Franz Guenthner*
(Munich)

NOTE ON ELEMENTARY SENTENCES AND A TOOL TO FIND THEM

Abstract. This note tries to provide arguments why the proper description of sentences in a natural language (in particular in computational applications) needs to be able to identify the correct «propositional form» (predicate-argument structure, government pattern, subcategorization frame, ...) of the sentence. An analysis which does not provide this is comparable to a lexical analysis of a sentence which labels certain words as «unknowns». The recognition of the predicate-argument structure that a sentence expresses thus needs to have recourse to a dictionary of such structures (much in the sense of a Mel'čukian DEC).

1. Government Patterns and Elementary Sentences

1.1. Introduction

This note is about an on-going research project that has been heavily influenced by Igor Mel'čuk's thinking about the semantics [1] of natural languages, cf. for instance (Mel'cuk 2010) among many other papers. The project runs under the name of «corpus calculus» and is intended to be an approach to linguistic description that interacts with a large variety of theoretical and practical questions about language understanding in a unified and mutually informative way. The basic idea of the corpus calculus approach is to consider most aspects of language analysis and applications of language analysis as following from a small number of principles that guide the analytical steps. This idea is not really new; contrary to most existing grammatical frameworks (and with very few exceptions like Mel'čuk's), we do not think that the best way to describe a language is in terms of some grammar rule format of the kind that has been popular since the middle fifties in the form of (axiomatic) rewriting rules. Since that time most linguists have come to believe and continue to believe that grammatical rules should be of the form X → W, where X is a syntactic category and W a sequence of categories or lexical elements. What constitutes a syntactic category varies from one school of linguistic thought to an-

[1] The other two main direct influences on this project are Z. Harris and M. Gross with respect to their emphasis on the importance of the role of the notion of «elementary sentences».

CIS — University of Munich. Oettingenstr. 67, Munich
gue@cis.uni-muenchen.de

other, but this does not really affect the basic assumptions. What is in our opinion the problem with the received view of grammatical description is simply this: there has been no clear view of what we should take to be the fundamental «units of linguistic analysis». Traditional (and modern) linguistics has clung to «words» as the minimal descriptive elements. What is even worse, it is typically just «simple» words that are manipulated and not the much larger number of «complex words» that need to be taken into account[2]. On our view it is rather the notion of an «elementary sentence» which should play the fundamental role in linguistic analysis. We assume here that the distinction between «elementary» and «complex» sentence is sufficiently clear; thus we can concentrate in the context of this note on how to describe the former.

The drawback to the traditional answers is that the choice of the «word» (as a lexical unit) as the fundamental item influences the rest of the thinking in a way that is detrimental in at least two different ways:

(A) The harmfulness of «bottom-up» thinking in linguistics: with only one exception, all traditional and modern linguistic frameworks start in one way or another with the idea of «words» being the basic units of linguistic descriptions. A large majority of typical problematic issues in linguistics, e.g. «ambiguity», «vagueness», «scope», and countless others, are in our opinion due to the way we think about the descriptive and explanatory context within which judgments and descriptive annotations are to be made. Individual words by themselves carry almost no interesting information and typically need to be supplemented by ad-hoc indications about the (combinatorial) context of possible occurrences of the word. The drawback of this method is obvious: given the individual word as a unit, the value of any description typically depends on the choice of the contexts; but most of the time these remain completely unanalyzed, so that the status of the descriptions produced is far from reliable. What is even worse, the choice of the «word» as the basic unit leads linguists very often to assume descriptive constraints and relations about the behavior of slightly «ambiguous» words, even when there is little or — as is most often the case — no connection at all between the many uses of the words in different contexts.

As already pointed out, the unreflected dependence on the individual word has also led directly to what must be the greatest methodological (if not theoretical) error of modern (Chomskyan and post-Chomskyan) linguistics when it comes to capturing the behavior of words in terms of so-called «syntactic rules» which supposedly operate on the individual words and their combinations. We will not review

[2] Mel'čuk has championed this point from the beginning with his emphasis on collocational elements in language; from a dictionary point of view, the most significant contributions in this area have come from M. Gross.

the decades of naive uses of phrase structure grammars given as examples of how to go about describing the structure of grammatical sentences. Nor will we here comment on the mathematical foundations of rewrite mechanisms in general. This will have to be done elsewhere. What needs to be pointed out, however, is that very popular recent trends in statistical approaches (e.g. probabilistic phrase structure grammars), instead of questioning the assumption of the usefulness of phrase structure in the first place, simply assume that using larger numbers of these structures derived from (very small and unrepresentative) treebanks might constitute a way out. Of course, no set of phrase structures rules (no matter how large) — which does not include information about the general lexico-semantic structure of the sentences they purport to describe — will ever come close to an adequate description of the nature of the sentences in a natural language.

