Linguistics behind the mirror

Karel Oliva

Institute of the Czech Language AS CR, v. v. i.
Letenská 123/4, Praha 1 - Malá Strana, CZ - 118 51, Czech Republic

Abstract. *A natural language is usually modelled as a subset of the set T^* of strings (over some set T of terminals) generated by some grammar G. Thus, T^* is divided into two disjoint classes: into grammatical and ungrammatical strings (any string not generated by G is considered ungrammatical). This approach brings along the following problems:*

- *on the theoretical side, it is impossible to rule out clearly unacceptable yet "theoretically grammatical" strings (e.g., strings with multiple centre self-embeddings, cf.* The cheese the lady the mouse the cat the dog chased caught frightened bought cost 10 £*),*
- *on the practical side, it impedes systematic build-up of such computational lingustics applications as, e.g., grammar-checkers.*

In an attempt to lay a theoretical fundament enabling the solution of these problems, the paper first proposes a tripartition of the stringset into:

- *clearly grammatical strings,*
- *clearly ungrammatical strings,*
- *strings with unclear ("on the verge") grammaticality status*

and, based on this, concentrates on

- *techniques for systematic discovery and description of clearly ungrammatical strings,*
- *the impact of the approach onto the theory of grammaticality,*
- *an overview of simple ideas about applications of the above in building grammar-checkers and rule-based part-of-speech taggers.*

1 Introduction

Apart from deciding on the membership of a particular string σ in a particular language L, a formal grammar is usually assigned an additional task: to assign each string from the language L some (syntactic) structure. The idea behind this is that the property of having a structure differentiates the strings $\sigma \in L$ from all "other" strings $\omega \notin L$, i.e. having a structure differentiates sentences from "nonsentences". Due to this, the task of identifying the appurtenance of a string to a language (the set membership) and the task of assigning the string its structure are often viewed as in fact identical. In other words, the current approach to syntactic description supposes that any string $\omega \in T^*$ which cannot be assigned a structure by the respective grammar is to be considered (formally) ungrammatical. Closely linked to this is also the presupposition that the borderline between strings which are grammatical and those which are ungrammatical is sharp and clear-cut.

Even elementary language practice (e.g., serving as a native speaker – informant for fellow linguists, or teaching one's mother tongue) shows that this presupposition does not hold in reality. The realistic picture is much more like the one in Fig. 1: there are strings which are considered clearly correct ("grammatical") by the native speakers, there are other ones that are doubtless incorrect (out of the language, "informally ungrammatical", unacceptable for native speakers), and there is a non-negligible set of strings whose status wrt. correctness (acceptability, grammaticality) is not really clear and/or where opinions of the native speakers differ (some possibly tending more in this, others more in the other direction, etc.).

Assuming the better empirical adequacy of the picture in Fig. 1, the objective of this paper will be to propose that a syntactic description of (some natural) language L should consist of:

- a formal grammar G defining the set $L(G)$ of doubtlessly grammatical strings ($L(G) \subseteq L$). Typically, the individual components of G (rules, principles, constraints, ...) are based on a structure assigned to a string, either directly (mentioning, e.g., the constituent structure) or indirectly, operating with other syntactically assigned features (such as subject, direct object, etc.). Since the description of the "clearly correct" strings via such a grammar is fairly standard, it will not be further dealt with here,
- a formal "ungrammar" U defining the set $L(U)$ of doubtlessly ungrammatical strings. Typically, any individual component ("unrule") of U would be based on lexical characteristics only, i.e. it would take recourse neither to any structure of a string nor to other syntactic characteristics (such as being a subject etc.), not even indirectly.

