LOCAL GRAMMARS IN CORPUS CALCULUS

Franz Guenthner
CIS, Universität München

1 Some Basic Tenets of Corpus Calculus

1.1 Knowledge of Language with and without a Corpus

For a very long time most investigations in linguistics were "example based" in the sense that linguists were mainly interested in exhibiting generic observations about linguistic structure; more recently (since the 1950s) linguists have moved beyond the systematic collection of examples as the primary result of analysis towards constructing sets of "rules", i.e. grammars. This has lead many researchers to assume that a "language" can best be viewed as the ("infinite") set of sentences generated or characterized by a corresponding grammar. This seems still to be the most widely accepted view of what linguistic description and explanation should be concerned with. The even more recent trend in computational linguistics that is exemplified by various statistical and machine-learning approaches to language analysis views the primary goal of the algorithms deployed to be the construction of some set of grammar rules. And these rules (whether probabilistic or not) are essentially of the context-free or enhanced context-free variety.

As we shall argue here, this general assumption is not necessarily the best way to approach the problems of language description. For one thing, the many attempts to formulate systematic sets of syntactic (and sometimes also semantic) rules have been more than unsuccessful in providing grammatical descriptions that can cater to any realistic testing or applications; not only does no existing "large-scale" grammar come close to capturing even the most frequent sentential forms of a specific natural language (a brief look at current machine translation systems or any other natural language system involving a "parser" amply illustrates this state of affairs), and even when parsers return analyses it is more often than not questionable whether these analysis are indeed the analyses we should want to have!!

Writers of grammars should at least have been expected to take a closer look at the way the grammars they propose match up with the observable material contained in the abundant collections of language we have now come to call "corpora". There has of course also been a trend to take corpora into account from the beginning in linguistic description and many quite different ways have been suggested in connection with the role of corpora in linguistics and especially in computational linguistic thinking.

Without going into details here, it is very interesting to observe that there even seems to be a kind of confluence in progress between the many corpus linguistic approaches and the diverse recent statistical trends (especially in statistical machine translation); what is proposed under such different nominations like "example-based", "memory-based", "data-oriented", etc. in fact amounts to take corpora as the starting point for a number of ways to construct grammars from existing sentences and sometimes also to evaluate these grammars in turn on portions of the corpora.

But all these approaches have in common that the role of corpora (and the algorithmic techniques used to analyze and exploit them) is essentially one of convenience: writing grammars by hand is assumed to be to time-consuming, too expensive, too difficult in a number of other ways...

1.2 The Notion of a Corpus

Almost by definition a language is learned in the context of previous manifestations of that language; every learner and every speaker of a language has access in a large variety of ways to previous utterances of that language. In fact, one way to say that someone speaks a particular language is simply to say that he has had access to a "sufficient" amount of data from that language. For any language and any speaker there is the (virtual) corpus of utterances which contains all the available material for that language (for the sake of simplicity we assume for the moment that these are given in some written form) and the subset of these utterances that the particular speaker has in fact been exposed to or has himself produced over time. Obviously the latter is only a tiny portion of the former and it is quite different from speaker to speaker. What a speaker can understand (or produce) in that language depends directly on these two corpora. It is hard to imagine that someone could master a given natural language without tremendous exposure to samples of the
language. It is on the basis of these samples that previously not encountered utterances can be understood in the sense that they can be systematically associated with or reduced to utterances in the sample corpus.

1.3 The Role of Elementary Sentences

Any such huge virtual corpus derived from concrete archives will contain a variety of elements; for the sake of the present discussion we assume that we can single out the subcorpus consisting only of "elementary sentences". By

elementary sentence we mean utterances that contain only one predicational part. Of course, this is by no means a trivial separation to accomplish, but it can certainly be approximated in a number of ways. Everything we will cover below can be applied to complex sentences as well, but it is preferable to limit the discussion to elementary sentences at this point. Restricting our discussion to English, let us assume that the initial virtual corpus can be viewed as a corpus consisting only of elementary sentences.

The first important observation concerns the form of these elementary sentences. We assume without argument at this point that sentences in any natural languages are either simple or complex. For simple sentences, we assume further that they can be decomposed into a predicational and an argument part (or better, an argument list).

Elementary sentences in languages like English come in three different guises: their predicational part is either a) verb-based (the predicate is realized by a semantic verb); b) adjective, noun or preposition-based (the predicated is realized by an adjectival, nominal or prepositional form); c) a frozen predicate.

In addition to being able to distinguish between these three varieties of predicational realizations, what is important to observe is that these need to be listed as dictionary entries (we call such entries predicate-argument structures).

Here are some simple examples of the classes above:
- Verb-based: *he kissed her*
- Adjective-based: *he is tired*
- Noun-based: *he gave a talk about it*
- Preposition-based: *he is under it*
- Frozen Predicate: *she blew her stack*

One of our basic assumptions is that any elementary sentences must belong to one of the three classes above and that the corresponding sentential schemas must be enumerated in one the basic lexica of the language (here English) in question.

