Witold Kieraś

SCHWYZERTÜÜTSCH, BAMBARA, AND CONTEXT-FREE LANGUAGES

Originally published as "Schwyzertüütsch, bambara i języki bezkonstekstowe," Studia Semiotyczne 27 (2010), 135–149. Translated by Rafał Jantarski.

1. INTRODUCTION

In his oft-cited, if rarely read book, Noam Chomsky (Chomsky 1956) ventured to propose a hierarchy of formal grammars. He defined them as rules for rewriting strings of terminal and nonterminal symbols into different strings of terminal and nonterminal symbols, thus giving birth to what is today known as the Chomsky hierarchy. The author himself claimed the theory applied exclusively to formal languages, but in his considerations he also happened to formulate a problem relating to natural languages, namely, to which formal grammar category descriptive grammars of natural languages belong? While restraining from siding with any answer to the problem, Chomsky sparked a long philosophical and linguistic discussion revolving around complex structures in natural languages.

At the very beginning of his considerations, Chomsky rejected two categories of the hierarchy, judging them inadequate for description of natural languages: type-0 grammars (unrestricted grammars, generating recursively enumerable languages), and type-3 grammars (regular grammars). The first was deemed too broad. Assuming, which Chomsky does, that adequate description of natural languages can be done via formal methods, the fact that natural languages can be described by unrestricted grammars to generate any formal language is practically useless. Regular grammars are rejected as too weak, a claim first introduced in *Syntactic Structures* (Chomsky 1957), an earlier book which appeared in print later than the hierarchy theory. The problem, therefore, relates to only two levels: context-free and

context-sensitive grammars.

2. MOTIVATIVES

There are several reasons why it would be interesting to ask whether context-free grammars are all we need to describe any natural language.

The first one is purely practical: essentially, recognition (and parsing) of context-free languages is multinomial, i.e. practically calculable. For context-sensitive languages, however, such calculation is problematic. In the IT industry, from data search to automated translation, there is an increasing need for processing texts and recordings that appear in natural languages. It would be good to know whether our computing capabilities are robust enough to handle natural languages.

Second, there are some general theoretic questions concerning the relationship between the traditional linguistic description and formal description. Having established purely formal properties of natural languages, we may be well positioned to try to judge whether their respective descriptive grammars are adequate.

Third, there are philosophical questions regarding mechanisms and computational capabilities of the human brain. If, on a daily basis, we use structures that machines are unable to process in a reasonable time, one may come to the conclusion that the brain mechanisms responsible for linguistic competences have computational capabilities far exceeding those of machines.

Note, however, one thing. Any (valid or invalid) argument for natural languages not being context-free always points to a specific syntactic structure in a specific language. Implicit or explicit conclusions drawn from that kind of reasoning are that finding one natural language that defies complete description in context-free grammar must mean by necessity that no natural language is ever context-free. This assumes linguistic universalism, inherent to both Chomsky's concept and generativism at large (Mecner 2005), although many linguists, psychologists and philosophers find universalism controversial. That said, it is not without reason to explore whether formal languages can be useful in describing natural languages, even if one is not subscribing to universalist inclinations.

3. INVALID ARGUMENT — ENGLISH COMPARATIVE

Early on in the discussion, some contributors formulated arguments that Gazdar and Pullum billed as folklore (Pullum, Gazdar 1982). It is not

the goal of this paper to discuss invalid arguments, but it might be of use to mention one less banal example, particularly because it was provided by Chomsky himself.

One popular artificial context-sensitive language is the *xx*-language. Assume a nonempty alphabet $\Sigma = \{a,b\}$. Language $\mathrm{L} = \{xx : x \in \Sigma^+\}$ is an *xx*-language. More generally, *x*-strings can be separated by any given string of symbols, but for a language to belong to the *xx*-type it needs to have two identical strings of alphabetical symbols (Hopcroft et al. 2006). ***xx***- and similar languages are often used to demonstrate that some language phenomena are not context-free.

In 1963, Chomsky argued that the syntax of English comparative is context-sensitive. Consider the following:

(1) That one is wider than this one is deep.