Unlike the standard approach, such a description allows also for the existence of a non-empty set of strings which belong to neither clearly grammatical nor clearly ungrammatical strings – more formally, such a description allows for a nonempty set $T^* \setminus (L(G) \cup L(U))$. Apart from this, the explicit knowledge of the set $L(U)$ of ungrammatical strings allows

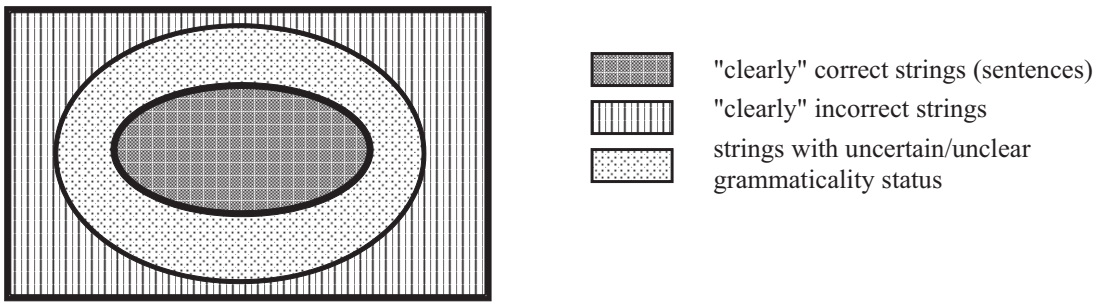

Fig. 1.

for straightforward development of important applications (cf. Sect. 4).

2 The unrules of the ungrammar

The above abstract ideas call for methods for discovering and describing the "unrules" of the "ungrammar". In doing this, the following two points can be postulated as starters:

- grammaticality/ungrammaticality is defined for whole sentences (i.e. not for subparts of sentences only, at least not in the general case)
- ungrammaticality occurs (only) as a result of violation of some linguistic phenomenon or phenomena within the sentence.

Since any "clear" error consists of violation of a language phenomenon, it seems reasonable that the search for incorrect configurations be preceded by an overview and classification of phenomena fit to the current purpose.

From the viewpoint of the way of their manifestation in the surface string, (syntactic) phenomena can be divided into three classes:

selection phenomena: in a rather broad understanding, selection (as a generalized notion of subcategorisation) is the requirement for a certain element (a syntactic category, sometimes even a single word) $E1$ to occur in a sentence if another element $E2$ (or: set of elements $\{E2, E3, \ldots, En\}$) is present, i.e. if $E2$ (or: $\{E2, E3, \ldots, En\}$) occur(s) in a string but $E1$ does not, the respective instance of selection phenomena is violated and the string is to be considered ungrammatical.

Example: in English, if a non-imperative finite verb form occurs in a sentence, then also a word functioning as its subject must occur in the sentence (cf. the contrast in grammaticality between *She is at home.* vs. **Is at home.*).

(word) order phenomena: word order rules are rules which define the mutual ordering of (two or more) elements $E1, E2, \ldots$ occurring within a particular string; if this ordering is not kept, then the respective word order phenomenon is violated and the string is to be considered ungrammatical.

Example: in an English *do*-interrogative sentence consisting of a finite form of the auxiliary verb *do*, of a subject position filled in by a noun or a personal pronoun in nominative, of a base form of a main verb different from *be* and *have*, and of the final question mark, the order must necessarily follow the pattern just used for listing the elements, or, in an echo question, it must follow the pattern of a declarative sentence. If this order is not kept, the string is ungrammatical (cf. *Did she come?*, *She did come?* vs. **Did come she?*, etc.).

agreement phenomena: understood broadly, an agreement phenomenon requires that if two (or more) elements $E1, E2, \ldots$ cooccur in a sentence, then some of their morphological characteristics have to be in a certain systematic relation (most often, identity); if this relation does not hold, the respective instance of the agreement is violated and the string is ungrammatical. (The difference to selection phenomena consists thus of the fact that the two (or more) elements $E1, E2, \ldots$ need not cooccur at all – that is, the agreement is violated if they cooccur but do not agree, but it is not violated if only one of the pair (of the set) occurs, which would, however, be a violation of the selection.)

Example: the string **She does it himself.* breaks the agreement relation in gender between the anaphora and its antecedent (while the sentences *She does it herself.* and *She does it.* are both correct – mind here the difference to selection).

