)^*$ and P contains at least one symbol from V_N. (Cf. Revesz 1983)

NB: The symbols of V_N are called *nonterminal* symbols and are usually denoted by capital letters. The symbols of V_T are called *terminal* symbols and are denoted by small letters. The sets V_N and V_T are disjoint (i.e., $V_N \cap V_T = \varnothing$). The nonterminal symbol S is called the *initial* symbol and is used to start the derivations of the strings/words of the language. The ordered pairs in F are called *rewriting rules* or *productions* and are written in the form $P \rightarrow Q$ where the symbol $\rightarrow$ is not in $V_N \cup V_T$.

Derivation in one step: Given a grammar G=(V_N, V_T, S, F) and two words X, $Y \in (V_N \cup V_T)^*$, we say that *Y is derivable from X in one step*, in symbols $X \underset{G}{\Rightarrow} Y$, iff there are words P_1 and P_2 in $(V_N \cup V_T)^*$ and a production $P \rightarrow Q$ in F such that $X=P_1PP_2$ and $Y=P_1QP_2$.

Derivation: Given a grammar G=(V_N, V_T, S, F) and two words X, Y in $(V_N \cup V_T)^*$, we say that Y is derivable from X, in symbols $X \overset{*}{\underset{G}{\Rightarrow}} Y$, iff X=Y or there is some word Z in $(V_N \cup V_T)^*$ such that $X \overset{*}{\underset{G}{\Rightarrow}} Z$ and $Z \overset{*}{\underset{G}{\Rightarrow}} Y$.
(In other words, the relation $\overset{*}{\underset{G}{\Rightarrow}}$ is the reflexive and transitive closure of $\underset{G}{\Rightarrow}$.)

Language: The language generated by G is defined as

$$L(G) = \{P | S \overset{*}{\underset{G}{\Rightarrow}} P \} \cap V_T^*$$

That is, the language generated by G contains exactly those words (sentences) which are derivable from the initial symbol S and contain only terminal symbols.

(Adapted from Revesz 1983.)

Types of grammars/languages: A generative grammar $G = (V_N, V_T, S, F)$ is said to be of *type i* if it satisfies the corresponding restrictions in the following list:

i = 0: No restrictions [Unrestricted rewriting system/Phrase Structure Grammar (PSG)]
i = 1: Every rewriting rule in F has the form $Q_1AQ_2 \rightarrow Q_1PQ_2$, with Q_1, Q_2 and P in $(V_N \cup V_T)^*$, $A \in V_N$, and $P \neq \varepsilon$ (ε is an "empty" element where for every X, $\varepsilon X = X\varepsilon = X$), except possibly for the rule $S \rightarrow \varepsilon$, which may occur in F, in which case S does not occur on the right-hand sides of the rules. [Context-sensitive PSG]
i = 2: Every rule in F has the form $A \rightarrow P$, where $A \in V_N$ and $P \in (V_N \cup V_T)^*$. [Context-free PSG]
i = 3: Every rule in F has the form either $A \rightarrow PB$ or $A \rightarrow P$, where $A, B \in V_N$ and $P \in V_T^*$. [Regular/Finite-state grammar]

Correspondingly, a language is said to be of type *i* if it is generated by a type *i* grammar. The class of type *i* languages is denoted by L_i. [See Chomsky 1959.]

The Chomsky Hierarchy: L_3 (RL) $\subset$ L_2 (CFL) $\subset$ L_1 (CSL) $\subset$ L_0 (REL) (Chomsky 1963)

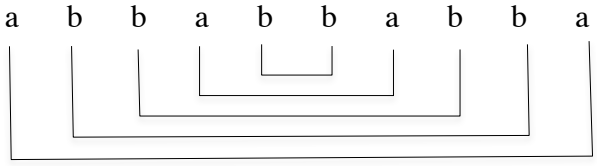
The Pumping Lemma for Regular Languages and why natural language is not regular:

♦ Let L be an infinite regular/finite-state language. Then there are strings x, y, z such that $y \neq \varepsilon$

and $xy^nz \in L$ for each $n \geq 0$.

