SYNTACTIC NORMALIZATION OF SPONTANEOUS SPEECH*

Hagen Langer, University of Bielefeld, W-Germany

ABSTRACT

This paper presents some techniques that provide a standard parsing system for the analysis of ill-formed utterances. These techniques are feature generalization and heuristically driven deletions.

PROBLEM

Generally the development of grammars, formalisms and natural language processors is based on written language data or, sometimes, not real data at all, but invented 'example sentences'. This holds for both computational and general linguistics. Thus many parsing systems that work quite well for sentences like 1a. and 1b. fail, if they get applied to the authentic data in 2a. and 2b.:

1a. die Grundform ist nicht eckig
the basic form is not angular

1b. das blaue habe ich als Waage auf dem grünen liegen
I have got the blue one lying upon the$_{DAT}$ green$_{DAT}$ one$_{DAT}$ like a balance

2a. die die Grund die Grundform sind is nich is nich eckig
the the basic the basic form are is not is not angular

2b. das blaue hab ich als Waage auf das grüne liegen
I have got the blue one lying upon the$_{ACC}$ green$_{ACC}$ one$_{ACC}$ like a balance

To native recipients the utterances in 2. appear to be more or less defective, but interpretable expressions. Moreover, the interpretation of 2a. or 2b. might require even less effort than, for instance, understanding an absolutely grammatical 'garden path sentence'. Since utterances like 2a. and 2b. occur quite frequently in spontaneous speech, an approach to parsing everyday language has to provide techniques that cover repairs, ungrammatical repetitions (2a.), case-assignment violation (2b.), agreement errors and other phenomena that have been summarized under the label 'ill-formed' in earlier research (Kwasny/Sondheimer 1981, Jensen et al. 1983, Weischedel/Sondheimer 1983, Lesmo/Torasso 1984, Kudo et al. 1988). Though the present paper will adhere to this terminology, it should be emphasized that it is not presupposed that there are any general criteria precise enough to tell us exactly whether some utterance is 'ill-formed' relative to a natural language. Let us assume, instead, that some utterance U is 'ill-formed (defective, irregular, ...) with respect to a grammar G' iff U is not a sentence of the language specified by G. Since, for instance, repairs exhibit a high degree of structural regularity (cf. Schegloff et al. 1977, Levelt 1983, Kindt/Laubenstein in preparation) one might prefer to describe them within the grammar and not within some other domain (e.g. within a production/perception model). Therefore the concept 'ill-formed' is used as a relational term that always has to be re-defined with respect to the given context.

There have been two main directions in the prior research on ill-formedness. The one direction has focussed on the problem of parsing ill-formed input in restricted domain applications, such as natural language interfaces to databases or robot assembly systems (Lesmo/Torasso 1984, Selfridge 1986, Carbonell/Hayes 1987). Though the techniques developed in that field seem to be quite adequate for the intended purposes, the results are not directly transferable to the interpretation of spontaneous speech, since the restrictions affect not only the topical domains but also the linguistic phenomena under consideration: e.g. the CASPAR parser (cf. Carbonell/Hayes 1987) is restricted to a subset of imperatives, Lesmo/Torasso (1984) achieve interpretations for ill-formed word order only at the price of neglecting long distance dependencies etc.

The other main direction has been the 'relaxation'-approach (Kwasny/Sondheimer 1981, Weischedel/Sondheimer 1983). The basic idea is to relax those grammatical constraints an input string does not meet, if a parse would fail otherwise. The main problem of this approach is that relaxing constraints (i.e. ignoring them) makes a grammar less precise. Thus, for instance, a noun phrase that lacks agreement in number is analysed as a noun phrase *without* number and it remains unexplicated how this analysis might support a further interpretation. Surprisingly, none of these papers concentrates on real life

*I am indebted to Dafydd Gibbon, Hans Karlgren and Hannes Rieser for their comments on earlier drafts of this paper. This research was supported by the Deutsche Forschungsgemeinschaft. Some aspects are discussed in more detail in Langer 1990.

spontaneous speech (most of them are explicitly concerned with written man-machine communication).

