C. CULY

FORMAL PROPERTIES OF NATURAL LANGUAGE AND LINGUISTIC THEORIES

1. Introduction

Over the past 40 years, various formal properties of natural language and linguistic theories have been studied, most notably weak generative capacity (cf. Savitch et al. 1987)[1] and time complexity (Barton et al. 1987). However, the standard closure properties (union, intersection, complementation, concatenation, concatenation closure, homomorphism, inverse homomorphism, and intersection with a regular set) of natural language have not been studied.[2] A novel proof technique will be used to argue that natural language is not closed under any of these operations. Finally, the ramifications of these non-closure facts for linguistic theory will be discussed.

2. Definitions

Before turning to the non-closure properties of natural language, it is necessary to clarify what aspect of natural language I am considering. I will be considering languages as sets of strings of the lexical categories ("pre-terminals") corresponding to well-formed sentences. I will assume that there is a finite set of lexical categories that languages can draw from. Furthermore, the interpretations of the lexical categories are the same across languages. For example, if N is the category "noun" in English, then it will be the category "noun" in Bambara, French, Yoruba, etc.

[1] I'd like to thank Chris Albert, Oduntan Bode, Rob Chametzky, Matthew Dryer, Ron Kaplan, Bill Ladusaw, Alexis Manaster Ramer, John Mugane, Stanley Peters, David M. W. Powers, Vaughn Pratt, Geoff Pullum, Daniel Radzinski, Johan van Benthem, Tom Wasow, members of the LINGUIST electronic List, and audiences at UC Santa Cruz, Stanford, and MOL 4 for helpful discussion. None of them are responsible for any errors. A portion of this work was done while I was a visiting scholar in the Linguistics Department of Stanford University, to whom I am very grateful. An earlier version of this paper was presented at MOL 4 (Mathematics of Language).

[2] Ristad (1986) makes the same claim for union, concatenation, Kleene closure, substitution, and intersection with regular sets, but provides no proof. Thanks to Geoff Pullum for bringing this to my attention.

Linguistics and Philosophy **19**: 599–617, 1996.

This view of language is somewhat unusual in the mathematical linguistics literature, though not unprecedented (cf. Kornai and Pullum 1990). Most arguments concerning the properties of natural language consider languages to be sets of strings of words. The reason for choosing lexical categories over words here is that syntax refers to grammatical categories and not individual words, and it is really the syntax of natural language that I am interested in.

Furthermore, we can note that even where strings of words have been considered in previous work on the generative capacity of natural language syntax, many other strings of words could have been used. For generative capacity of natural language syntax, what matters is not the particular words that are used, but the classes (categories) that they represent. Readers who are uncomfortable with this assumption can pick representatives of the lexical classes to carry the arguments over to the word level.[3]

This view of language is also crucial in making the closure properties interesting. For example, it is clear, or so it seems, that the union of the sentences of two natural languages will in general not be another (possible) natural language. This has as much to do with the lexical peculiarities of languages as anything else. However, once we abstract away from the particular words used, it is no longer as obvious that languages are not closed under, for example, union.

Obviously, natural language sentences have structure, which I am ignoring. In this regard, the enterprise here is parallel to the study of weak generative capacity. If the non-closure results follow without reference to structure, then we have established a strong result about natural language based on the weakest of representations.[4]

I will also not consider the categories in direct quotations to be part of the strings to be considered, e.g. I said "You should be glad". The two reasons for this are that we are interested in the use of categories, not their mention, and that non-language can be quoted, e.g. I said "srkrvbntqvn".

There are two further assumptions about language that I will be making. In particular, I will assume that all sentences consist of at least one category (excluding the empty string as a sentence), and that every lan-

[3] Of course, some care does need to taken to pick appropriate instances of the lexical categories.

[4] It is a separate question whether the sets of structures of natural languages are closed under the various operations. Most of the (relevant) arguments here can be carried over to structures with little modification. See below for further discussion.

guage consists of at least one sentence (excluding the empty set as a natural language).[5]

A couple more definitions are in order as background. I will be considering the set of all natural languages. A set of languages with at least one non-empty language is called a family of languages (not to be confused with a group of historically related languages). I will show that natural language (as a family of languages), is not closed under a variety of operations, which will make it an anti-AFL (abstract family of languages). The definitions for AFL and anti-AFL are given in (1–2).[6]

(1) Abstract family of languages (AFL)
A family of languages closed under union, (concatenation,) positive concatenation closure ("Kleene +"), non-erasing homomorphism,[7] inverse homomorphism, and intersection with a regular set. An AFL is a full AFL if it is closed under "Kleene *" and arbitrary homomorphism.

