

Efficiency of Morphological Analysis in ALEP

Axel Theofilidis

IAI

Martin-Luther-Str. 14

66111 Saarbrücken, Germany

axel@iai.uni-sb.de

Abstract

Starting from a brief outline of the approach to morphological analysis supported by the ALEP linguistic sub-system, we will present the results of a small-scale experiment that aims at quantifying the parse time that is consumed by morphotactic analysis relative to overall parse time. This experiment was based on the German LS-GRAM grammar implemented in ALEP and providing a full account of inflectional morphology. Tracing the process of morphotactic parsing with this grammar revealed a high degree of non-determinism being introduced by the process of word segmentation. Our experiment shows that parsing efficiency significantly decreases due to this kind of non-determinism. Options will be presented of how to reduce non-determinism of morphological analysis in ALEP.

1 Morphological Analysis in ALEP

The ALEP linguistic sub-system supports a two-stage approach to morphological analysis consisting of the operations of morphographemic and morphotactic analysis. Morphographemic analysis is concerned with recognizing morpheme strings constitutive of word forms and, by this, segmenting word forms into sequences of morpheme strings. Morphotactic analysis, on the other hand, is concerned with combining morphemes to word form representations and, by this, interpreting the relations which hold between morphemes in terms of morpho-syntactic information being established.

To illustrate the two operational steps consider the German word form *'kaufen'*. The first operation recognizes the word form *'kaufen'* as being composed of the morpheme strings *'kauf'* and *'en'*. The second operation interprets the relation between the morpheme strings *'kauf'* and *'en'* as a relation between

the verb stem *'kauf'* and the 1st or 3rd person plural verb agreement or infinitive suffix *'en'*.

Morphographemic analysis is performed by a two-level algorithm establishing characterwise correspondences between surface word strings and underlying lexical morpheme strings. Such correspondences are modelled in terms of two-level rules consisting of a rule identifier and a two-level description possibly annotated with constraints over variables or with lexical filters as suggested in (Trost(1990)). The two-level rule shown in Figure 1, for instance, accounts for the possibility of any surface sequence of two alphabetical characters being separated by a morpheme boundary (marked by the diacritic symbol '+'), provided that the lexical filter applies which claims that the morpheme to the left of the morpheme boundary is licensed to figure in non-terminal position (it should be noted that lexical filters, in the ALEP two-level formalism, are generally applied wrt. to the morpheme string to the left of the morpheme boundary; we will come back to this point shortly). Thus, rule 'segment' strives for word segmentation:

```
tlm_rule(  
segment,  
[X] [] [Y] => [] [+] [],  
morphol[last no],  
[X in alphabet, Y in alphabet] ).
```

Figure 1: ALEP two-level rule

Morphotactic analysis, on the other hand, is performed by either one of two head driven parsing algorithms establishing immediate dominance relations between word form descriptions and their morphological constituents in terms of context-free re-write rules. Attachment of affixes, for instance, will be accounted for by a re-write rule representing the abstract immediate dominance scheme shown in Figure 2. It should be noted that morphotactic analysis in ALEP is fully integrated with phrase structure parsing such that output structures will generally

be trees having span from the sentence node to morpheme leaves.

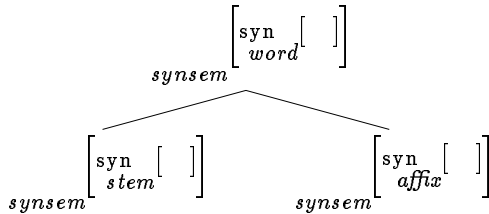


Figure 2: *Affixation scheme*

2 Non-Determinism in Morphological Analysis

Currently, a strict division of labour is assumed in ALEP between morphographemic and morphotactic analysis in that it is not foreseen to anticipate morphotactic constraints during morphographemic analysis already. The reason is that, like most two-level formalisms, the ALEP two-level formalism does not support expressing informational constraints between any two adjacent word segments obtained by morphographemic analysis. Lexical filters may be annotated only once per two-level rule, and within rules inserting a morpheme boundary as the right-hand side element of the two-level description lexical filters are always interpreted wrt. the lexical string to the left of the morpheme boundary symbol. By consequence, it is not possible to express any selectional or co-occurrence restrictions wrt. two adjacent morphemes recognized during morphographemic analysis. For any inflected word form being segmented during morphographemic analysis, for instance, it is not possible to claim that the segments share their part-of-speech specification.