Chomsky argues that it is ungrammatical to say

(2) *That one is wider than this one is wide,

its grammatical equivalent being

(3) That one is wider than this one is.

He concludes that sentences with recurring adjectives[1] are incorrect, and that the right form requires a different adjective in each part of the sentence. He then goes on to argue that English comparative is not context-free because it creates an *xy*-language in which two constituent parts must differ and which very much resembles an *xx*-language. Let's assume a vocabulary $\Sigma = \{a,b,\alpha,\beta,\gamma\}$.

$$L' = \{\alpha x \beta y \gamma : x,y \in L \wedge x \neq y\}$$

is an *xy*-language, where L is any language with words consisting of a and b. Chomsky claims that if L' is context-sensitive, then so is English, by virtue of having such structures as (1). But he does not provide any arguments to substantiate the claim that *xy*-languages are context-sensitive, although his reasoning perhaps works on the implied premise that *xy* is context-sensitive

[1]Note that one adjective is in comparative, while the other is not, although in this particular example the difference is irrelevant.

by the same token as *xx* is. Gazdar and Pullum, however (Pullum, Gazdar 1982), came up with the following context-free grammar for an *xy*-language:

(4) a. S → $aS'\gamma$|aS"γ
b. S' → $CS'C|D\beta|\beta D$
c. S'' → $AB'|BA'$
d. A→ $CAC|a(D)\beta$
e. B→ $CBC|b(D)\beta$
f. A' → $a(D)$
g. B' → $b(D)$
h. C → $a|b$
i. D → $C(D)$

This makes *xy*-languages context-free and Chomsky's argument goes by the board.

Gazdar and Pullum note that we can generate an infinite number of independent grammars for any given language, therefore to argue convincingly that a language is context-sensitive one cannot demonstrate that all its grammars are non-context-free. However, trivial as this remark may be, many authors quoted by Gazdar and Pullum seemed to have overlooked this simple fact. Now, a convincing argument that a certain language is context-sensitive needs to depend on formal properties of context-free languages. The three most popular are:

– pumping lemma for context-free languages,
– closure under homomorphism,
– closure under intersection with regular languages.

It is particularly the latter that produces the most powerful and popular arguments, those being Shieber's argument based on Swiss German, the argument based on Dutch, and Culy's argument based on Bambara spoken in Mali.

4. SCHWYZERTÜÜTSCH AND SHIEBER'S ARGUMENT

Schwyzertüütsch (also: Schwyzerdütsch or Schweitzerdeutsch) is a group of Allemanic dialects of German used in Switzerland and Lichtenstein. They dominate the spoken language, while Standard German remains the preferred option in writing (although St. Gallen and Zürich publish books in

local Allemanic dialects). In his 1985 paper, Shieber explored some interesting syntax structures in Schwyzertüütsch which are not found in Standard German. These are present only in subordinate clauses and concern syntax requirements for verbs. Much like in Polish, verbs in Schwyzertüütsch may require, apart from the subject in the nominative, that other phrases also have specific grammatical case, namely accusative and dative. Further, some verbs may require other verbs (also a familiar Polish feature). This is why in subordinate clauses there are characteristic strings of verbs, preceded by the string of valence requirements in specific cases. Let's now consider a few examples (since the focus is on subordinate clauses, we may assume that each sentence begins with *Jan säis das...*, which means "*Jan says that...*"):

(5) ...mer em Hans es huus hälfen aastriiche
...we Hans-DAT house-ACC helped paint
'...we helped Hans paint the house.'

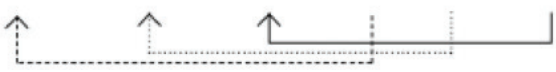
At the end of the phrase there are two verbs, *hälfen* and *aastriiche*, each requiring a different case, respectively dative and accusative. In Schwyzertüütsch, and particularly in the example above, case exponents are such words as *em* or *es*.

(6) ...mer d'chind em Hans es huus
...we the children-ACC Hans-DAT house-ACC
lönd hälfeaastriche.
let help paint.
'... we let the children help Hans paint the house.'