(2) Anti-AFL (cf. Salomaa 1973: 237)
A family of languages which is not closed under the operations of union, positive concatenation closure, non-erasing homomorphism, inverse homomorphism, and intersection with a regular set (i.e. a family of language not closed under any of the AFL operations with the possible exception of concatenation).

All of the more familiar families of languages are AFLs, and most are full AFLs, as seen in (3). In addition, several of them are closed under intersection. In this light, it is rather striking that natural language is an anti-AFL. It suggests that natural language is in some sense very different from these other families of languages.

[5] Cf. Wasow 1978 who thinks all finite sets should be excluded as possible natural languages.
[6] Cf. Salomaa 1973. Hopcroft and Ullman 1979 include concatenation as one of the AFL operations. Since the operations are not completely independent, the difference in definitions is not important. I have chosen Salomaa's, since the definition of anti-AFL depends on his definition of AFL.
[7] A homomorphism h is non-erasing if for all x, $h(x) \neq e$ [the empty string].

(3) Properties of some families of languages

Family	AFL	Full AFL	Closed under intersection	Closed under complementation
Regular	√	√	√	√
Context free	√	√		
Index	√	√		
Context-sensitive	√		√	√
Recursive	√		√	√
Recursively enumerable	√	√	√	

3. Non-Closure Properties of Natural Language

In this section I will argue that natural language is not closed under any of the standard operations. In order to prove this statement, I would have to have knowledge of the nature of all (possible) natural languages, which is easier said than done. All of the proofs will be of the form: If natural language[8] has property P, then it is not closed under operation O. If these properties do indeed hold of natural language, then I have proven something about natural language. Of course, other properties could also lead to the same results. I will start with the non-closure properties which are easiest to argue for, and hence have the strongest arguments in their favor, and work my way up to the more difficult ones.

Before turning to the first non-closure property, we will need the following property of natural languages.

NL UNIVERSAL 1. No natural language consists only of sentences of strings containing only one category (i.e. a subset of C^+), where C is some category.

Note that while most, if not all, languages will have some sentences consisting of only one category (e.g. English "Go hide!"), the claim here is that no natural language will consist only of such sentences. Implicit here is the claim that every natural language will make use of more than one lexical category.[9]

[8] Or any family of languages for that matter.

[9] The two types of lexical categories likely to be found in all languages are some kind of noun and some kind of verb.

We can now argue that natural languages are not closed under intersection with regular sets, even if the intersection is infinite.

PROPOSITION 1. Natural languages together with the empty set are not closed under intersection with finite or infinite regular sets, even when the result of the intersection is infinite.

ARGUMENT. There are two cases, where the regular set is finite[10] and where it is infinite. In both cases the regular set will be a subset of a set of the form C^+.

Case 1: Finite regular set. Let $\mathbf{R} = \{V\}$. Let **L** be English. Then $\mathbf{L} \cap \mathbf{R} = \mathbf{R}$ (cf. Stop!). But by NL Universal 2, **R** is not a natural language.

Case 2: Infinite regular set. Now let $\mathbf{R} = (V)^+$. Let **L** be Donno Sɔ. DS has sentences that consist of sequences of verbs, as shown in (4).

(4) Sequences of verbs as sentences in DS

Verbs	Example	Gloss
1	Bojɛu	You are going
2	Bojɛu gim	I said you are going
3	Bojɛu gim giu	You said I said you are going
4	Bojɛu gim giu gim	I said you said I said you are going

These sentences can have any number of verbs, so $\mathbf{L} \cap \mathbf{R} = \mathbf{R}$. But by NL Universal 1, **R** is not a natural language. √

NL Universal 1 will also provide an argument against closure under homomorphism. Non-closure under substitution by regular sets follows immediately from non-closure under homomorphism.

PROPOSITION 2. Natural languages are not closed under homomorphism.

ARGUMENT. Let $h(x) = C$, some lexical category, for all x, where x is a lexical category. For any natural language L, $h(L) \subseteq C+$, which is not

[10] Note that if we did not include the empty set with the natural languages, we would have a simple proof of the finite case, since the empty set is regular, and is the intersection of itself with any NL, and we have excluded the empty set as a possible natural language.

a natural language by NL Universal 1. We can also note that *h* is non-erasing. √

COROLLARY. Natural languages are not closed under substitution (by regular sets).