The present paper focusses the problem of *normalization*, i.e. how to define the relation between ill-formed utterances (e.g. 2a. and 2b.) and their well-formed 'counterparts' (1a. and 1b.). A sentence is an adequate normalization of an ill-formed utterance, if it corresponds to our intuitions about what the speaker might have intended to say. This is, of course, not observable, but a request for repetition (which typically does *not* give rise to a literally repetition in case of an utterance like 2a.) might serve as a suitable test.

In the present approach normalization is based on solely syntactic heuristics, not because syntactic information is regarded to be sufficient, but as a starting point for further work. Thus, the normalizations achieved on the basis of these heuristics serve as *default* interpretations that have to be evaluated using additional information about the linguistic and situational context. The empirical background is a corpus of authentic German dialogues about block worlds that has been recorded for the study of coherence phenomena (cf. Forschergruppe Kohärenz [ed.] 1987).

I will discuss three heuristics that are used in an experimental normalization system, called NOBUGS (NOrmalisierungskomponente im Bielefelder Unifikationsbasierten Analysesystem für Gesprochene Sprache - *normalization component of a Bielefeld unification-based speech analysis system*). The core of NOBUGS is a left-corner parser that interprets a GPSG-like formalism encoded in DCG notation. The grammars used with NOBUGS are very restrictive and exclude everything that is beyond the bounds of written standard German. But in combination with the heuristics I will discuss now the system is capable of handling a wider range of phenomena including morpho-syntactic deviations, explicit repair and ungrammatical repetitions.

MORPHO-SYNTACTIC DEVIATIONS

Morpho-syntactic deviations make up a considerable proportion of errors both in spoken and written German (German has a much more complex inflectional morphology than English).

The basic principle of this approach to normalization is as follows:

> Try to find out which properties of a given input string make a parse fail and use the given grammatical knowledge to alter the input string minimally so that it is as similar as possible to its initial state but without the properties that caused the failure.

What is meant by that can easily be seen if we consider an example where the property that makes a parse fail is evident, e.g. the string 'John sleep', which lacks the NP-VP-agreement concerning person and number that is required by the following rule:

cat = S		cat = NP	cat = VP
person = X_1	⟶	case = nom	person = X_1
num = X_2		person = X_1	num = X_2
		num = X_2	

This rule is not applicable to 'John sleep', since there are no lexical entries for 'John' and 'sleep', respectively, that have unifiable specifications for person and number, and this makes the whole parse fail.

The strategy to account for strings like 'John sleep' consists of three steps:

Step 1: Collect all lexical entries that match with the words of the input string and generalize them by substituting variables for their morpho-syntactic specifications (case, number, gender etc.).

Step2: Parse the string using the generalized lexical entries instead of the completely specified entries.

Step3: If the parse with generalized specifications is successful, the problem with the input string is morpho-syntactic (agreement error or case-assignment violation). Collect all preterminal categories (most of them still contain variable morpho-syntactic specifications) and try to unify them with full-specified lexical entries. At least one matching entry will belong to some item different from the corresponding word in the input string. In that case replace the original word by the matching item. If there are many different sets of matching entries choose the one that requires the least number of substitutions and output it as the default normalization (if there are many sets of matching entries that require the same least number of substitutions the normalization is ambigous. In that case output all of them).

Returning to our example string 'John sleep', let us assume that the grammar consists just of the rule stated above and the following lexical entries:

John: person = 3, num = sg, cat = np, case = nom
sleep: person = 3, num = pl, cat = vp
sleeps: person = 3, num = sg, cat = vp

Generalizing the lexical entries for the input string 'John sleep' will produce two new entries:

John: person = VAR_1, num = VAR_2, cat = np, case = VAR_3
sleep: person = VAR_4, num = VAR_5, cat = vp

A parse using these entries will be successful. The application of the rule unifies the variable specifica-

tions for number and person and instantiates case nominative in the NP. The preterminal categories resulting from the parse are:

person = VAR_1	person = VAR_1
num = VAR_2	num = VAR_2
cat = np	cat = vp
case = nom	

Though the crucial specifications (*person* and *num*) are still variable the difference is now that there are the same variables in both categories. The (only) set of lexical entries that match with these preterminal categories requires the replacement of 'sleep' by 'sleeps' and thus 'John sleeps' is the normalization of 'John sleep'.