One last remark about the word-centeredness of linguistic thinking: one would think that after hundreds of years of linguistic discussion of words and their interaction with other words, at least a solid and useful enumeration of all the words in common use in a given language (e.g. German, English, Russian, etc.) would have seen the light of day. Needless to say, this is by no means the case. Probably the first detailed and (relatively complete) morphological dictionary for a natural language is only 50 years or so of age [3]. More recently, work done at the LADL in Paris under the direction of Maurice Gross has shown that the notion of the «simple word» is much too crude an approximation of what constitutes a «lexical item» in a language [4]. This work points out that the large majority of lexical units consists of «multi-word items». Again, this is not the place to discuss the question as to why there is so far no dictionary in sight (for any language) of such compound lexical items, or as to what would need to be done to construct such a dictionary [5]. So when such rules (no matter how great their number) are used, the values of the syntactic category variables are never known and are therefore greatly prone to error. In the case of probabilistic grammars this means that it is never possible to assign any kind of serious «reliability score» to an analysis (apart from an incidental corpus-derived one...).

[3] We are referring to Zalizniak's morphological dictionaries of Russian, which seem to be the grandmothers of all electronically useful Russian dictionaries ever since their initial formulation.

[4] M. Gross was, as far as we are aware, the first person to underscore the importance of (morphological) dictionaries of simple and compound lexical items (and in general of frozen expressions in all sysntactic categories).

[5] It is not hard to see that without such a dictionary any serious large-scale application in the area of natural language processing (from machine translation to question-answering, for example) will be hopelessly awash in erroneous results (a brief glance at any translation system — statistical or rule-based or whatever — will make this more than clear. Cf. for extensive discussion Guenthner [in preparation])

(B) The observational limits of «bottom-up» thinking: if the individual word is the main candidate for the starting point of linguistic descriptions, then it is hard to know when such a candidate has been adequately described. Not only because the individual word might have quite a number of uses that are difficult to separate (or are completely independent) but also because it is usually not clear what the limits of an explanation should be. It is for this reason for instance that (again apart from Mel'čuk's and M. Gross's work) so-called idiomatic constructions have had an extremely hard if not impossible role in most grammatical frameworks. If one specifies the behavior of individual words in terms of some syntactic rules — which mention only their syntactic categories — instead of some more fundamental properties associated with the lexical items themselves — then one can hardly be sure what interpretation of the «word» (or better: «use of the word») is indeed intended in a specific utterance.

The main point I want to draw attention to, therefore, is that in linguistic analysis we need to be able — whenever a linguistic operation is performed — to say that the result can be identified as belonging to a previously known class. In order to be able to do this, we need to require not only that all elements involved in such an operation are previously known (e.g. are enumerated in some dictionary in the system) but also that the result of the operation is also in some previously known class. A much discussed linguistic example is a typical case of this: Chomsky (and many others, e.g. Tesnière) argued many years ago (against the statistical approach to language analysis) that a sentence like

(a) colorless green ideas sleep furiously

was indeed a «grammatical» sentence of English, in spite of the fact that (until he coined the sentence) it had never been uttered before and therefore had no statistical weight whatsoever... It is easy to see what led Chomsky to this observation: the above sentence could be taken to be an instantiation of a sequence of syntactic categories like

(b) A A N V ADV

where *colorless*, *green* are assumed to be adjectives, *ideas* a noun, *sleep* a verb and *furiously* an adverb, and where something like these «syntactic» rules

(c) S ⇒ NP VP
NP ⇒ A A N
VP ⇒ V ADV

would account for the overall grammaticality of this sentence... This grammaticality judgment embodies a number of completely spurious dogmas, all of which have to do with the belief that individual words are sufficient for morphological, syntactic and semantic predictions and that abstract rules like those in (c) can be meaningful in terms of the categories themselves. In our view, none of these dogmas are sound at all, nor is the conclusion (that (a) is indeed a «grammatical sentence of English») justified (no matter how it is arrived at): *colorless green ideas sleep furiously* is as much a grammatical sentence as a scarecrow in a field dressed up in human clothing is a human being. It is not grammatical because the main ingredient for sentencehood is absent: **the sentence does not exhibit a predicate argument structure**, a topic to which we now turn.