Another property of natural language is needed before moving on to more non-closure properties.

NL UNIVERSAL 2. There are lexical categories which cannot constitute a sentence alone, e.g. complementizers, plural markers, and case markers.

While not all languages use these lexical categories (e.g. English does not use the plural marker), in languages that do, these categories cannot be sentences themselves. As an illustration of a plural marker, consider the Bambara examples in (5). The plural marker is *w* (phonetic [*u*]), and occurs at the end of the noun phrase.[11]

(5) Plural marker in Bambara

jakuma = cat	jakuma w = cats
jakuma finman = black cat	jakuma finman w = black cats

An example of a case marker is given in (6). The Donno So (Dogon) possessive is marked by *mɔ̃*, which occurs after the possessed NP (cf. Embree 1993).

(6) Possessive case marker in Donno So (Kervran 1982: 33)

Indɔ mɔ̃ kubɔ le numɔ le
person.def poss foot and hand and
bɛbaa boli
become without strength go.pst
This person's feet and hands became without strength

We can now argue against closure under inverse homomorphism.

PROPOSITION 3. Natural languages are not closed under inverse homomorphism.

[11] Note that there is no phonetic or phonological reason why it should not occur alone, since it is homophonous with the third person plural pronoun.

ARGUMENT. Let h be as follows:

$h(C) = NV$, where C is one of the lexical categories from NL Universal 2 which cannot be a sentence.[12]

$h(x) = x$, x a lexical category different from C

Recall that h^{-1}(English) = $\{s \mid h(s)$ is English$\}$. In particular we can note that $h(C) = NV$, which is in English (cf. Pat left). Thus, h^{-1}(English) contains C, which is not a well formed sentence in any natural language by NL Universal 2. This makes h^{-1}(English) not a natural language. √

NL Universal 2 also allows us to show that natural languages are not closed under complementation. Proposition 4 is actually a slightly stronger result.

PROPOSITION 4. If U is the universal set of lexical categories, and L is a natural language, then $U^+ - L$ is not a natural language. In other words, the complement of any natural language is not a natural language.

ARGUMENT. Let C be one of the categories from NL Universal 2 which cannot be a sentence, and let L be any natural language. By hypothesis, $C \in U$, the universal set of lexical categories. Since C is not a possible sentence, $C \in U^+ - L$. But again since C is not a possible sentence, $U^+ - L$ is not a possible natural language. √

That natural languages are not closed under complementation follows immediately from Proposition 4, since the complement of any particular language is not a natural language.

COROLLARY. Natural languages are not closed under complementation.

Turning now to non-closure under concatenation, we can make the following observation about natural languages.

NL UNIVERSAL 3. There is an upper bound on the minimum sentence length in natural language.

Every language has a minimum sentence length. For English, it is one (cf. Go!). There may be some languages with a minimum sentence length of

[12] For concreteness, we could take C to be Pl[ural].

2 or possibly even 3, for example if auxiliaries are always required in a particular language. However, no language will have only sentences consisting of, e.g. at least 10 categories.

PROPOSITION 5. Natural languages are not closed under concatenation.

ARGUMENT. Let n be the upper bound on minimum sentence length. By the assumption that all sentences consist of at least one category, $n \geq 1$. Let **L** be a natural language whose minimum sentence length is n. **LL** is not a natural language, since its minimum sentence length is $2n > n$, which contradicts the hypothesis that n is the upper bound on minimum sentence length. √

NL UNIVERSAL 4. No natural language allows the orders Noun Demonstrative Numeral Adjective and Noun Adjective Numeral Demonstrative within the NP, but not some other order of Demonstrative, Numeral, and Adjective after the noun in the NP.[13]

PROPOSITION 6. Natural languages are not closed under union.

ARGUMENT. Gĩkũyũ is a language with Noun Demonstrative Numeral Adjective order, as seen in (7), while Yoruba is a language with Noun Adjective Numeral Demonstrative order, as seen in (8). In both languages, this order of the postnominal modifiers is the only one possible.[14]

(7) Word order in Gĩkũyũ NP

[Mbarathi icio igiri njiru] niirorire
10horse 10Dem 10two 10black vanished
Those two black horses vanished

(8) Word order in Yoruba NP

[13] This is a special case of the following universal, the essence of which was proposed by Matthew Dryer (p.c.):

> If *X*, *Y*, and *Z* are elements of a constituent in a language, and the language has the orders *XYZ* and *ZYX*, then it will also have some other order of *XYZ* in that constituent.