In (6), there are three verbs at the end of the phrase: the first requires accusative, the second — dative, and the third — accusative. In theory, there are no limits for construction of such sentences. Examples (7) and (8) graphically illustrate relationships in (5) and (6).

(7) ...merem Hans es huus hälfen aastriche.

(8) ... mer d'chind em Hans es huus lönd hälfe aastriche.



Knowing how the structure works, we may now read through Shieber's argument. Having provided linguistic data, Shieber enumerates the following four properties of Schwyzertüütsch:

– Swiss-German subordinate clauses can have a structure in which all the verbs follow all the noun phrases;
– Among such sentences, those with all dative noun phrases preceding all accusative noun phrases, and all dative-subcategorizing verbs preceding all accusative-subcategorizing verbs are acceptable;
– The number of verbs requiring dative objects must equal the number of dative noun phrases and similarly for accusatives;
– An arbitrary number of verbs can occur in subordinate clauses such as (5) or (6).

Now, assume any given language L that satisfies these claims (e.g. Schwyzertüütsch) and contains sentences such as (5). Consider the following homomorphism f, where

$$
\begin{aligned}
f\ (\text{"d'chind"}) &= a \\
f\ (\text{"em Hans"}) &= b \\
f\ (\text{"lönd"}) &= c \\
f\ (\text{"hälfe"}) &= d \\
f\ (\text{"Jan säitdasmer"}) &= w \\
f\ (\text{"eshuus"}) &= x \\
f\ (\text{"aastriche"}) &= y \\
f\ (s) &= z \text{ - otherwise}
\end{aligned}
$$

When intersecting $f(L)$ with the regular language $\mathrm{r} = wa^*b^*xc^*d^*y$, we arrive at the language $f(L) \cap \mathrm{r} = wa^m b^n x c^m d^n y$ which does not fit into the context-free category (it's a classic example of context-sensitive language, see Hopcroft et al. 2006).

Since context-free languages are closed under homomorphism and intersection with regular languages (see Hopcroft et al. 2006), also L is not context-free. Therefore, also languages that contain structures like (5) are not context-free. This concludes Shieber's argument.

5. DISCUSSION

In his paper, Shieber identified several counterarguments to challenge his

reasoning, and tried to refute them. It seems, however, that he wasn't really committed to the task and didn't fully explore their disruptive potential.

5.1 CASE IS NOT SYNTACTIC

One of the potential counterarguments offered by Shieber goes as follows: maybe verb case-marking (used to make the Schwyzertüütsch argument) is of semantic, not syntactic, nature. This would naturally imply that a context-free grammar could be used to describe Swiss German. This, however, is at odds with traditional research in inflectional languages. As far as inflection goes, Schwyzertüütsch is closer to Polish than to English, and both Polish and German linguists consider case marking as a syntactic issue. Making a semantic problem out of it is very much possible, but highly problematic.

One may of course go as far as to claim that inflection, semantics, or word formation is merely a matter of convention and preferred point of view. But even those leaning toward semantic interpretation of various structures in Polish wouldn't go as far as to consider case-marking in semantic terms. This seems to chime with everyday intuitions. Sentences where verb valence does not correspond with subject or object, e.g.

(9) * Jaś lubi jabłek. [John likes apples(gen.)]

are not perceived as semantically derailed, but grammatically incorrect. Without stirring much controversy one may argue that a regular user of Swiss German has a comparable perception of similar structures in his native speech. Shieber approached his informants with similar examples, and all deemed them ungrammatical (not just semantically bizarre).

5.2 OTHER CONSTITUENT ORDERS ARE POSSIBLE

Schwyzertüütsch, much like Polish, doesn't have a strictly fixed order, therefore the examples above explore just one from among a number of acceptable variations. Furthermore, there are reasons to believe that other orders are more natural. That is not to say that the structure itself is incorrect, Schwyzertüütsch permits such order and its grammar should be robust enough to describe it. Shieber's argument holds even if other structures are possible; his method is to consider one subset of sentences in Swiss German, concluding that it cannot be generated by context-free grammars.