1.2. Predicate Argument Structures

It seems at first glance relatively easy to define the related notions of «predicate argument structure» (PAS) and «elementary sentence». We saw above how Mel'čuk specified the form and role of his so-called «government patterns», a notion which is both theoretically and notationally very close to what we refer to as PAS here. One of the intuitions underlying our approach to this topic is that there should be a very close (and observable) association between attested sentences and the predicate argument structure they express. The PAS are the basic semantic entities that we can associate with real sentences and in particular with «simple» or «elementary» sentences; so the basic question therefore is how we can tell which PAS is expressed by which sentence, the first axiom being that every elementary sentence necessarily expresses a PAS. A PAS consists in general of two parts: a «predicational part» and an argument part; the argument part is an ordered sequence of semantically-typed argument variables. Which system of semantic categories is most appropriate to specify the arguments is still an open research question [6].

Elementary sentences come in three varieties:

i) Elementary sentences that express the predicate of the PAS by a (simple or compound) verb; this form is typically the one linguists have prioritized, even though, as we shall see, in terms of the numbers of items, this is in fact the smallest class. The PAS in these cases consists of the predicate part (a verb and perhaps some prepositional element) as in

> X KISS Y, X SELL Y to Z for U (a 4-place argument PAS where the fourth argument is the price involved), X TALK to Y about Z

Of course, the arguments X, Y, etc. can be arbitrarily complex, but they reduce to simple (typed) argument placeholders. There is no PAS in English of the form

[6] Some progress towards answering this question has been made in Le Pesant and Mathieu-Colas (ed.) (1998) and Langer (1996).

X SLEEP (where X has a semantic category appropriate to «idea»)[7]

ii) Elementary sentences that express the predicate of the PAS by a single lexical unit that is not a verb, e.g. a noun, an adjective, a preposition or an adverb.

iii) Elementary sentences that express the predicate of the PAS by the entire elementary sentence schema. These are often called «frozen predicates» and, contrary to popular belief, they constitute the largest set among the three types. (Examples would be constructions like «X give Y a hard time», «X have the time of his life»; we estimate that languages like English or German have up to 50 000 such frozen predicates...

The fact that no more or less complete enumerations (for any language) of the ways to expression predications have been produced is thus in our opinion the main obstacle to real progress in just about every area of natural language analysis, and in particular, in natural language processing.

2. Corpus Calculus as an Alternative to Traditional Grammar

The reason for singling out the notion of «predicate argument structures» as the most central one for linguistic analysis is not only that it is the backbone of all semantic analysis, but equally importantly, that it leads to an approach to thinking about language which brings together many different objectives and methods in a new framework, one which we have called «corpus calculus» (Guenthner 2005a; 2005b). Corpus calculus is based on the simple idea that language comprehension, language learning and many other processes are best understood as activities either having to do with relating new sentences to previously learned ones (and the representations associated with them) or with adding new constructions (and the representations associated with them) to the already acquired ones in a systematic manner. These operations all concern the way new utterances can be accommodated in terms of other related ones. Let us consider the case of how the principles of corpus calculus apply to the situation of dealing with new (previously unheard) sentences and let us also just consider the case of «elementary» utterances. We write K_L ? S to express that a sentence in the language L can be decomposed (via the type of operations described below) to elements of K; by K we understand a corpus of sentences in L enhanced with whatever intermediate results used in previous decompositions, where S is an arbitrary elementary sentence of L. (We leave out reference to L in what follows.) We distinguish several different types of operations that

[7] I am not at all saying that metaphorical (or otherwise analogically produced) sentences do not exist or cannot be interpreted in some way; but this is only possible because we know the the literal interpretation of the sentence is nonsense.

can be used to show that K ? S. (A much more detailed discussion of the principles and examples of their application is contained in Guenthner (2005a), which is available from the author.)

Principle 1: K ? S iff $S \in K$; this principle simply says that previously processed sentences count as already understood; this principle also goes a long way to explaining the use and processing of many formulaic and other highly frozen expressions.