In this case, *Y* is Numeral, and *X* and *Z* are Demonstrative and Adjective, with constituent being NP.

[14] Gĩkũyũ also allow the demonstrative to precede the noun, and Yoruba allows a plural marker/pronoun to precede the noun. Thanks to John Mugane for the Gĩkũyũ example and discussion, and to Oduntan Bode for the Yoruba example and discussion.

[Ile giga meji yii] dara
house tall two this nice
These two tall buildings are nice

Now the union of Gĩkũyũ and Yoruba will have sentences of the form *N* Dem Num Adj *V* and *N* Adj Num Dem *V*, but none of the other 4 orders of Dem, Num and Adj following the *N*. This violates NL Universal 4, so the union of Gĩkũyũ and Yoruba is not a natural language. √

Another non-closure property is intersection. The relevant universal is given below:

NL UNIVERSAL 5. If a natural language has an imperative construction, it will allow an argument to accompany an imperative verb in a simple sentence (e.g. Eat your vegetables!)

PROPOSITION 7. Natural languages are not closed under intersection.

ARGUMENT. DS is a strict verb-final language, so all arguments of the verb precede it:

Miñ bara	* Bara miñ
1sg-ac help	help 1sg-acc
Help me!	(Help me!)

Of course, in English, non-subject arguments follow the verb. Thus, **L** = English ∩ DS will not contain any imperatives with an accompanying argument, violating NL Universal 5. √

Of all the basic closure properties, closure under positive Kleene closure (Kleene^{+}) is the most difficult to argue against.[15] There are two reasons for this. First, since we are dealing only with strings of categories, and not constituent structure, two concatenated sentences can often be reanalysed as a type of subordination. For example, consider the sequence *N V N V*, which corresponds to "Pat left" concatenated with "Lee arrived", i.e. "Pat left Lee arrived" which is not a sentence.[16] However it also corresponds to "Pat said Lee arrived", which is a perfectly well-formed sentence. It is the lack of constituent information which allows this reanalysis.

[15] Natural languages are trivially not closed under Kleene*, since for any language *L*, *L** contains the empty string, and all sentences have at least one category, by hypothesis.
[16] But see below.

The second reason that closure under positive Kleene closure is difficult to argue against is that a wide variety of languages do allow sentences to be concatenated to form other sentences.[17] There are two types. The first type corresponds to coordination with "and" in English. Even English allows this to a certain extent, in "I'll check this room; you check that one". In these cases, the whole sequence can be given intonation that corresponds to a single sentence (roughly, there is no fall on "room", but only on "one"). The second type of concatenation is in lists, as in "I came, I saw, I conquered". Here, each element of the list descends in pitch from the preceding one.

If natural languages generally allow any type of sentence to occur in either of the concatenation constructions, then each language will be its own positive Kleene closure,[18] and natural language will obviously be closed under positive Kleene closure.

However, it is not clear that concatenation of arbitrary sentences is possible. In particular, questions and possibly imperatives seem to resist concatenation. So something like "Where are you? Are you coming?" seems to have only one possible intonation, where the two clauses are independent sentences. Similarly, "Who are you? What are you doing? Where are you going?" does not seem to have a list intonation, but only one with three sentences. If questions cannot be concatenated (formulated as NL Universal 6), then we can construct an argument that natural language is not closed under positive Kleene closure.[19]

NL UNIVERSAL 6 (tentative). Questions cannot be concatenated to form a sentence.

PROPOSITION 8. Natural languages are not closed under Kleene closure.

ARGUMENT. Fula has two types of matrix question markers.[20] The sentence final marker *na* can be used to indicate a simple yes-no question,

[17] I am grateful to a member of the LINGUIST list for information on this point, and particularly to David M. W. Powers for pointing out lists, as discussed below.

[18] We can note that for natural language to be closed under Kleene closure, at least one language must be its own Kleene closure: Given a natural language L, L^+ must be a natural language, but $(L^+)^+ = L^+$.

[19] Of course, this would also provide another argument that natural language is not closed under concatenation.

[20] These markers are not used with embedded questions.

as is shown in (9a). There is also a sentence initial marker *kora* which has the force of expecting a positive reply, as shown in (9b).

(9) Question markers in Fula

a. Sentence final

A nani na?
2sg hear-pst QF
Did you hear?

b. Sentence initial

Kora a nani?
QI 2sg hear-pst
I trust that you heard?