Note that this strategy is not, in principle, limited to morpho-syntactic features. It might be useful for phonological and semantic normalization, as well.

EXPLICIT REPAIR

When people detect an error during an utterance they often try to correct it immediately. This, in general, makes the utterance as a whole ungrammatical. The structure of an utterance containing a self repair is often:

Left context - reparandum - repair indicator - reparans right context.

The reparandum is the part of the utterance that is to be corrected by the reparans. Typical repair indicators are interjections like 'uh no', 'nonsense', 'sorry' etc. The following example from our corpus shows that structure (note that the left context is empty in the original German version):

Den linken eh Quatsch den roten stellst du links hin
reparandum indicator reparans right context

You put the left one eh nonsense the red one to the left
left c. reparandum indicator reparans right context

A plausible normalization of this utterance would be 'Den roten stellst du links hin' (*'You put the red one to the left'*). This normalization differs from the original utterance in that the reparandum and the repair indicators have been deleted. The strategy to cover this type of repair is to scan the input string $w_1w_2...w_n$ until a repair indicator sequence $w_iw_{i+1}...w_j$ is found ($1 < i < j \leq n$). If there is such an explicit signal, then there probably is something wrong immediately before the repair sequence. But it is not clear what the reparandum is. Possibly the reparandum is just the word immediately before the repair indicator sequence or a longer substring or even the whole substring $w_1w_2...w_{i-1}$. Which deletion of a substring $w_kw_{k+1}...w_j$ gives a grammatical sentence can only be decided by the grammar. Thus it is necessary to parse the results of the alternative deletions beginning with $w_1...w_{i-2}$ $w_{j+1}...w_n$ and incrementing the length of the deleted substring until the parse succeeds. If the deletion of a substring $w_kw_{k+1}...w_j$ makes a parse successful and if there is no other deletion of a substring $w_lw_{l+1}...w_j$ such that $k < l$ then $w_1w_2...w_{k-1}w_{j+1}w_{j+2}...w_n$ is the normalization of the input string.

If applied to the utterance 'You put the left one eh nonsense the red one to the left' the first deletion gives 'You put the left the red one to the left' which is not accepted by the parser. The second alternative tried ('You put the the red one to the left') fails, too. But the third attempt ('You put the red one to the left') is accepted by the parser and thus considered as the normalization of the original utterance.

UNGRAMMATICAL REPETITIONS

Ungrammatical repetitions of single words or longer stretches occur quite frequently in spontaneous speech. As long as a sequence is repeated completely and without any alteration it is easy to detect the redundant duplication and remove it from the input string to get a normalized version. The problem is with incomplete repetitions and repetitions that introduce new lexical items:

```
Some blocks some red blocks are small.
\_______/ \_____________/
 part 1      part 2

Some red some blue blocks are small.
\_____/ \_____________/
 part 1     part 2
```

The deletion of the substrings indicated as 'part 1' in the utterances above, respectively, would yield a suitable normalization. Utterances of this kind are in many respects like the explicit repairs discussed above, but they lack indicators. Typically, part 2 is similar to part 1 in that at least some words occur in both substrings. Moreover, part 1 and part 2 often belong to the same category (e.g. NP in the utterances above). This similarity motivates the following heuristic:

> The input string $w_1w_2...w_n$ is scanned for two different occurrences, say w_i and w_j ($1 \leq w_i < w_j < w_n$), of the same lexical item. w_i and w_j are permitted to differ in their inflectional properties, since an unsuitable inflection of w_i might have been the reason to repeat it in proper inflexion as w_j (e.g. 'He takes took a block'). If such a repetition is

found the substring beginning with the first occurrence up to the word immediately before the second occurrence (i.e. $w_i w_{i+1} \ldots w_{j-1}$) is parsed. If the parse is succesful and yields some category C for the substring, the next step is to find a prefix of $w_j w_{j+1} \ldots w_n$ that belongs to the same category C. If such a prefix exists and $w_1 w_2 \ldots w_{i-1} w_j w_{j+1} \ldots w_n$ is accepted as a grammatical sentence it is considered to be the suitable normalization.