Principle 2: K ? S iff $\exists S' \in K$ such that $S \approx_s S'$ and $S' \in K$; this principle (first discussed by Harris in the 50s) places heavy emphasis on the role of «distributional» structure. Essentially, this principle says that new sentences can be formed (and understood) by «substituting» ($\approx_s$) appropriate constituents for constituents of previously processed sentences. In many ways this principle accounts for the only real creativity or «newness» effects available to us as speakers.

Principle 3: K ? S iff $\exists S' \in K$ such that $S \approx_p S'$ and $S' \in K$; this principle, which says that another important way in which sentences can be systematically related is by «permuting» ($\approx_p$) constituents. Again, Harris' notion of a transformation is the first attempt to describe such transformational phenomena. Needless to say there are hundreds of already observed particular forms that such transformations can take (cf. for instance (Blanco 2010) for some discussion.).

Principle 4: K ? S iff $\exists S' \in K$ such that $S \approx_{gf} S'$ and $S' \in K$; this principle corresponds to what has probably been most discussed in the linguistic description tradition and also in methods of (second-language) learning: namely, all the ways we can make «paradigmatic» variations of the elements of language. The subscript *gf* stands for «grammatical function», i.e. all systematic morpho-grammatical operations available in a particular language L. However — as in the case of transformations — no systematic inventory of such operations for an individual language seems to have been made (but cf. Mel'čuk's course on morphology for the best current overview). Of course, calling these operations «grammatical functions» should not be understood to mean that they do not play a semantic role, the contrary is of course true. A particularly impressive illustration of what we mean here is given in Gross (1999), a paper which characterizes (by finite state means) the literally thousands of forms that grammatically complex verbs can take in English; the approach in Gross' paper shows how to enumerate essentially the complete range of auxiliary and quasi-auxiliary forms that can precede a verbal form (whether it expresses a predicate or not) in the elementary sentence, e.g from forms like «might have been *Ving*» to «should have made a plan to V» and more than a thousand similar constructions. The fact that these can indeed be enumerated and used in the application of this principle in concrete cases (in particular, in automated analyses) suggests that there are interesting ways to immensely increase both the precision and speed of such analyses.

Principle 5: K ? S iff $\exists$S' $\in$K such that S $\approx_{\mathrm{lf}}$ S' and S' $\in$K; this principle is due in its entirety to Mel'čuk's fundamental insight into the ubiquity and complexity of the use of so-called «lexical functions» in the grammatical description of natural languages. As in the case of grammatical functions, the range and variety of the use of lexical functions is highly restricted and cannot in principle be guessed or inferred independently from the lexicon of the language in question. Among the many merits of Mel'čuk's analysis of the more than 60 systematic (syntagmatic and paradigmatic) meaning relationships we should underline his very early recognition of the role of «support» verbs in the expression of predication; that verbs like «pay» or «give» in sentences like «X paid a lot of attention to this problem» or «X gave me some advice» are not the real predicates in these sentences has still not been widely acknowledged. A brief glimpse at the entries in such databases as FrameNet, Verbnet, Propbank, the Collins verb pattern classification and many others reveals many entries where this distinction has not been made and where taking these uses of «pay», «give» and about 500 other cases leads to erroneous lexicon entries. In the context of corpus calculus, lexical functions give rise to a new class of inference relations based on the notion of paraphrase, without doubt the most innovative use of lexical functions proposed by Mel'čuk (cf. (Zangenfeind 2010) for a good summary.)

Principle 6: In showing that K ? S via operations of the types mentioned above, we need to keep the basic predicate of S invariant. That is, all reasoning used to showing that a given elementary sentence **may only involve related sentences (in K) that either share the same predicate as S** or sentences S' that are reachable from the predicate in S by paraphrases. Strangely enough — apart from Mel'cuk's explanatory and combinatorial dictionary system — there have been few attempts to address the question of the behavior of individual predicate argument structures in all the necessary detail[8], (cf. also Schuster 2010).

We can summarize the main thrust of what the corpus calculus approach tries to address in an extremely simple observation: any serious approach to language analysis must have an account of all the entities that are part of the analysis; this is obviously true for individual words (and compounds, etc.) but much more so for the **primary units of understanding**, namely, predicate argument structures. Any analysis of a sentence purely in terms of syntactic category labels and some form of structural bracketing is useless if it does not include an identification of the lexical items and above all of the predicate argument structure it expresses. Just imagine a

[8] This is all the more ironic in that formal semantic approaches to natural language semantics have been very popular, but where it should have been obvious that the study of mathematical structures typically deals with the very detailed study of just a few predicates and their properties, e.g. $\in$, $\leq$, etc.

situation where all communication in a given situation consisted of sentences of the form of such so-called grammatical sentences like our «colorless friends». But currently there is no framework of natural language analysis (computational or otherwise) which contains the information necessary to carry out such analyses on a large scale. Mel'čuk's efforts are obviously completely in the right direction, but the description of little more than 300 semantically interesting items in the DEC (for French) is certainly less than what will eventually be necessary.