Eg. 1. The language $L = \{a^nb^n : n \geq 0\}$ is not regular. (Chomsky 1956) Reason: Given the pumping lemma, there are strings x, y, and z with $y \neq \varepsilon$ such that $xy^nz \in L$ for each $n \geq 0$. There are three possibilities for y, and in each case we can show that L must contain some string not of the correct form. Case 1: y consists entirely of a's. Then $x = a^p$, $y = a^q$, and $z = a^rb^s$, where $p, r \geq 0$ and $q, s > 0$. But then L must contain $xy^nz = a^{p+nq+r}b^s$ for each $n \geq 0$, and at most one of these strings has equal numbers of a's and b's. [I.e., all the b's are contained in the z part, and as y is pumped, the number of a's in the string will increase while the number of b's remains the constant. Thus, we will continually be producing strings which have more a's than b's in them, which cannot be in the language a^nb^n.] Case 2: y consists entirely of b's; this is similar to Case 1 (but here the number of b's outstrips the number of a's). Case 3: y contains both a's and b's. Then for $n>1$, xy^nz has an occurrence of b preceding an occurrence of a and therefore cannot be in L. [NB: This language can be generated by the Context-free PSG $G = (V_N = \{S\}, V_T = \{a, b\}, S, F = \{S \rightarrow ab, S \rightarrow aSb\})$.][See Partee, et al. 1990, and other standard textbooks.]

Eg. 2. The language $L = \{xx^R \mid x \in \{a,b\}^*\}$ (where x^R denotes the reversal/mirror image of the string x) is not regular. (Chomsky 1956) In strings of this language, the i^{th} symbol from the left must match the i^{th} symbol from the right:



Reason: This language can be shown not to be regular by first intersecting it with the regular language aa*bbaa* to give $L=\{a^nb^2a^n \mid n \geq 1\}$ and showing that the latter is not regular by means of the pumping lemma. [Since regular languages are known to be closed under intersection (as well as under other operations), if the "mirror-image" language were regular, L would be also, which is not the case.]

☛ Nested dependencies cannot be represented by regular grammars.

Eg. 1. (Chomsky and Miller 1963)

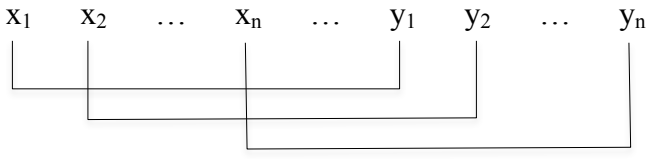
Anyone$_1$ who feels that if$_2$ so$_3$ many more$_4$ students$_5$ whom we$_6$ haven't$_6$ actually admitted are$_5$ sitting in on the course than$_4$ ones we have that$_3$ the room had to be changed, then$_2$ probably auditors will have to be excluded, is$_1$ likely to agree that the curriculum needs revision.

The Pumping Lemma for Context-free PS languages:

♦ Let G a Context-free PSG. Then there is a constant K, depending on G, such that any string *w* in L(G) of length greater than K can be rewritten as $w = uvxyz$ in such a way that either *v* or *y* is nonempty and uv^nxy^nz is in L(G) for every $n \geq 0$.

Eg. 1. The language $L = \{a^nb^nc^n \mid n > 0\}$ is not a Context-free PS language. Proof: omitted (see, for example, Lewis and Papadimitriou 1981: 126) [NB: This language can be generated by the Context-sensitive PSG $G = (V_N = \{S, A, B, C\}, V_T = \{a, b, c\}, S, F = \{S \rightarrow aSBC, S \rightarrow aBC, CB \rightarrow AB, AB \rightarrow AC, AC \rightarrow BC, aB \rightarrow ab, bB \rightarrow bb, bC \rightarrow bc, cC \rightarrow cc\})$.]