This overview of classes of phenomena suggests that each string violating a certain phenomenon can be viewed as an extension of some minimal violating

$$\prec \oplus \left(\left[\begin{array}{c} \text{cat: n} \\ \text{gender:fem} \end{array} \right] \vee \left[\begin{array}{c} \text{cat: pron} \\ \text{pron\_type:pers} \\ \text{gender:fem} \end{array} \right] \right) \oplus \textit{himself} \oplus \succ \qquad (1)$$

string, i.e. as an extension of a string which contains only the material necessary for the violation. For example, the ungrammatical string *The old woman saw himself in the mirror yesterday,* if considered a case of violation of the anaphora-agreement relation, can be viewed as an extension of the minimal string *The woman saw himself,* and in fact as an extension of the string *Woman himself* (since for the anaphora-agreement violation, the fact that some other phenomena are also violated in the string does not play any role).

This means that a minimal violating string can be discovered in each ungrammatical string, and hence each "unrule" of the "formal ungrammar" can be constructed in two steps:

- first, by defining an (abstract) minimal violating string, based on a violation of an individual phenomenon (or, as the case might be, based on combination of violations of a "small number" of phenomena)
- second, by defining how the (abstract) minimal violating string can be extended into a full-fledged (abstract) violating string (or to more such strings, if there are more possibilities of the extension), i.e. by defining the material (as to quality and positioning) which can be added to the minimal string without making the resulting string grammatical (not even contingently).

The approach to discovering/describing ungrammatical strings will be illustrated by the following example where the sign '$\prec$' will mark sentence beginning (an abstract position in front of the first word), and '$\succ$' will mark sentence end (i.e. an abstract position "after the full stop").

Example: As reasoned already above, the abstract minimal violating string of the string *The old woman saw himself in the mirror yesterday* is the following configuration (1) (in the usual regular expression notation, using feature structures for the individual elements of the regular expression, '$\vee$' for disjunction, the sign '$\oplus$' for concatenation, and brackets '('and')' in the usual way for marking off precedence/grouping).

This configuration states that a string consisting of two elements (the sentential boundaries do not count), a feminine noun or a feminine personal pronoun followed by the word *himself*, can never be a correct sentence of English (cf., e.g., the impossibility of the dialogue *Who turned Io into a cow? *Hera himself.*)

Further, such a minimal violating (abstract) string can be generalized into an incorrect configuration of unlimited length using the following linguistic facts about the anaphoric pronoun *himself* in English:

- a bound anaphora must cooccur with a noun or nominal phrase displaying the same gender and number as the pronoun (with the binder of the anaphor); usually, this binder precedes the pronoun within the sentence (and then it is a case of a true anaphor) or, rarely, it can follow the anaphor (in case of a cataphoric relation: *Himself, he bought a book.*)
- occassionally, also an overtly unbound anaphora can occur; apart from imperative sentences *(Kill yourself!)*, the anaphor must then closely follow a *to*-infinitive (*The intention was only to kill himself.*) or a gerund *(Killing himself was the only intention.)*.

Taken together, these points mean that the only way how to give the configuration from the string (1) at least a chance to be grammatical is to extend it with an item which

- either, is in masculine gender and singular number
- or is an imperative or an infinitive or a gerund and stands to the left of the word *himself*.

This further suggests that – in order to keep the string ungrammatical also after the extension – no masculine gender and singular number item must occur within the (extended) string, as well as no infinitive or gerund must appear to the left of the word himself.

This can be captured in a (semi-)formal way (employing the Kleene-star '*' for any number of repeated occurrences, and '$\neg$' for negation) as follows.

In the first step, the requirement of no singular masculine is to be added (2), in the second step, the prohibition on occurrence of an imperative or an infinitive (represented by the infinitival particle *to*) or a gerund to the left of the word *himself* will be expressed as in (3). This is then the final form of description of an abstract violating string. Any particular string matching this description is guaranteed to be ungrammatical in English.