3. Corpus Construction and Extracting Patterns from Corpora

There have been dozens of attempts over the last few decades to assemble representative collections of «verb patterns» (Cf. Filatova 2011 for a survey). Most of them have been based on manual processing (e.g. using manually constructed treebanks or simply dictionaries), others have tried to automatically extract these patterns from corpora.

In order to investigate the structure and variety of propositional forms in detail, we have designed a corpus database as well as a corpus-indexing and query system of a new kind. These systems are meant to be used in an experimental environment for deploying and testing the corpus calculus framework. The database is to contain as many instances of (partially) analyzed elementary sentences as possible. We started out with the following procedure. From a very large raw corpus of English text that we have accumulated over more than 10 years (at present over 80 billion running words and counting), we continuously extract sentences of length from 2 to 10 words, having the property that we are able to analyze a sentence-initial noun phrase[9] followed by a complex verb structure (exhibiting one of the thousands of auxiliary and quasi-auxiliary elements and additional arguments). The latter we normalize to the base form of the verb in question. This first extraction process resulted in more than 250 million (partially analyzed) sentences of length of 2 to 10 words. For most of our experiments we use a subset of these sentences containing around 100 million sentences of length 5 to 8 words, all having the form X <verb> W, where X represents the sentence-initial noun phrase, <verb> represents a specific verb, and W a sequence of length from 3 to 6 words. The working assumption is that most elementary sentences of English (apart from a small number of longer frozen predicational forms) should appear somewhere in sentences of this length. In a second step we applied the publicly available tree tagger to these sentences. Altogether, this corpus, which we call the PRO-Corpus (referring to the fact that

[9] For the parsing of the noun phrases we use a local grammar of simple English noun phrases (not including relative clauses, but capturing more than 3000 complex determiners).

the initial noun phrase has been «pronominalized») contains about fifteen thousand verb lemmas in the initial verb position [10].

Our goal is to introduce into the PRO-Corpus as many instances of elementary sentences as possible [11] and to continue to «normalize» the W part, ultimately arriving at completely reduced elementary sentence schemas expressing a predicate argument structure [12].

For rapid querying, hypothesis-testing and corpus calculus operations, we designed a new index structure based on suffix array constructions with which we can create parallel suffix arrays to index not only the strings in the corpus but also various levels of morpho-syntactic, semantic information and other types as well in parallel. Queries using this index structure can therefore involve different kinds of constraints at the same time (cf. Goller (2010) for details on the index structure). Here is an example query

SB X #talk# <IN> <DT> <NN>SE

which will find all (more than 3000) sentence schemas in the corpus of length 5 with *talk* as the main verb followed by a determiner and a (singular) noun [13]. The query system allows us to extract constrained hit lists in a large variety of ways. For instance, we can ask for the (frequency) list of all prepositions (referred to by the tag <IN>) with a query like

SB X #talk# [<IN>] <DT> <NN> SE

where [ ] specifies the context for which we want the list of distributions. It takes a few milli-seconds to extract the following result from the entire index of over 100 million sentences:

[10] In reality of course more verbs, since there are quite a lot of particle verbs which have not been oincluded in this count.

[11] The way this corpus is being automatically constructed has a number of drawbacks (e.g. it presently includes many sentences with non-argumental *it*-forms, it does not include any sentences with propositional subjects, etc.) We plan to remedy this in the near future.

[12] A crucial step in this procedure is the identification of non-argument noun phrases that are typically part of adverbial phrases, which need to be enumerated. These include dates, locatives and many other kinds of modifiers and connectives. We have started to compile a static list of such adverbials (*from time to time*, *in the first place, over the years*, and tens of thousands of others.

[13] SB and SE are markers that indicate sentence beginnings and endings, respectively. The frequency indications in the tables that follow refer to the PRO-Corpus and they obviously do not correpond to the raw frequency counts one is used to usual statistics on (non-pretreated) corpora. More about this in Guenthner (2005b).