Let L be Fula. Then the string Pro V QF QI Pro V is in **LL**. However, I argue that this string cannot be a well formed sentence in any language.

By NL Universal 6, the string cannot be analysed as the concatenation of two questions, so there would have to be another analysis of it. In particular, it would have to be analysed as a type of subordination. The two possibilities are given in (10).

(10) Possible reanalyses of Pro V QF QI Pro V

a. [[Pro V QF] QI Pro V]
b. [Pro V QF[QI Pro V]]

Taking (10a) first, there is nothing to rule out [Pro V QF] as a subordinate clause, since embedded questions often are marked in the same way as matrix questions. But then QI is not initial in its clause. A similar argument can be made concerning (10b).[21]

In this section, I have argued that natural languages are not closed under any of the standard set operations. In doing so, I have argued for the following theorem.

ANTI-AFL THEOREM. The family of natural languages is an anti-AFL.

[21] It might be argued that the two question markers are not members of distinct lexical categories. If this is so, then the relevant string is Pro V Q Q Pro V, we have the same two analyses, and the same problem as in (10), since *Q* has to occur at a sentence edge. Sentence internal question particles have different cross-linguistic properties (Greenberg 1963), so presumably are a different category.

ARGUMENT. This follows from Propositions 1, 2, 3, 6, 8 and the definition of anti-AFL. √

It is worth noting that arguments for the non-closure results for most of the operations involving only one natural language (finite intersection with regular sets, homomorphism, inverse homomorphism, complementation)[22] are, or can be extended in obvious ways, to stronger propositions which show that the result of applying the operation to *any* natural language results in something which is not a natural language.[23] Proposition 4, involving complementation, is an example.

The operations involving more than one natural language (concatenation, union, and intersection) may well not hold of particular pairs of natural languages. In other words, it may be possible to find pairs of natural languages such that their concatenation (respectively, union, intersection) is itself a natural language.

Consider the hypothetical example of a language which adds a lexical category, for example an optional question marker, but makes no other simultaneous changes. Call the older language **L**, and the newer language **L**′. Clearly $\mathbf{L} \subset \mathbf{L}'$. But then $\mathbf{L} \cup \mathbf{L}' = \mathbf{L}'$ and $\mathbf{L} \cap \mathbf{L}' = \mathbf{L}$ are both natural languages. Bambara may be in the process of providing exactly this example. While it already has a sentence final question marker *wa*, some people have apparently added a sentence initial question marker *esiki* (< French *est-ce que*).

4. Consequences of the Non-Closure Results

The non-closure properties and their proofs raise a host of issues. These issues include the relevancy of the mathematical properties, the nature of the proofs, the nature of the appropriate object of study, and ways in which linguistic theories can/should capture these generalizations. These issues will be discussed in turn.

4.1. *The Mathematical Properties and the Proofs*

From a purely formal point of view, the relevancy of the mathematical properties is clear. When natural languages are taken to be sets of certain types of elements, then they are mathematical objects and as such, the

[22] The case where the intersection is infinite may be true as well, but that remains an open question. The case of Kleene closure is another one that may be true for all languages.

[23] I would like to thank Johan van Benthem for discussion on this point.

mathematical operations relevant to sets and to the elements are indeed relevant. In addition, these formal results can lead to further formal questions. One obvious question is whether there are operations under which natural language *is* closed. Mirror image is one candidate.[24] Another type of question is to look for subregularities in language, e.g. a characterization of the infinite regular sets that can be subsets of a natural language.[25,26]

We can also note that Chomsky 1981 (p. 11) has suggested it is possible that there are only a finite number of natural languages.[27] He also mentions that the finiteness of the number of natural languages makes some (but not necessarily all) mathematical questions (e.g. learnability) uninteresting. In the appendix it is shown that most of the non-closure properties do not follow simply from having a finite family of languages. In other words, for most operations, there is a finite family of languages which is closed under that operation, and there is a finite family of languages which is not. Thus, the non-closure properties are an example of properties that are not rendered trivial by restricting our attention only to finite families of languages.

The formal results are relevant in another way: they can inform other aspects of linguistic inquiry. To take an example suggested by an anonymous referee, suppose it turns out that the concatenation of English and French is in fact a (possible) natural language. Since we know that natural language is not closed under concatenation, this is a surprising result. We should then investigate the properties of English and French that lead to their concatenation being a natural language. Without the non-closure results, we would have no reason to be surprised at the resulting concatenation, and no reason to investigate further why it is legitimate. Thus, these non-closure results could lead to new areas of investigation.