Eg. 2. The language $L = \{xx \mid x \in \{a,b\}^*\}$ is not a Context-free PS language. (Chomsky 1956) This language exhibits "cross-serial" dependencies, for strings of length $2n$ in the i^{th} and $(n+i)^{th}$ symbols must match. Proof: omitted. (See for example Partee, et al. 1990: 503) [NB: This language can be generated by a Context-sensitive PSG. See Chomsky 1959, 1963.]



☛ Cross-serial dependencies cannot be represented by Context-free PSGs.

Whether natural language is Context-free or not: Controversial. See, among others, Chomsky 1956, Chomsky 1959, Bar-Hillel and Shamir 1960, Chomsky 1963, Postal 1964, Pullum and Gazdar 1982, Bresnan et al. 1982, Hingginbotham 1984, Shieber 1985, etc.

"… It is not difficult to construct languages that are beyond the range of description of [Σ, F] grammars [our Context-free/Context-sensitive PSG; NF]. In fact, the language L_3 of (12iii) [i.e., L={xx | x∈ {a, b}*}, our Eg. 2. above (the so-called "copying language")] is evidently not a terminal [i.e., PS] language.[9] [note 9 says that "[t]his is not true if the grammar contains rules rewriting a symbol in non-null context…" That is, the language L is generated by a Context-sensitive PSG. See above.] I do not know whether English is actually a terminal language or whether there are other actual languages that are literally beyond the bounds of phrase structure description. Hence I see no way to disqualify this theory of linguistic structure on the basis of consideration (3) [i.e., weak generation]. When we turn to the question of the complexity of description…, however, we find that there are ample grounds for the conclusion that this theory of linguistic structure is fundamentally inadequate." (Chomsky 1956: 116 (page reference is to the 1965 corrected version))

The Weak Generative Capacity of Transformational Grammars:

Transformations ϶ {Adjunction, Substitution, Deletion} (Deletion is subject to the "condition on recoverability" (cf. Chomsky 1965: 144-145, 177ff; Peters and Ritchie 1973: 62)).

The "standard theory" is assumed (particularly, the "cyclic application" of transformations).

Then, we have the following theorems by Peters and Ritchie 1973 (see also Salomaa 1971):

Theorem 5.1 : Every recursively enumerable language is generated by some context-sensitive based transformational grammar, and conversely.

Theorem 5.2 : Every recursively enumerable language is generated by some context-free based

transformational grammar, and conversely.

♦ L_3 (RL) $\subset$ L_2 (CFL) $\subset$ L_1 (CSL) $\subset$ recursive languages $\subset$ L_0 (=recursively enumerable language)

What do these theorems mean for linguistic theory? (Matthews 1979, Perrault 1984)

"The source of nonrecursivity in transformationally generated languages is that transformations can delete large parts of the tree, thus producing surface trees that are arbitrarily smaller than the deep structure trees they were derived from." (Perrault 1984: 167)

We need a stronger condition on deletion operations than the original "recoverability condition" (Peters and Ritchie 1973: 71-72, 81-83). Peters and Ritchie show that if the rules of a transformational grammar meet a certain condition (what Peters 1973: 382 calls the "survivor property"), the language generated by such a transformational grammar will be recursive. Further, Peters 1973 shows that all the rules proposed by linguists do in fact meet this condition.

The more fundamental question is: Must natural language be recursive?