1525 about
610 to
289 with
126 like
124 of
113 in
82 on
42 through
38 at
32 over
31 for
29 into
25 during
18 around
16 as
16 up
16 after
11 throughout
8 before
8 from
6 without
5 by
5 outside
5 since
4 off
4 above
4 out
4 across
4 against
2 via
2 thru
2 along
2 until
1 till
1 upon
1 under
1 towards
1 toward
1 near
1 because
1 behind
1 beside
1 beyond

We can also use the query syntax to extract cooccurrence results directly, as for instance in

SB X #talk# [<IN>] <DT> [<NN>]SE

where the result consists of all the pairs of prepositions and nouns, thus giving us interesting clusters like (we list just a few of them):

with [..]woman
with [..]man
with [..]guy
with [..]teacher
with [..]player
with [..]girl
with [..]doctor
with [..]team
with [..]supervisor
with [..]reporter
with [..]person
with [..]patient
with [..]pastor
with [..]officer
with [..]minister
with [..]lawyer
with [..]instructor
with [..]ghost
with [..]gentleman
with [..]friend
with [..]editor
with [..]dude
with [..]day
with [..]counselor
with [..]consultant
with [..]clerk
with [..]church
with [..]caller
with [..]boy
with [..]audience
with [..]attorney
with [..]accent
with [..]wolf
with [..]witness
with [..]winner
with [..]wind
with [..]wheeze
with [..]week
with [..]weather
with [..]wall
with [..]volunteer
with [..]voice)
with [..]victim
with [..]vet
with [..]vendor
with [..]user
with [..]union
with [..]unicorn
with [..]traitor
with [..]trainer
with [..]townsfolk
with [..]tower
with [..]tom
with [..]tailor
with [..]ta
with [..]surgeon
with [..]superintendent
with [..]submitter
with [..]stutter
with [..]student
with [..]stranger
with [..]spendthrift
with [..]spectrum
with [..]speaker
with [..]solicitor
with [..]smile
with [..]slur
with [..]skit
with [..]sister
with [..]shopkeeper
with [..]sense
with [..]senator
with [..]secretary
with [..]saviour
with [..]sand
with [..]salesperson
with [..]rocket
with [..]robot
with [..]representative
with [..]recruiter
with [..]rebuilder
with [..]ranger
with [..]purpose
with [..]public
with [..]psychologist
with [..]psychiatrist
with [..]prospect
with [..]prosecutor
with [..]proprietor
with [..]promoter
with [..]professor
with [..]professional
with [..]producer
with [..]proctor
with [..]principal
with [..]prince
with [..]priest
with [..]press
with [..]president
with [..]preacher

with [..]pope
with [..]pollster
with [..]policeman
with [..]police
with [..]pilot
with [..]photographer
with [..]party
with [..]participant
with [..]parrot
with [..]owner
with [..]organizer
with [..]opposition
with [..]operator
with [..]nose
with [..]nominee
with [..]newcomer
with [..]neighbor
with [..]musician
with [..]mother
with [..]morning

It is easy to see that this cluster tells us that one major schema for X <talk> Prep Y is the schema X <talk> with Y_human. If we then apply a dictionary that knows about the semantic types of nouns, it will be easy to identify the few non-human items above and decide on their function (e.g. *talk with a slur*, etc.).

Here is a list of the most frequent sequences of syntactic patterns for the subcorpora of length 5, 6 and 7 respectively (we indicate the number of different patterns per sentence length in parentheses and the frequency of the patterns themselves):