Let's now turn to the pragmatic implications of this situation. Shieber's argument is informed by the idea that the verb string will be ordered according to the case — first go all verbs with dative, then all with accusative (or the other way round). Noun phrases must be ordered accordingly. It is not required, however, that the noun phrase required by *i*-verb is positioned on the *i*-slot in the string of noun phrases. Shieber himself provides examples where noun phrases swap their slots, while their respective verbs stay in the previous order. In highly inflectional languages, unconstrained word order is a fairly common occurrence. But if this is the case, sentences considered by Shieber are extremely ineffective in terms of pragmatics, as structures have neighbouring phrases in the dative and accusative. The number of possible syntactic interpretations of such a sentence rises exponentially, relative to the length of verb strings, because each verb needs to be interpreted against each phrase that has the required case. Hence, they will hardly ever be used in real life, the sheer number of possible interpretations making it pragmatically inefficient.

5.3 CLAUSES ARE BOUNDED IN SIZE

Another counterargument provided by Shieber (later accommodated by others) is that verb strings are limited in number — which would mean that structures could work under a context-free grammar. Indeed, it would be rather unusual to use more than five verbs in a single sentence. But if we were to further this reasoning, we would be quickly compelled to conclude that the natural language structures that we perceive as recurrent and potentially infinite are, in fact, finite and constrained. We may even go as far as to conclude that there is, say, an upper limit of simple sentences that can be linked with coordinators. But it would be equally legitimate to say that, since natural languages are finite, they can be described by both context-free and regular grammars. Such a defense of context-freeness is however difficult to accept.

Further, one must separate two things: adequate theoretical description of a language and implementation of the theory in question. Implementation permits simplification due to technical limitations, but a robust theory should be free from such shortcuts.

One more pragmatic remark: extension of such Swiss German structures is possible by application of verbs with specific valence requirements. They must be able to link to a noun phrase in its specific case and another verb in the infinitive. Polish has only a handful of those, and one may assume

that Schwyzertüütsch is not entirely different. Again, building longer structures of this sort is pragmatically (although not grammatically) constrained. But again, given the pragmatic constraints outlined above, the phenomenon in question is extremely rare.

6. SIMILAR ARGUMENTS

Let me now reiterate arguments based on Dutch and Bambara, two languages often analyzed in this context. Dutch came to attention early on in the discussion, but the resulting arguments were often dismissed as not being entirely relevant. I will present a later version of the Dutch argument, but each of those builds on cross-seriality, also present in Swiss German. Bambara provides another interesting example, with its arguments being not of syntactic, but morphologic nature.

6.1 CROSS-SERIALITY IN DUTCH

One of the earliest examples of cross-seriality was found in Dutch. Several authors explored structures that are quite similar to the ones existing in Schwyzertüütsch, but for various reasons those interpretations were challenged (Pullum, Gazdar 1982, among others). Dutch arguments may not be adding anything new to what has already been said in relation to Swiss German, but I shall nevertheless briefly discuss one of them, provided by Alexis Manaster-Ramer (Manaster-Ramer 1987).

In Dutch, cross-seriality occurs, like in Schwyzertüütsch, in subordinate clauses and in certain types of interrogative. For the sake of greater argumentative diversity, I will focus on the latter. Consider the following:

(10) of Jan Piet Marie zag kussen?
Did John Peter Mary saw kiss?
Did John see Mary kiss Peter?

As we can see, there is a verb string preceded by noun phrases satisfying valence requirements of those verbs. Adding a structure with coordination, we arrive at the following:

(11) Of Jan Piet Marie horde ontmoeten en zag omhelzen?
Did John Peter Mary heard meet and saw embrace?
Did John hear that Peter met Mary and embraced her?

The important thing in this example is the relationship between the number of noun phrases (*Jan*, *Piet*, *Marie*) and verb phrases in two verb strings (*horde ontmoeten* and *zagomhelzen*). We have two intersecting structures here: cross-seriality between verbs and their valence requirements, and coordination. The result is the following interrogative in Dutch:

$$\mathrm{NP}^n \quad \mathrm{V}^n \quad \& \quad \mathrm{V}^n.$$

Let's consider the following language:

$$L = \{\text{Of Jan } N^n \text{ Marie horde } V^n\text{ontmoeten zag } W^n\text{omhelzen}\},$$

where N = {Joop, Alexander, Jan, Wim, Piet, Marie, Willem, ...), V = {horen, zien, helpen}, W = {laten, leren}. We can easily see that, by applying a homomorphismtransposing N, V, W from L into a, b, c and other symbols from L into the empty symbol $\in$, we arrive at the language $a^n b^n c^n$, which is context-sensitive (more specifically, it belongs to index languages, a subcategory of context-sensitive languages, see Hopcroft et al. 2006).