Yes: Putnam 1961, Levelt 1974
Not necessarily: Matthews 1979, Chomsky 1980

"…, it seems that this [i.e., the Peters-Ritchie result] is one kind of argument that might support the thesis of recursiveness, though obviously it would not, and is not intended to show that recursiveness is a logically necessary property of [human] language. In short, while language may be recursive, there is no reason to suppose that this must be so." (Chomsky 1980: 122)

"Notice that it might be a reasonable move to abstract away from grammars and consider only the languages that they generate. Most of the work in mathematical linguistics does just that, and results have been attained that even have some empirical significance. … But it is important

to recognize that when we move from grammar to language we are moving a step further away from the [actual] mechanisms; we are moving to a higher stage of abstraction, and perhaps illegitimately so, …"　(Chomsky 1980: 86)

Chomsky goes on to claim that the question cannot even be properly addressed:

"A third kind of problem that arises when we consider E-languages is that sets have definite properties.　Thus it seems to be sensible to ask whether the E-languages made available by some theory of language have specific formal properties: Are they context-free languages, or recursive languages for which a decision procedure is available, or recursively enumerable languages that can be generated by a finite procedure, or denumerable?　All of these proposals have been advanced, and denied, but what is relevant here is that they are taken seriously. … But the main point to notice is that these discussions are simply not *about* anything, they have no content or substance, until we define the notions E-language and "weak generation" [of natural languages] somehow.　That has never been done, or even attempted, to my knowledge, in any serious way, …" (Chomsky 1988: 47; emphasis original)

◆ If we accept Chomsky's view, then empirically meaningful mathematical studies should be done not about "weak generation," but rather about "strong generation."

👉　E 言語、I 言語、理論、理論の提示、算術の理論 vs. 原子理論、等の関連する議論については、『チョムスキー言語基礎論集』pp. 250-256（およびその前後の議論）を参照。

Strong and Weak Generation

"Let us say that a grammar *weakly generates* a set of sentences and that it *strongly generates* a set of structural descriptions […].　Suppose that the linguistic theory T provides a class of grammars G_1, G_2, …, where G_i weakly generates the language L_i and strongly generates the system of structural descriptions Σ_i.　Then the class $\{L_1, L_2, \ldots\}$ constitutes the *weak generative capacity* of T and the class $\{\Sigma_1, \Sigma_2, \ldots\}$ constitutes the *strong generative capacity* of T."

(Chomsky 1965: 60)

Defining L(G) as the language generated by a grammar G (see above) and Σ(G) as the set of structural descriptions specified by G, we have

WGC (G) = L(G)
WGC (T) = {L(G_1), L(G_2), …}, where T provides {G_1, G_2, …}
SGC (G) = Σ(G)
SGC (T) = {Σ(G_1), Σ(G_2), …}, where T provides {G_1, G_2, …}

Now we can define "equivalence" in weak and strong generative capacity in terms of *identity* (Chomsky and Miller 1963):

-Two grammars G_1 and G_2 are equivalent in WGC iff they generate the same set of strings, i.e., iff $L(G_1)=L(G_2)$.

-Two theories T_1 and T_2 are equivalent in WGC iff for every grammar G_i provided by T_1 there is a grammar G_i' provided by T_2 such that $L(G_i) = L(G_i')$ and for every grammar G_i' provided by T_2 there is a grammar G_i provided by T_1 such that $L(G_i') = L(G_i)$.

-Two grammars G_1 and G_2 are equivalent in SGC iff they generate the same set of structural descriptions, i.e., iff $\Sigma(G_1) = \Sigma(G_2)$.

-Two theories T_1 and T_2 are equivalent in SGC iff for every grammar G_i provided by T_1 there is a grammar G_i' provided by T_2 such that $\Sigma(G_i) = \Sigma(G_i')$ and for every grammar G_i' provided by T_2 there is a grammar G_i provided by T_1 such that $\Sigma(G_i') = \Sigma(G_i)$.

Kuroda 1976: 308: "…formal grammars, in general, can be strongly equivalent without being essentially identical."

Equivalence in terms of isomorphism: Two grammars G and G' are equivalent in SGC if there is an isomorphism between their derivation sets, i.e., a bijective mapping which preserves certain given functions or relations.