3 Ungrammar and the theory of grammaticality

An important case – mainly for the theory of grammaticality – of a minimal violating string is three fi-

$$\prec \oplus \left(\neg \left[\begin{array}{c}\text{number: sg}\\ \text{gender:masc}\end{array}\right]\right)^{*} \oplus \left(\left[\begin{array}{c}\text{cat: n}\\ \text{gender:fem}\end{array}\right] \vee \left[\begin{array}{c}\text{cat: pron}\\ \text{pron\_type:pers}\\ \text{gender:fem}\end{array}\right]\right) \oplus \left(\neg \left[\begin{array}{c}\text{number: sg}\\ \text{gender:masc}\end{array}\right]\right)^{*} \oplus \textit{himself} \oplus \left(\neg \left[\begin{array}{c}\text{number: sg}\\ \text{gender:masc}\end{array}\right]\right)^{*} \oplus \succ \quad (2)$$

$$\prec \oplus \left(\neg \left(\left[\begin{array}{c}\text{number: sg}\\ \text{gender:masc}\end{array}\right] \vee [\text{v\_form} : (\text{imp} \vee \text{ger})] \vee \left[\begin{array}{c}\text{cat:part}\\ \text{form:to}\end{array}\right]\right)\right)^{*} \oplus \left(\left[\begin{array}{c}\text{cat: n}\\ \text{gender:fem}\end{array}\right] \vee \left[\begin{array}{c}\text{cat: pron}\\ \text{pron\_type:pers}\\ \text{gender:fem}\end{array}\right]\right) \oplus \left(\neg \left(\left[\begin{array}{c}\text{number: sg}\\ \text{gender:masc}\end{array}\right] \vee [\text{v\_form} : (\text{imp} \vee \text{ger})] \vee \left[\begin{array}{c}\text{cat:part}\\ \text{form:to}\end{array}\right]\right)\right)^{*} \oplus \textit{himself} \oplus \left(\neg \left[\begin{array}{c}\text{number: sg}\\ \text{gender:masc}\end{array}\right]\right)^{*} \oplus \succ \quad (3)$$

nite verbs following each other closely, i.e. the configuration $VFin + VFin + VFin$. Such a configuration appears, e.g., in the sentence *The mouse the cat the dog chased caught survived* which is a typical example of – in its time frequently discussed – case of a multiple centre self-embedding construction. The important point concerning this construction is that it became the issue of discussions since

- one the one hand, this construction is – (almost) necessarily – licensed by any "reasonable" formal grammar of English, due to the necessity of allowing in this grammar for the possibility of (recursive) embedding (incl. centre self-embedding) of relative clauses
- on the other hand, such sentences are unanimously considered unacceptable by native speakers of English (with the contingent exception of theoretical linguists ☺).

The antagonism between the two points is traditionally attributed to (and attempted to be explained by) a tension between the langue (grammar, grammatical competence) and the parole (language performance) of the speakers, that is, by postulating that the speakers possess some internal system of the language but that they use the language in a way which deviates from this system. Such an assumption is generally a good explanation for such (unintentional) violations of langue (i.e. of grammaticality) in speech as, e.g., slips of tongue, hesitations and/or repetitions, etc., but it can hardly be used sensibly in case there are no extralinguistic factors and, above all, where the sentences in question correspond to the langue (to the grammatical description). This demonstrates that what is really at stake here is the correctness of the general understanding of the langue (and not a problem of a particular grammar of a particular language).

The difference in methods of ruling sentences with multiple centre self-embedding out of the language drives us to the fact that the standard view of langue – and hence that of a grammar – and the view advocated in this paper differ considerably:

- the standard approach to langue, which allows for specification of the set of correct strings only (via the grammar), has no means available for ruling out constructions with multiple centre self-embedding (short of ruling out recursion of the description of relative clauses, which would indeed solve the problem, however, would also have serious negative consequences elsewhere),
- the approach proposed, by allowing for explicit and most importantly independent specifications of the sets of correct and of incorrect strings as two autonomous parts of the langue, allows for ruling out constructions involving multiple centre self-embedded relative clauses (at least in certain cases); this is achieved without consequences on any other part of the grammar and the language described, simply by stating that strings where three (or more) finite verbs follow each other immediately belong to the area of "clearly incorrect" strings.