6.2 MORPHOLOGY IN BAMBARA

Bambara, or Bamana, a Niger-Congo language belonging to the Mande group, is used primarily in Mali. It is spoken by ca. 2.7 million people, with another four million using it as lingua franca. Bambara inspired Christopher Culy (Culy 1985) to come up with a proof that language generating word-formation structures in Bambara (i.e. language over the set of morphemes) is not context-free. However, in the context-freeness dispute this argument is weaker because the generative syntax theory (under which the problem was formulated) assumes that the vocabulary (all possible word structures in the language) is already given. It is therefore of no interest to those preoccupied with sentence structures or parser constructors.

Culy combines two word formation constructions from Bambara to make his case. In first, the noun is duplicated to create a non-definite structure. Two identical nouns are separated by the morpheme *o*, giving N *o* N, which translates into "whatever N" or "whichever N."

(12) wulu o wulu
dog dog
"whichever dog"

(13) malo o malo

rice rice
"whichever rice"

(14) *wulu o malo
dog rice

The above examples show that on both sides of *o* there must be the same noun, other configurations must be dismissed as incorrect.[2]

The other interesting structure in Bambara is an agentive structure N + TV + *la*, which translates into "*one who TVs Ns*".

Consider the following examples:

(15) wulu + nyini + la = wulunyinina
dog search for
"one who searches for dogs" i.e. "dog searcher"

(16) wulu + filé + la = wulufiléla
dog dog watch
dog "one who watches dogs" i.e. "dog watcher"

(17) malo + nyini + la = malonyinina
dog rice search for
dog "one who searches for rice" i.e. "rice searcher"

(18) malo + filé + la = malofiléla
dog rice watch
dog "one who watches rice" i.e. "rice watcher"

Words in (15) and (17) end with *na*, not with *la*, because some sound clusters in Bambara change *l* to *n*, but this morphological phenomenon is irrelevant to the argument.

Agentive structure is recursive, which means that produced words produced can in turn function as its arguments:

(19) wulunyinina + nyini + la = wulunyinina nyinina
dog searcher search for
"one who searches for dog searchers"

(20) wulufiléla + nyini + la = wulufilélanyinina

[2]Generally, reduplication frequently occurs in morphology of many languages (both in word formation and flection). The above is an example of full reduplication (one repeats the whole word, as opposed to partial reduplication, where one repeats e.g. a morpheme). In Indo-European languages full reduplication doesn't occur, but elsewhere it is quite common and serves various purposes.

dog watcher search for
"one who searches for dog watchers"

Nouns formed in the second structure can be then embedded in the first structure:

(21) wulunyinina o wulunyinina
dog searcher dog searcher
"whichever dog searcher"

(22) wulunyininanyinina o wulunyininanyinina
one who searches for one who searches for
dog searchers dog searchers
"whoever searches for dog searches"

And so forth...

Thus, we arrive at a structure similar to the one explored by Shieber in his Schwyzertüütsch argument. Let B be Bambara vocabulary (a complete set of words, and by extension a set of morpheme strings). Let R be the following set:

$$R = \{\text{wulu(filéla)}^h\text{(nyinina)}^i \quad \text{o} \quad \text{wulu(filéla)}^j\text{(nyinina)}^k\text{:}$$

$$h,\, i,\, j,\, k \geqslant 1\}$$

Intersection of B and R produces the following:

$$B\text{'} = B \cap \text{r} = \text{wulu(filéla)}^m\text{(nyinina)}^n \quad \text{o} \quad \text{wulu(filéla)}^m\text{(nyinina)}^n\text{:}$$

$$m,\, n \geqslant 1\}$$

B' has the general from of $\{a^m b^n a^m b^n\text{: } m,\ n \geqslant 1\}$, which makes it a context-sensitive language. Context-free languages are closed under intersection with regular languages (R being regular), therefore, if B' is not context-free, then neither B can be context-free.