1.4 Elementary Sentences and Predicate-Argument Structures

Simple inspection of the examples above reveals that the structure of the types elementary sentences goes very much hand in hand with their built-in semantic structure: we view the semantic structure of elementary sentences to consist of a predicate and argument part. This is the second basic axiom in our framework: to every elementary sentence corresponds a predicate-argument structure. Obviously different elementary sentences can be related to the same predicate-argument structure (PAS from now).

1.5 A Grammar is Not Necessarily a "Set of Syntactic Rules"

Assuming our corpus to consist only of elementary sentences, the task of analyzing these sentences can thus be seen to be one of determining for each of these elementary sentences which sentential schema (i.e. which of the three classes) corresponds to it and more interestingly which PAS the elementary sentence exhibits. This kind of analysis (or rather this kind of reduction or simplification) is not what is typically accomplished by grammatical rules of the usual kind.

The main difference is that on our view the analysis of a sentence should come up with the predicate-argument structure associated with the utterance and not with some analysis tree that only matches syntactic categories.

2 Corpus Calculus Principles

One of the central claims behind Corpus Calculus is that the production and comprehension of new sentences consists in operations that show how such sentences are systematically relatable to previously produced sentences. We write

$$K ? s$$

to express that the sentence s can be derived from a corpus K via a number of operations that we will discuss below. In general, showing that a sentence is already implicitly contained in K is a demonstration of its "grammaticality" with respect to K and derivatively with respect to the language of K. Clearly this notion can also be "quantified" in a number of interesting respects which we will not discuss further here.

2.1 The Repetitive Nature of Language

It is obvious that such a notion of linguistic description presupposes that there is a tremendous amount of repetition - at various levels - of linguistics elements which we can capitalize on; one of the up-shots of deploying corpus calculus techniques massively will be a realistic description of the kinds of repetitiveness that are of interest.

2.2 Inferring Sentencehood

Contrary to popular belief it is not completely obvious what the basis for a well-founded concept of a "well-formed" or "grammatical" sentence should be. There have been many different answers to this question but almost none of them have taken the evidence provided by corpora (i.e. by the mass of previous utterances) into account. What should count as a sentence in any given language should obviously be derived from a detailed inspection of what sentences have been produced.

So a first and trivial answer to when a corpus provides evidence for a sentence belonging to a language is Prinicple 0 of Corpus Calculus which states that K ? *s* if s is an element of K; we will need to distinguish pure elementhood from more sophisticated notions of what it means for *s* to occur in K.

2.3 Principle 1: Distribution

Almost no interesting sentences s can be derived from our all-purpose corpus K via Principle 0 (with the exception of course of some so-called non-standard sentences like "ok" or "good night" etc.).

A more interesting principle (the principle of "distribution" or of "substitution" says that *s* follows from K if there is an s' in K such that s is a substitution instance of s' with respect to certain subsequences, e.g. noun phrase arguments. In other words, a sentence like "John despises his teacher" can be derived from K if there is a sentence in K which modulo licensed substitutions is identical to K; e.g. "Max despises the president".

2.4 Principle 2: Permutation

The principle of permutation guarantees that "the students read the paper on statistical methods" can be inferred from a sentence in K of the form for instance "many papers were read by the students"; this principle captures the majority of observations related to various transformational approaches to linguistic description.

2.5 Principle 3: Lexical Functions

Many years ago Igor Mel'čuk put forth an extremely fruitful view on paraphrastic relations in language based on a set of "lexical functions"; it is obvious that many sentences can be related on the basis of the connections encoded in these lexical functions. These insights can be used to show that a sentence like "He lectured on his favorite topics" can be viewed as an instance of "Maxwell gave lectures on quantum mechanics". Paraphrases based on lexical functions are thus central to the operations of Corpus Calculus.

2.6 Principle 4: Grammatical Meanings

Probably most well-known and most entrenched among the linguistic operations (essentially of an analogy kind) are systematic regularities based upon grammatical paradigms or grammatical series as we shall call them here. These series are interesting both with respect to individual predicate-argument structures as well as across different predicate-argument structures. Consider the parallels between "John smokes" and "John didn't smoke" and "John didn't smoke" and "John didn't drink". Many if not most linguistic treatises obfuscate the difference between the two sets of patterns!
4

2.7 Principle 5: Respecting PAS

When we want to show that a sentence *s* is already implicitly contained in a corpus K we need to make sure that we are not trading apples for pears. There is little point to make an observation about an instance of the principles above when we jump from a sentence based on a predicate P to a sentence based on a predicate Q. Any of the above "regularities" (e.g. permutations involving a passive) only make deep sense when applied to the SAME predicate argument structure. Spurious similarities only lead to confusion.

3 What are Local Grammars?

3.1 The Notion of a Local Grammar

Maurice Gross introduced the notion of local grammars about 20 years ago in connection with the low-level description of many grammatical phenomena that escape a systematic description in terms of abstract syntactic rules.

Over the years this notion has become more and more clear and it has been employed in the description of a large number of specific grammatical constructions; interesting examples are for example the grammar of noun phrases, of temporal adverbs, etc.