By solving the problem of unacceptability of the strings involving three (and more) finite verbs following each other via the formal ungrammar, the approach proposed enforces a refinement of perspective of the general description of grammaticality and ungrammaticality. In particular, from now on the Fig. 1 above has to be understood as depicting the situation in the language (understood as set of strings) only, i.e. without any recourse to the means of its description (i.e. without any recourse to a grammar and, in particular, to the coverage of a grammar). The coverage of the two grammar modules introduced above (the "grammar of the correct strings" and the "ungrammar of the incorrect strings"), i.e. the stringsets described by the components of the grammar describing the "clearly correct" and the "clearly incorrect" strings, should be rather described as in Fig. 2.

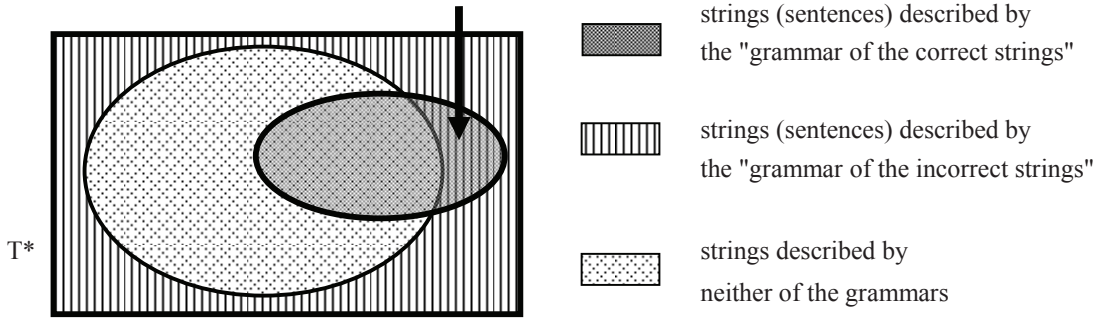

Fig. 2.

The crucial point is the part of this picture pointed out by the arrow (where dense dots and vertical bars overlap). This area of the picture is the one representing strings which are described by both components of the grammar, i.e. strings which are covered both by the description (grammar) of the correct strings and by the description (ungrammar) of the incorrect strings. At first glance, this might seem as a contradiction (seemingly, some strings are considered correct and incorrect simultaneously), but it is not one, since the true situation described in this picture is in fact two independent partitionings of the set of strings T^* by two *independent* set description systems, each of which describes a subset of T^*. Viewed from this perspective, it should not be surprising that some strings are described by both of the systems (while others are described by neither of them). The fundamental issue here is the relation of the two description systems (the grammar and the ungrammar) to the pretheoretical understanding of the notion of grammaticality as acceptability of a string for a native speaker of a language. Traditionally, all the strings were considered grammatical which were described by the grammar of the correct strings. In the light of the current discussion, and mainly of the evidence provided by the multiple centre self-embedding relative constructions, this definition of grammaticality should be adjusted by adding the proviso that strings which are covered by the description of incorrect strings (by the ungrammar) should not be considered grammatical (not even in case they are simultaneously covered by the grammar of the correct strings). This changes the perspective (compared to the standard one), by giving the ungrammar the "veto right" over the grammaticality of a string, but obviously corresponds to the language reality more closely than the standard approach.

Viewed from the perspective of a grammatical description considered as a model of a linguistic competence, the previous discussion can be summed up as follows:

- (formally) grammatical strings are strings described by the grammar but not by the ungrammar
- (formally) ungrammatical strings are strings described by the ungrammar
- strings whose grammaticality is (formally) undefined are strings which are described neither by the grammar not by the ungrammar.

4 Applications

In the previous sections, rather theoretical issues concerning the general view of grammaticality and means of description of grammatical/ungrammatical strings were dealt with. The task of finding the set of strictly ungrammatical strings has also a practical importance, however, since for certain applications it is crucial to know that particular configuration of words (or of abstractions over strings of words, e.g., configurations of part-of-speech information) is guaranteed to be incorrect.