As indicated before, this argument is considered to be weaker, as is concerns morphology rather than syntax. Note, however, that such structures are very rigid in terms of word order, while syntax structures based on case requirements permit a more liberal approach, as, traditionally, word order in inflectional languages is more free.

There is another reason tohave a closer look at Culy's argument as he identifies an interesting problem: in which category under Chomsky's

hierarchy one should classify not only syntax of natural languages, but also other linguistic subsystems. In a brief answer to that question we should first note that subsystems have different complexity. The lowest level, phonetics, is undeniably the simplest subsystem of them all. Ever since the emergence of modern linguistics, phonetics has occupied a separate place within the field. This is because it focuses on a relatively small and finite number of units, making the calculation, from Culy's point of view, the least problematic. The next level is morphology: inflection and word formation. In Polish, automatic inflectional analysis is performed via finite-state machines (i.e. regular grammars), a practical and effective solution to the problem. Derivative (word-formation) analysis in Polish has so far made little progress, but there is nothing to indicate that it would be of greater complexity than inflectional analysis. Another linguistic subsystem, syntax, is explored in the greater part of this paper. Undoubtedly, syntax analysis would require at least the power enabled by context-free grammars. Some arguments presented in this paper show that natural language structures are too complex for context-free grammars. Finally, semantics: this level seems the most complex of all, but the level of complexity remains difficult to estimate. It seems interesting (but not entirely surprising) that subsequent levels of linguistic systems show increasing complexity.

7. SUMMARY

The paper has explored the most popular arguments against context-freeness of natural languages. It is interesting to note that for the greater part of the 50 years since formulation of this problem it has been tacitly assumed, without really demonstrating it, that natural languages require power that only context-sensitive languages can provide. However, when the matter is attended to with due care, it suddenly appears that this perception is far from self-evident. Further, the strongest arguments supporting the case are derived from rather exotic languages. Take Shieber's argument, which bases itself on a language that has almost no presence in writing and exists exclusively in its spoken variation. Therefore, even if, generally, natural languages require the power of context-sensitive grammars, this covers linguistic phenomena occurring rather infrequently, which at the end of the day makes the problem negligible in practical applications. As it is, context-sensitiveness in languages is rather rare — over the years quite a large group of linguists and philosophers tried to find it, but so far the results have been modest to say the least. This would mean that it is a rather undesired phenomenon — perhaps due to difficulties that our brains

experience while processing such structures.

Bibliography

1. Chomsky, Noam (1956) "Three Models for the Description of Language." *IRE Transactions on Information Theory*, vol. IT-2, No. 3: 113-124.

2. Chomsky, Noam (1957) *Syntactic Structures*. The Hague: Mouton & Co. Publishers.

3. Chomsky, Noam (1963) "Formal Properties of Grammars." In *Handbook of Mathematical Psychology*, Vol. II, Robert D. Luce, Robert R. Bush, Eugene Galanter (eds.). New York: Wiley.

4. Culy, Christopher (1985) "The Complexity of the Vocabulary of Bambara." *Linguistics and Philosophy* 8: 345-351.

5. Hopcroft, John E., Rajeev Motwani and Jeffrey D. Ullman (2006) *Introduction to Automata Theory, Languages, and Computation*. Boston: Addison-Wesley.

6. Manaster-Ramer, Alexis (1987) "Dutch as a Formal Language." *Linguistics and Philosophy* 10: 221-246.

7. Mecner, Piotr (2005) *Elementy gramatyki umysłu*. Kraków: Universitas.

8. Pullum, Geoffrey and Gerald Gazdar (1982) "Natural Languages and Context-Free Languages." *Linguistics and Philosophy* 4: 471-504.

9. Schieber, Stuart (1985) "Evidence Against the Context-Freeness of Natural Language." *Linguistics and Philosophy* 8: 333-343.