S5 (18550 patterns)	S6 (135217 patterns)	S7 (579367 patterns)
582330 X VV DT JJ NN	258117 X VV IN DT JJ NN	236630 X VV DT NN IN DT NN
469062 X VV IN DT NN	210544 X VV DT NN IN NN	100352 X VV DT NN IN PP$ NN
299118 X VV DT NN NN	148883 X VV TO VV DT NN	92691 X VV DT JJ NN IN NN
189555 X VV IN PP$ NN	141413 X VV IN DT NN NN	68469 X VV IN DT NN IN NN
163842 X VV VV DT NN	141059 X VV RB IN DT NN	66276 X VV VVN IN DT JJ NN
137028 X VV DT NN RB	140480 X VV DT JJ NN NN	63068 X VV TO VV IN DT NN
132721 X VV PP$ JJ NN	131227 X VV VVN IN DT NN	56322 X VV PP$ NN IN DT NN
107972 X VV DT JJ NNS	117475 X VV RB VV DT NN	55880 X VV TO VV DT JJ NN
106704 X VV IN JJ NNS	106844 X VV NN IN DT NN	54148 X VV DT NN IN JJ NN
99761 X VV IN JJ NN	80226 X VV PP IN DT NN	53193 X VV RB IN DT JJ NN
96758 X VV IN DT NNS	79417 X VV VV DT JJ NN	48611 X VV RB VVN IN DT NN
79206 X VV IN PP$ NNS	78326 X VV DT JJ JJ NN	47899 X VV DT NN IN JJ NNS
77067 X VV RB IN NN	72313 X VV DT NN IN NNS	46405 X VV RB VV DT JJ NN
76520 X VV VVN IN NN	67523 X VV TO VV PP$ NN	43463 X VV IN DT JJ NN NN
76469 X VV NN IN NN	65986 X VV IN PP$ JJ NN	42045 X VV DT NNS IN DT NN
71560 X VV VV PP$ NN	63627 X VV DT NN IN PP	39687 X VV NN IN DT JJ NN
69639 X VV PP$ NN NN	57972 X VV VVG IN DT NN	35136 X VV DT JJ NN IN NNS
67142 X VV IN NN NN	55963 X VV VV IN DT NN	35076 X VV DT NN IN NN NN
59745 X VV JJ CC JJ	50191 X VV DT RB JJ NN	34419 X VV DT NN IN DT NNS
59594 X VV TO VV NN	49862 X VV DT NN NN NN	33141 X VV VVN IN DT NN NN

Such sequences could easily serve as the basis for traditional syntactic rules (as is often done with treebank derived rules) but as one can quickly see, the individual schemata above are far from being sufficient to really classify individual sentences correctly since the patterns do not say anything about their propositional forms!

In fact, given the possibilities of our indexing technology and the corresponding query mechanisms, we can even find candidates for possible schemata semi-automatically by first extracting reasonable candidates with appropriate syntactic patterns and then verifying the results with additional filters and by manual inspection. We have done this by applying (relatively simple) noun phrase grammars to the corpora in question. Here are two example search patterns that we have used for this purpose on the set of sentences of length 5, 6, 7 and 8 (ARG is a pattern alias for a very small noun phrase grammar in regex form):

SB X v @4 ARG SE and
SB X v @2 ARG @2 ARG SE

The first pattern finds all instances of sentences with a subject noun phrase followed by a verb and then followed by zero to 4 arbitrary words before a noun phrase at the end of sentence (using the same noun phrase grammar as described above); the pattern is intended to find elementary sentences with two noun phrases as argument candidates). The second pattern is used to find elementary sentences with three noun phrases as candidates for arguments. In applying the extraction queries we also replace the strings matched by the noun phrase grammar represented by the ARG macro in the query by noun phrase variables. E.g. the sentence *X ⟨meet⟩ the president*, or *X ⟨talk⟩ about the game to John* will be represented in the result by *X ⟨meet⟩ Y*, or *X ⟨talk⟩ about Y to Z* in the result. The first pattern yields over seven million (!) such reduced patterns, most of which of course are not — for various reasons — directly relevant for consideration as two-place predicate argument relations. Here are the top 100 sentence schemata for the first query:

310378 X have Y
277427 X be Y
68684 X get Y
56234 X see Y
52237 X make Y
47938 X find Y
39498 X use Y
37578 X be in Y
37211 X need Y
30488 X like Y
26599 X do Y
24440 X want Y
21535 X know Y
20738 X take Y
20207 X call Y
19324 X become Y
18473 X go to Y
17194 X love Y
17031 X look at Y
14321 X receive Y
13962 X be on Y
13299 X be also Y
13189 X wear Y
13019 X lose Y
12506 X create Y
11506 X do with Y
11310 X meet Y
11268 X give Y
11241 X hear Y
11172 X enjoy Y
10244 X live in Y
10184 X buy Y
10132 X start Y
9638 X offer Y
9494 X reach Y
9294 X play Y
9269 X enter Y
9062 X do in Y
8969 X come to Y
8561 X join Y
8476 X develop Y
8407 X win Y
8348 X return to Y
8224 X think of Y
8209 X talk about Y
8130 X build Y
8079 X know about Y
7967 X feel Y
7959 X pick up Y
7880 X write Y
7853 X miss Y
7644 X be at Y
7590 X run Y
7579 X work on Y
7527 X be part of Y
7359 X say Y
7003 X get to Y
6959 X leave Y
6671 X work in Y
6585 X keep Y
6527 X provide Y
6484 X read Y
6477 X begin Y
6420 X remember Y
6403 X hold Y
6360 X want to be Y
6207 X get into Y
6199 X share Y
6165 X add Y
6130 X look like Y
6126 X go into Y
6121 X go through Y
6031 X choose Y
5946 X visit Y
5722 X see in Y
5717 X think about Y
5717 X support Y
5712 X have seen Y
5701 X set up Y
5688 X deal with Y
5645 X change Y
5549 X be just Y
5536 X attend Y
5521 X have on Y
5471 X pay for Y
5375 X be with Y
5332 X will make Y
5304 X work for Y