A second way in which the non-closure results can inform other aspects of linguistic inquiry is illustrated by the Bambara example discussed earlier. Recall that Bambara seems to be in the process of acquiring a sentence initial question marker. If no other changes occur, then the older stage of Bambara (B_1) is a subset of the newer one (B_2). But if that's

[24] Note that I am not suggesting that any particular language is its own mirror image, but that the mirror image of a natural language might be another (possible) natural language.

[25] This was suggested by Johan van Benthem.

[26] Another formal property of natural language that has been previously explored is the Constant Growth property (Joshi 1985). It says (roughly) that there are not arbitrarily large gaps in the lengths of sentences in a language. A related, stronger property proposed by Geoff Pullum in a series of talks is the String-Length Density property, namely that every language has sentences of every (finite) length.

[27] But see Pullum (1983) for commentary.

true, then $B_1 \cup B_2 = B_2$ and $B_1 \cap B_2 = B_1$. Since natural language is not closed under union or intersection, these results are exceptional. Various questions arise, such as what kinds of language change result in a subset relation between the different stages and which subset relations (and hence changes) are possible in natural language.

Turning to the proofs themselves, we see that they have an interesting property. The proofs themselves are all very simple, but the properties of natural language that they rely on are not at all trivial. Most striking in this regard are the word order universal (NL Universal 4) and the imperative-argument universal (NL Universal 5). We also saw that the question of closure under Kleene closure hinges on the status of lists, an area which has not received much linguistic attention. This technique of appealing to natural language universals to prove formal properties about natural language is, to the best of my knowledge, novel. This technique can also be applied to other situations, as will be shown below.

It is perhaps worth pointing out again that these proofs are unusual in that they treat natural languages as strings of lexical categories, not strings of words. As discussed above, this uncommon view is crucial to making the closure properties (more) interesting. In fact, to the best of my knowledge, proofs of the closure properties of these sets have never been given, since people have never considered these sets, and have been concerned with sentences as strings of words.[28]

4.2. *The Object of Study*

The non-closure results also raise the issue of what the appropriate object of study is. The norm in the mathematical linguistics literature has been to treat natural languages as sets of strings (usually of words). However, a case can be made for treating natural language as sets of structures (e.g. trees). Certainly, structures of one sort or another are the primary objects of syntactic theory.

What we can notice is that the arguments made in the proofs can be carried over to structures with little, if any modication, when the operations are well-defined. Consider for example non-closure under homomorphism. The idea of the proof was to map every category to a single (non-sentential) category. We can do the same thing with (node) labels in structures, and construct the same kind of homomorphism. Thus, the proofs offer new techniques that can be applied to other problems, in

[28] Eric Ristad (p.c.) confirmed that he had in mind sentences as strings of words when he made his statement of non-closure properties, as mentioned in fn. 2.

addition to the non-closure results about natural languages as sets of strings.

An additional issue that is raised by this work is that of grammaticality. It is standardly assumed in the mathematical linguistics literature that grammaticality is binary – a string is either grammatical or it is not. This paper follows in that tradition.

However, this assumption about grammaticality has been increasingly questioned (cf. Chomsky 1986, Manaster Ramer 1993). If grammaticality is a gradient feature, then the question is which strings form the appropriate object of study. The simplest distinction to make is the one made here, namely fully grammatical sentences versus everything else. If we loosen this restriction, we run the risk of including any string, and making the question of what is a possible natural language moot. If a principled reason can be given for a different distinction, then the proofs here can be reexamined to see if they are affected by the change in definition of a language.[29]

4.3. *Capturing the Generalizations*

Before turning to how linguistic theories can/should capture the non-closure generalizations, it is worth pointing out that natural language is not unique in having these non-closure properties. There are many types of developmental system (L system) families of languages which have the same non-closure properties.[30] Aravind Josh (p.c.) notes that the family of programming languages also seems to have the same non-closure properties.

What these families of languages have in common is their functionality: natural languages are (primarily) for expressing ideas; L-systems are (primarily) for expressing biological growth; and programming languages are (primarily) for expressing algorithms. However, there are also other non-functional families of languages which also have the non-closure properties. For example, let $L_n = a^n b^+$, and let F be the family of all such languages. It is simple to show that F shares all the non-closure properties that natural language has. It still could be the case, though, that the functionality of the three previous families of languages is somehow "responsible" for their having the non-closure properties. Why this should be is puzzling.