SGC(G) = SGC(G') iff, for the functions f_1, …, f_n on Σ(G) and the corresponding functions f'_1, …, f'_n on Σ(G'), there is a bijective mapping Ψ, Ψ: $\Sigma(G) \rightarrow \Sigma(G')$, such that if $\Psi(D_i) = D_i'$, $D_i \in \Sigma$(G) and $D_i' \in \Sigma$(G'), then, for all j, $1 \leq j \leq n$, $f_j(D_i) = f'_j(D_i')$.

Equivalence in terms of homeomorphism: "Given two tree languages, K and K', and a function *f* from K to K', we can ask if, for each set of phrases $\prod$ of K, there is a set of phrases $\prod'$ of K' sufficiently large so that *f* will be continuous from **T**/$\prod$(K) [a topology generated by $\prod$; NF] to **T**/$\prod'$(K'), or from **T***/$\prod$(K) to **T***/$\prod'$(K'). We will say that *f* is structurally continuous (structurally *-continuous) if for each finite set of phrases $\prod$ of K there exists a finite set of phrases $\prod'$ of K' such that *f* is continuous from **T**/$\prod$(K) to **T**/$\prod'$ (K') (from **T***/$\prod$(K) to **T***/$\prod'$(K')); *f* is called a structural *homeomorphism* (**homeomorphism*) if *f* is bijective and if *f* and its inverse are structurally continuous (structurally *-continuous)." (adapted from Kuroda 1976: 314)

強生成力に関する初期の頃からの有力なアプローチとして形式的べき級数（formal power series）の概念を用いるものがある。Chomsky and Schuetzenberger 1963、Gross and Lentin 1970, Salomaa and Soittola 1978 等を参照。

Formal power series:

“Schuetzenberger’s notion of representing sets enumerated by a generative process in terms of formal power series is well motivated for the study of language. As has been mentioned several times, we are ultimately interested in studying processes that generate systems of structural descriptions rather than sets of strings; that is, we are ultimately interested in strong rather than weak generative capacity. The framework just sketched provides a first step toward this goal, since it takes account of the number of structural descriptions assigned to a string (though not the structural descriptions themselves). It also provides a particularly natural way of approaching the study of nondeterministic transduction.” (Chomsky 1963: 406)

実際、1950 年代後半から 1960 年代初頭にかけての研究の全てにおいて、言語学的に有意義な結果はたった 1 つだけだと思います。… その結果とは、文脈自由文法を、（ここ

が肝心なのですが）強い意味で等価な非決定性プッシュダウン・オートマトンへ写像するための構成的手続が存在するという事実です。　（『生成文法の企て』、p. 349）

基本的に、強生成力に関する数理言語学的研究は為されていないことになります。とにかくあまりに複雑になりすぎるんです。実際、強生成力が研究されている唯一のやり方というのは…、構造を持った表現を取り上げて、それを線形連鎖として扱うという方法です。 でも、そうなると弱生成力に戻ってしまい、言語学的動機付けや意義を失うことになってしまいます。　（同書、p. 351）

[註：引用頁等、書誌情報の詳細のいくつかは省略した]

Selected References

Textbooks:

Gross, M. and A. Lentin. 1970. *Notions usr les grammaires formelles*. Gauthier-Villars Éditeur.

Harrison, M. A. 1978. *Introduction to Formal Language Theory*. Addison-Wesley.

Hopcroft, J. E. and J. D. Ullman 1979. *Introduction to Automata Theory*, *Languages, and Computation*. Addison-Wesley.

Kasami, T., et al. 1988. *Keishiki Gengo Riron*. Denshi Joho Tsushin Gakkai.

Lewis, H. R. and Papadimitriou, C. H. 1981. *Elements of the Theory of Computation*. Prentice-Hall.

Partee, B. H., et al. 1990. *Mathematical Methods in Linguistics*. Kluwer Academic Publishers.

Revesz, G. 1983. *Introduction to Formal Languages*. Dover.