5270 X consider Y
5235 X understand Y
5234 X agree with Y
5217 X try Y
5215 X participate in Y
5212 X complete Y
5204 X form Y
5135 X put on Y
5122 X be given Y
5118 X follow Y
5057 X maintain Y
4955 X wait for Y
#6539 X give me Y
#5629 X give him Y
#5571 X give you Y
#5452 X call it Y
#5217 X think it was Y

Some of the schemata found obviously need to be removed (e.g. *X give me Y*) because they will also be captured by the second query), but it should be obvious that almost none of these schemata (and this goes for the less frequent ones as well) can be taken to express a two-place predicate argument structure as such. Each one needs to be dealt with individually both as far as the number of arguments and the type of the arguments is concerned. (We need of course to check whether a given observed pattern is not in fact a reduced form of a more complex structure, e.g. *X sell Y to Z* compared to *X sell Y to Z for U*, where the latter is in fact the maximal form.

In addition to the argument patterns found in terms of non-sentential arguments, we need to enumerate predicate-argument structures for expressions taking proposition-valued arguments as well. So far, we have concentrated on verbal patterns for which we have compiled a list of altogether almost 300 syntactic structures that can be used for the enumeration of predicate-argument structures.

4. Summary

Like Igor Mel'čuk, we view the notion of a «propositional form» (and the related «government pattern») as playing a central role in linguistic analysis if not the most basic one. One of the goals of the project reported on here is to transform our very large corpus — operation per operation — into an inventory of «elementary sentences» or better «elementary sentence schemas», each exhibiting as closely as possible its predicate-argument structure.

Bibliography

Blanco 2010 — *Blanco X.* Propriétés transformationnelles unaires en lexicographie informatique // META. 2010. P. 42—57.

Goller 2010 — *Goller J.* Exploring text corpora using index structures. PhD thesis, CIS. University of Munich, 2010.

Gross 1999 — *Gross M.* Lemmatization of compound tenses in English // Lingvisticae Investigationes. 1999. 22. P. 71—122.

Guenthner 2005a — *Guenthner F.* Corpus Calculus in a Nutshell. Unpublished CIS-Report. 2005.

Guenthner 2005b — *Guenthner F.* Local Grammars in Corpus Calculus // Dialog 2005. M., 2005. P. 616—620.

Guenthner (in preparation) — *Guenthner F.* Electronic Dictionaries, Tagsets and Tagging. Munich: Lincom, (in preparation).

Filatova 2011 — *Filatova N.* Vergleich von computerlinguistischen Verbklassifikationen. Master's Thesis. CIS, LMU. Munich, 2011.

Langer 1996 — *Langer S.* Selektionsklassen und Hyponymie im Lexikon. Semantische Klassifizierung von Nomina für das elektronische Wörterbuch CISLEX. CIS-Report. 1996.

Mel'čuk 2010 — *Mel'čuk I.* The Government Pattern in the Explanatory Combinatorial Dictionary // *De Schryver G.-M.* (eds.). A Way with Words: Recent Advances in Lexical Theory and Analysis. Menha Publishers, 2010.

Le Pesant, Mathieu-Colas 1998 — *Le Pesant D., Mathieu-Colas M.* (eds.). Langages 131. Les classes d'objets. Paris: Larousse, 1998.

Schuster 2010 — *Schuster J.* Towards Predicate-Driven Grammar. Munich: Lincom, 2010.

Zangenfeind 2010 — *Zangenfeind R.* Grammatik der Paraphrase. Munich: Lincom, 2010.