Returning now to the question of the consequences of the non-closure

[29] For a different approach to this problem, see Rounds et al. (1987).
[30] Cf. Rozenberg (1974).

properties for linguistic formalisms, the fundamental question that must be addressed is: What is the formalism supposed to do? There are two main answers to this question, and the consequences of non-closure will vary according to the answer.

One role that a linguistic formalism could have is allowing all and only the possible natural languages (and their grammars). This has been a theme of Chomsky's work from early on, as the quote in (11) illustrates. Other frameworks have been developed in a similar spirit, as evidenced by the quotation in (12).

(11) (Chomsky 1965: 31)
. . . we must require of such a [explanatorily adequate] linguistic theory that it provide for:
(i) an enumeration of the class $s_1, s_2, \ldots$ of possible sentences
(ii) an enumeration of the class $SD_1, SD_2, \ldots$ of possible structural descriptions
(iii) an enumeration of the class $G_1, G_2, \ldots$ of possible generative grammars

(12) (Gazdar et al. 1985: 4)
Our goal in the work that has led to GPSG has been to arrive at a constrained metalanguage capable of defining the grammars of natural languages, but not the grammar of, say, the set of prime numbers.

On this view of the role of formalism, the consequences of the non-closure properties are quite clear: a syntactic formalism needs to have properties such that the family of languages that it generates has the non-closure properties. There are two different ways in which a formalism could have such properties. One way is for the basic mechanisms themselves to guarantee the properties. For example, if the basic mechanism is context-free phrase structure rules, then the family of languages generated could well be within the context-free languages.

The other way for a formalism to have a property leading to non-closure is to have additional constraints in addition to the basic mechanisms. "Finite closure" of metarules in GPSG is an example of such a constraint. The basic mechanisms of ID/LP (phrase structure) rules, and metarules (generating ID rules from other ID rules) do not in and of themselves guarantee that the family of languages is within the context-free languages (cf. Uszkoreit and Peters 1986). However, limiting the application of metarules to forming a finite set of ID rules does guarantee the context-freeness of the languages generated (Gazdar et al. 1985: 66).

A second view of the role of formalism is hinted at in the quote from Gazdar et al. above. On this view, the formalism is a "metalanguage" used to describe grammars, and it does not itself necessarily guarantee any particular formal properties of natural language. The formal properties of natural language follow from linguistically motivated conditions (in contrast to the purely formal condition of finite closure of metarules) on the types of grammars expressible in the formalism.[31]

On this view of formalism, the consequences of the non-closure properties are also clear: linguistically motivated properties of grammars need to be found which then entail the non-closure properties. These properties might be the ones presented here as the NL Universals, or they could be other properties.[32]

What should be clear is that on either view of the role of formalism, there is work to be done in finding the appropriate constraints and/or mechanisms to account for the non-closure properties.

Appendix

For each operation below, I have given a finite family that is closed under the operation, and a finite family that is not closed ("open") under the operation. In cases where one type does not exist, I have given a demonstration of that fact. These results show that in general, just knowing that a family is finite does not guarantee any kind of closure, or non-closure, properties.

Union
Closed Family: $\{L\}$: $L \cup L = L$
Open Family: $\{a^*, b^*\}$: $a^* \cup b^* = a^* + b^*$

Intersection
Closed Family: $\{L\}$: $L \cap L = L$
Open Family: $\{a^*b, ba^*\}$: $a^*b \cap ba^* = b$

Complementation
Closed Family: $\{a^{2n}, a^{2n+1}\}$: $a^* - a^{2n} = a^{2n+1}$, $a^* - a^{2n+1} = a^{2n}$

[31] As Alexis Manaster Ramer has pointed out (p.c.) the difference between these two views may not be substantive, since imposing constraints on a formalism results in another formalism, no matter what the motivation of the constraints. However, this dicotomy is a commonly perceived one.

[32] This discussion harks back to Wasow (1978), who argues that we need to constrain the languages generated by linguistic formalisms and not just the linguistic formalism.

Open Family: $\{a^*b^*\}$: $(a + b)^* - a^*b^* \neq a^*b^*$

Concatenation
Closed Family: $\{a^*\}$: $a^*a^* = a^*$
Open Family: $\{a^*b^*\}$: $a^*b^*a^*b^* \neq a^*b^*$

Kleene closure
Closed Family: $\{a^*\}$: $(a^*)^* = a^*$

Note: $\{a^*, b^*\}$ is also closed, but not closed under concatenation

Open Family: $\{a^*b^*\}$: $(a^*b^*)^* \neq a^*b^*$

Homomorphism, inverse homomorphism
Closed Family: N.A.
Open Family:
If each language has a finite vocabulary.
Let $\Sigma = \cup\Sigma_i$ (the vocabularies)
Given a in Σ, a' not in Σ.
Let $h(a) = a'$, $g(a') = a$
F is not closed under h, g^{-1}.