Articles/Books:

Bar-Hillel, Y. and Shamir, E. 1960. Finite-State Languages: Formal Representations and Adequacy Problems. *The Bulletin of the Research Council of Israel*, volume 8F, pp. 155-166. [Reprinted in Bar-Hillel, Y. 1964. *Language and Information: Selected Essays on Their Theory and Application*. Addison-Wesley, pp. 87-98.]

Bresnan, J., R.M. Kaplan, S. Peters, and A. Zaenen. 1982. Cross-serial Dependencies in Dutch.

Linguistic Inquiry 13: 613-635.

Chomsky, N. 1956. Three Models for the Description of Language. *IRE Transactions on Information Theory*, IT-2, pp. 113-124. [Reprinted with corrections in Luce, R. D., et al., eds. 1965. *Readings in Mathematical Psychology*, volume II, pp. 105-124.]

Chomsky, N. 1959. On Certain Formal Properties of Grammars. *Information and Control* 2, pp. 137-167.

Chomsky, N. 1963. Formal Properties of Grammars. In Luce, R. D., et al., eds., *Handbook of Mathematical Psychology*, volume II, pp. 323-418.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. The MIT Press.

Chomsky, N. 1980. *Rules and Representations*. Columbia University Press.

Chomsky, N. 1987. Transformational Grammar: Past, Present, and Future. In Chomsky, N. *Generative Grammar: Its Basis, Development and Prospects*, pp. 33-80. Kyoto University of Foreign Studies.

Chomsky, N. and Miller, G. 1963. Introduction to the Formal Analysis of Natural Languages. In Luce, R. D., et al., eds., *Handbook of Mathematical Psychology*, volume II, pp. 269-321.

Chomsky, N. and Schuetzenberger, M. P. 1963. The Algebraic Theory of Context-free Languages. In P. Braffort and D. Hirschberg, eds., *Computer Programming and Formal Systems*. North-Holland, pp. 118-161.

Eilenberg, S. 1976. *Automata, Languages, and Machines*. Volume B. Academic Press.

Higginbotham, J. 1984. English is Not a Context-Free Language. *Linguistic Inquiry* 15, 225-234.

Kuroda, S.-Y. 1974. A Topological Study of Phrase-Structure Languages. *Information and Control* 30, 307-379.

Levelt, W. J. M. 1974. *Formal Grammars in Linguistics and Psycholinguistics*, volume 2, *Applications in Linguistic Theory*. Mouton.

Matthews, R. J. 1979. Are the Grammatical Sentences of a Language a Recursive Set? *Synthese* 40, 209-224.

Perrault, C. R. 1984. On the Mathematical Properties of Linguistic Theories. *Computational Linguistics* 10, 165-176.

Peters, P. S. 1973. On Restricting Deletion Transformations. In Gross, M, et al., eds.

The Formal Analysis of Natural Languages. Mouton.

Peters, P. S. and Ritchie, R. W. 1973. On the Generative Power of Transformational Grammars. *Information Sciences* 6, 49-83.

Postal, P. M. 1964. Limitations of Phrase Structure Grammars. In Fodor, J. and J. Katz, eds., *The Structure of Language*, pp. 137-151. Prentice-Hall.

Pullum, G. K. and Gazdar, G. 1982. Natural Languages and Context-free Languages. *Linguistics and Philosophy* 4, 471-504.

Putnam, H. 1961. Some Issues in the Theory of Grammar. In *Proceedings of Symposia in Applied Mathematics* 12, American Mathematical Society. [Reprinted in Putnam, H. 1975. *Mind, Language and Reality*, pp. 85-106. Cambridge University Press.]

Salomaa, A. 1971. The Generative Capacity of Transformational Grammars of Ginsburg and Partee. *Information and Control* 18, 227-232.

Salomaa, A. and Soittola, M. 1978. *Automata-theoretic Aspects of Formal Power Series*. Springer-Verlag.

Shieber, S. 1985. Evidence Against the Context-freeness of Natural Language. *Linguistics and Philosophy* 8, 333-343.