Let us apply this strategy to the utterance 'Some blocks some red blocks are small'. Scanning this input string from the left to the right will immediately find the repeated lexical item 'some'. The parse of the substring 'Some blocks' results in an NP and thus a prefix of 'some red blocks are small' is searched for which is also an NP. Such a prefix is found (i.e. 'some red blocks') and therefore 'some red blocks are small' is tested if it is a grammatical sentence and, indeed, it is.

RESULTS, CONCLUSIONS, FURTHER TASKS

The normalization strategies outlined in this paper make a given standard parsing system applicable to certain language phenomena that occur frequently in spontaneous speech, but deviate from the standards of written language. Additional rules, special grammar formalisms or fixed parsing algorithms are not required.

If the parse succeeds, the analysis assigned to a deviating input is not only some partial structure description, but a well-formed sentence including its complete syntactic structure.

Preliminary tests have shown that the normalizations achieved by the strategies discussed in this paper are plausible default interpretations in most cases. Bad normalizations result from the lack of phonological, semantic and world knowledge. A typical example is 'Take a red block oh no blue block' which gets incorrectly normalized into 'Take a red blue block', if the grammar accepts 'block' being specified by two different color adjectives. If it does not, trying the next alternative according to the explicit-repair strategy described above will yield the most plausible result 'Take a blue block'. Another way to avoid the wrong normalization is to consult additional phonological information about the input string. It is very probable that there is a contrastive stress upon 'blue' in the input utterance. Let us assume the rule: if there is a word with contrastive stress in a reparans sequence then there must be a suitable word in the reparandum sequence to which it is in contrast. This implies that 'red' must be part of the reparandum (and thus has to be deleted) and rules out the wrong normalization 'Take the red blue block'. A further task will be to find out how additional semantic and phonological information both in the grammar and in the normalization strategies can be used to make the normalization results more reliable.

REFERENCES

Carbonell, J.G./Hayes, P.J.: Robust parsing using multiple construction-specific strategies. In: Bolc, Leonard[ed.]: Natural language parsing systems. Berlin 1987. pp. 1-32. (Springer series symbolic computation - artificial intelligence).

Forschergruppe Kohärenz [ed.]: "n Gebilde oder was" - Daten zum Diskurs über Modellwelten. KoLiBri-Arbeitsbericht 2. Bielefeld 1987.

Jensen, K./Heidorn, G.E./Miller, L.A./Ravin, Y.: Parse fitting and prose fixing: getting a hold on ill-formedness. In: AJCL 9 (1983), 147-160.

Kindt, W./Laubenstein, U.: Reparaturen und Koordinationskonstruktionen. KoLiBri-Arbeitsbericht 20. (In preparation).

Kudo, I./Koshino, H./Chung, M./Morimoto, T.: Schema method: A framework for correcting grammatically ill-formed input. In: COLING 1988, 341-347.

Kwasny, S.C./Sondheimer, N.K.: Relaxation techniques for parsing ill-formed input in natural language understanding systems. In: AJCL 7 (1982), 99-108.

Langer, H.: Syntaktische Normalisierung gesprochener Sprache. KoLiBri-Arbeitsbericht 23. Bielefeld 1990.

Lesmo, L./Torasso, P.: Interpreting syntactically ill-formed sentences. In: COLING 1984, 534-539.

Levelt, W.J.M.: Monitoring and self-repair in speech. In: Cognition 14 (1983), 41-104.

Schegloff, E.A./Jefferson, C./Sacks, H.: The preference for self-correction in the organization of repair in conversation. In: Language 53 (1977), 361-382.

Selfridge, M.: Integrated processing produces robust understanding. In: CL 12 (1983), 161-177.

Weischedel, R.M./Sondheimer, N.K.: Meta-Rules as a basis for processing ill-formed input. In: AJCL 9 (1983), 161-177.