Intersection with a regular set
Closed Family: $\{\emptyset, \{a\}\}$: $L \cap R = \emptyset$ or $\{a\}$ (This is possible only if $\emptyset$ is in F.)
Open Family: $\{a^*\}$: if $R = a^{2n}$, then $L \cap R = R$

References

Barton, G. Edward Jr., Robert C. Berwick, and Eric Sven Ristad: 1987, *Computational Complexity and Natural Language*, MIT Press, Cambridge, MA.

Chomsky, Noam: 1965, *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

Chomsky, Noam: 1981, *Lectures on Government and Binding*, Foris Publishers, Cinnaminson, NJ.

Chomsky, Noam: 1986, *Knowledge of Langauge: Its Nature, Origins, and Use*, Praeger, New York.

Embree, Emily: 1993, 'The Morpheme 'mɔ̃' and the Dogon Genitive Construction', Paper presented at the 24th Annual Conference on African Linguistics.

Gazdar, Gerald, Ewan Klein, Geoffrey, K. Pullum, and Ivan A. Sag: 1985, *Generalized Phrase Structure Grammar*, Harvard University Press, Cambridge, MA.

Greenberg, Joseph: 1963, 'Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements', *Universals of Language*, MIT Press, Cambridge, MA, pp. 73–113.

Hopcroft, John E. and Jeffrey D. Ullman: 1979, *Introduction to Automata Theory, Languages, and Computation*, Addison and Wesley Publishing Company, Reading, MA.

Joshi, Aravind K.: 1985, 'Tree Adjoining Grammars: How Much Context-Sensitivity is Required to Provide Reasonable Structural Descriptions?', in David R. Dowty, Lauri Kartunnen, and Arnold M. Zwicky (eds.), *Natural Langauge Parsing: Psychological, Computational and Theoretical Perspectives*, Cambridge University Press, Cambridge, pp. 206–250.

Kervran, Marcel: 1982, *Dictionnaire Dogon: Donno So, Région de Bandiagara*, Paroisse Catholique, Bandiagara, Mali.

Kornai, András and Geoffrey K. Pullum: 1990, 'The X-Bar Theory of Phrase Structure', *Language* **66**(1), 24–50.

Manaster-Ramer, Alexis: 1993, 'Towards Transductive Linguistics', in Karen Jensen, George E. Heidorn, and Stephen D. Richardson (eds.), *Natural Language Processing: the PLNLP Approach*, Kluwer Academic Publishers, Boston, pp. 13–27.

Pullum, Geoffrey K.: 1983, 'How Many Possible Human Languages are There?', *Linguistic Inquiry* **14**(3), 447–467.

Ristad, Eric Sven: 1986, 'Defining Natural Language Grammars in GPSG', *24th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pp. 40–44.

Rounds, William C., Alexis Manaster-Ramer, and Joyce Friedman: 1957, 'Finding Natural Languages a Home in Formal Language Theory', in Alexis Manaster-Ramer (ed.), *Mathematics of Language*, John Benjamins Publishing Company, Philadelphia, pp. 349–359.

Rozenberg, Grzegorz: 1974, 'Theory of L Systems: From the Point of View of Formal Language Theory', in Grzegorz Rozenberg and Arto Salomaa (eds.), *L Systems*, Springer-Verlag, New York, pp. 1–23.

Salomaa, Arto: 1973, *Formal Languages*, Academic Press, Inc., Boston.

Savitch, Walter J., Emmon Bach, William Marsh, and Gila Safran-Naveh: 1987, *The Formal Complexity of Natural Language*, D. Reidel Publishing Company, Dordrecht.

Uszkoreit, Hans and Stanley Peters: 1986, 'On Some Formal Properties of Metarules', reprinted in Walter J. Savitch et al. (eds.), *Linguistics and Philosophy*, 227–250: **9**(4), 477–494.

Wasow, Thomas: 1978, 'On Constraining the Class of Transformational Languages', *Synthese* **39**, 81–104. Reprinted in Walter J. Savich et al. 1987, 56–86.

The University of Iowa
Department of Linguistics
Iowa City IA 52242
U.S.A.