Local grammars come in two main varieties: grammars for predicate-arguments structures and grammars for other constructions.

We need local grammars for the application of Principle 1 when we need to abstract from (complex) argument strings or for the application of Principle 2 when we need to compare permuted strings.

We also need local grammars for a very large set of so-called inserts, i.e. constructions which can occur at almost every position inside an elementary sentence realizing a predicate argument structure. Similarly, local grammars can easily characterize a huge set of adverbial modifications, e.g. the temporal adverbial phrases expressing the meaning of "when" something occurred.

3.2 Local Grammars for Predicates

Already in 1992 E. Roche pioneered the idea of producing the set of realizations of a predicate in terms of a local grammar; more recently work by S. Paumier has shown how these local grammars for a given predicate can be systematically generated from a standard linguistic description given for instance in the form of a syntactic table in the style of the LADL.

Trivially speaking a 4-place predicate like

X sells Y to Z for U

potentially has 5! permutation realizations (with respect to the predicate and the four argument positions) of which more than half are in fact observed.

These realizations can easily be captured and expressed by a local grammar.

3.3 Local Grammars for the Rest

Similar observations apply to the grammars of noun phrases, the grammar of modifiers (adverbials), inserts and other constructions.

4 The Principles in Action

4.1 A Simple Example

Let us have a closer look at the first sentence in this note:

For a very long time most investigations in linguistics were "example based" in the sense that linguists were mainly interested in exhibiting generic observations about linguistic structure

It has hardly likely that Principle 0 will lead us to accept this sentence as a bona fide instance because it has already occurred n times in K; most interesting sentences are unique as far as elementhood in K is concerned.

But like all sentences, this sentence is not that unique after all. Our sentence is not elementary, since it is composed of two sentences connected by "in the sense that":

- *S1: For a very long time most investigations in linguistics were "example based"*
- *S2: linguists were mainly interested in exhibiting generic observations about linguistic structure*

S1 unfortunately has not occurred in K either as far as we can tell; neither have any of its reductions:

- *most investigations in linguistics were "example based"* and
- *X were "example based"*

(NB: the replacement of "most investigations in linguistics" with the schematic argument variable X is an instance of Principle 1 in a special form.)

On the other hand,

- *most investigations in linguistics* and
- *X are "example based"*

do not occur as such directly in K either, but can be derived with intermediate rules. (Cf. the full version of this paper for details.)

So essentially this is simply a case of a new combinatorial association of already observed building blocks!

4.2 CC and Translation

We can generalize the basic insight of Corpus Calculus to many specific applications.

In the context of machine translation, the basic question boils down to asking whether

$$K_{L1}?\ s_{L2}$$

where K_{L1} is a corpus of utterances restricted to the language L1 and s_{L2} is an utterance in L2.

4.3 CC and Diachronic Linguistics

In diachronic linguistics the basic question boils down to asking whether

$$K_t?s_{t'}$$

where K_t is a corpus of utterances restricted to time periods before or equal to t and t′ is a time period after t.

4.4 CC as a Search Engine

In the context of search engines, the basic question is the same as before

$$K?s$$

where K is a set of sentences divided into documents. But we need to add a further condition to Principle 5 which states that the substitutions licensed by Principle 1 must all be "meaning preserving" and not just "structure-preserving". In other words, "Max hates Bill" licences "Jane hates Sue" from a pure CC point of view (the predicates are the same and the arguments have the right type), but obviously the latter does not follow from the former in a semantic sense.

With the further condition, only sentences involving X and Y in the appropriate positions can be used to derive our sentence, where Max = X and Bill = Y can be shown to be semantically equivalent (e.g. co-referential); other logical relations like instantiation etc. can of course come into play as well here. Cf. the work in the context of Hans Kamp's Discourse Representation Theory for many interesting leads on this matter.

4.5 More Applications

There are many more applications for Corpus Calculus as for instance in language learning that could be discussed and there are of course many specific operations and conditions that need to be spelled out for the above principles to be effective; we refer the interested reader to the full version of this paper.

5 Bibliography

1) Gross, G. & F. Guenthner Manuel d'analyse linguistique, in preparation
2) Gross, M. "On the Failure of Generative Grammar", Language, 1982.
3) Gross, M. Methodes en syntaxe, Herman, 1975
4) Gross, M. (1997). "The construction of local grammars". In: Finite-State Language Processing, E.
5) Roche ; Y. Schab`es (eds.), Cambridge, Mass./London: MIT Press, pp. 329-354.
6) Gross, M., 1999. "Lemmatization of compound tenses in English". Analyse Lexicale et Syntaxique,
7) Fairon, C´edrick (ed.), 71-122.
8) Guenthner, F. "Corpus Calculus in a Nutshell", CIS-Report, 2005.
9) Mel'čuk, I. (1997) "Vers une linguistique Sens-Texte". Le¸con inaugurale. Paris: Collège de France
10) Paumier, S. Recursive Automata for Syntactic Grammars, 2003.