The most prominent (or at least: the most obvious) among such tasks is (automatic) **grammar-checking**: the ability to recognize reliably that a string is ungrammatical would result in grammar-checkers with considerably more user-friendly performance than most of our present ones display, as they are based predominantly on simple patter-matching techniques, and hence they produce a lot of false alarms over correct strings on the one hand while they leave unflagged many strings whose ungrammaticality is obvious to a human, but which cannot be detected as incorrect since their inner structure is too complex or does not correspond to any of the patterns for any other reason.

Another practical task where the knowledge of the ungrammar of a particular language may turn into the central expertise needed is **part-of-speech tagging**, i.e. assigning morphological information (such as part-of-speech, case, number, tense, ...) to words

in running texts. The main problem for (automatic) part-of-speech tagging is morphological ambiguity, i.e. the fact that words might have different morphological meanings (e.g., the English wordform *can* is either a noun ("a food container") or a modal verb ("to be able to"); a more typical – and much more frequent - case of ambiguity in English is the noun/verb ambiguity in such systematic cases as *weight, jump, call, ...*). The knowledge of ungrammatical configurations can be employed for the build-up of a part-of-speech tagger based on the idea of (stepwise) elimination of those individual readings which are ungrammatical (i.e. impossible) in the context of a given sentence. In particular, each extended violating string with n constituting members (i.e. a configuration which came into being by extending a minimal violating string of length n) can be turned into a set of disambiguation rules by stipulating, for each resulting rule differently, $(n-1)$ constituting members of the extended violating string as unambiguous and issuing a deletion statement for the n-th original element in a string which matches the constituting elements as well as the extension elements inbetween them. Thus, each extended violating string arising from a simple violating string of length n yields n disambiguation rules.

Example: The two-membered minimal violating string ARTICLE + VERB, after being extended into the configuration (in the usual Kleene-star notation) ARTICLE + ADVERB* + VERB, yields the following two rules:

Rule 1:

find_ a_ string consisting of (from left to right):

- a word which is an unambiguous ARTICLE (i.e. bears no other tag or tags than ARTICLE)
- any number of words which bear the tag ADVERB (but no other tags)
- a word bearing the tag VERB

delete_ the_ tag VERB *from* the last word of the string

Rule 2:

find_ a_ string consisting of (from left to right):

- a word bearing the tag ARTICLE
- any number of words which bear the tag ADVERB (but no other tags)
- a word which is an unambiguous VERB (i.e. it bears only a single tag VERB or it bears more than one tag, but all these tags are VERB)

delete_ the_ tag ARTICLE *from* the first word of the string

The (linguistic) validity of these rules is based on the fact that any string matching the pattern part of the rule on each position would be ungrammatical (in English), and hence that the reading to be deleted can be removed without any harm to any of the grammatical readings of the input string.

It is important to realize that the proposed approach to the "discovery" of disambiguation rules yields the expected results – i.a. rules corresponding to the Constraint Grammar rules given in standard literature (e.g., it brings the rule for English saying that if an unambiguous ARTICLE is followed by a word having a potential VERB reading, then this VERB reading is to be discarded, cf. [1, p. 11], and compare this to the example above). The most important innovative feature (wrt. the usual ad hoc approach to writing these rules) is thus *the systematic linguistic method* of discovering the violating strings, supporting the development of all possible disambiguation rules, i.e. of truly powerful Constraint Grammars. It is also worth mentioning that the idea of the method as such is language independent – it can be used for development of Constraint Grammars for most different languages (even though the set of the developed rules will be of course language-specific and will depend on the syntactic regularities of the language in question).

References

1. F. Karlsson, A. Voutilainen, J. Heikikilä, and A. Antilla (eds.) *Constraint grammar – a language-independent system for parsing unrestricted text.* Mouton de Gruyter, Berlin & New York, 1995.