To illustrate this point, once more consider the German word form *kaufen*. Before it is segmented during the process of morphographemic analysis, this word form can unambiguously be identified as a verb form. Once segmented, however, it is no longer possible to separately determine for each of the two morphemes which particular part-of-speech category they instantiate in the given syntagmatic context. *kauf* qualifies both as a verb and noun stem (that is, if no treatment of derivation is adopted) and *en* qualifies as a verb, noun, or adjective inflectional affix. Thus, it is left to the operation of morphotactic analysis to figure out which of the alternative morphemes combine to constitute the word form *kaufen* and, in that, to determine the part-of-speech characteristics of this word form. Put more radically, morphographemic analysis results, in fact, in a loss of information, only in order to recover this information in a processing-intensive time-consuming manner during morphotactic analysis.

Against the background of this consideration, we have performed a small-scale experiment which aims at quantifying the effect which the mentioned kind of non-determinism residing in the domain of morphological analysis has on overall parse time. The experiment was based on the German LS-GRAM grammar which provides a rather complete account of inflectional morphology (cf. (Schmidt et al. (1996))).

We have chosen two sample sentences which were run once with the standard German LS-GRAM lexicon providing all required morpheme entries (i.e. stems and affixes), and a second time using a lexicon, where all entries for the stems constitutive of the inflected word forms occurring in the two sample sentences were commented out and replaced by fullform entries for the respective word forms. Thus, in the second run, non-determinism due to the availability of alternative morpheme entries has been removed at the level of lexical description, with the fullform entries being designed such that no change was required wrt. the structure component of the grammar. The two sentences are the following:

- (1) Dieser Erfolg überrascht in zwei Hinsichten.
- (2) Große Bereiche der Dasa leiden unter dem Rückgang des einst lukrativen Rüstungsgeschäfts.

As for sentence (1), there are two word forms marked by inflectional affixes: *überrascht* consisting of the stem *überrasch* and the affix *t*, and *Hinsichten* consisting of the stem *hinsicht* and the affix *(e)n*. As for sentence (2), it comprises five word forms marked by inflectional affixes: *große* consisting of the stem *groß* and the affix *e*, *Bereiche* consisting of the stem *bereich* and the affix *e*, *leiden* consisting of the stem *leid* and the affix *(e)n*, *lukrativen* consisting of the stem *lukrativ* and the affix *(e)n*, and finally *Rüstungsgeschäfts* consisting of the stem *rüstungsgeschäft* and the affix *s*.

For each of the affixes *t* and *(e)n*, the standard German LS-GRAM morpheme lexicon assumes 7 entries corresponding to distinct readings of these affixes in terms of part-of-speech characteristics. For the affixes *e* and *s*, it assumes 5, respectively 3, entries. Thus, non-determinism caused by multiple affix entries can be crudely quantified as follows for sentence (1), respectively sentence (2), relative to the overall number of words occurring in these sentences:

$$(3) \frac{7+7}{6} = 2.33$$

$$(4) \frac{5+5+7+7+3}{12} = 2.25$$

Figure 3 shows the performance figures obtained for the two sentences on the basis of the standard German LS-GRAM lexicon (TEST A) and the modified lexicon (TEST B) using both the basic and the record parser of ALEP.

Input : Dieser Erfolg überrascht in zwei Hinsichten.

	BASIC		RECORD	
	1st	all	1st	all
	solution	solutions	solution	solutions
TEST A	0.470	2.580	1.960	1.970
TEST B	0.410	1.420	1.230	1.230

Input : Große Bereiche der Dasa leiden unter dem Rückgang des einst lukrativen Rüstungsgeschäfts.

	BASIC		RECORD	
	1st	all	1st	all
	solution	solutions	solution	solutions
TEST A	29.090	39.760	6.960	6.960
TEST B	24.960	32.590	5.000	5.010

Figure 3: Parse results (secs.) of the two test runs using the basic and the record parser

The increase of efficiency to be observed with TEST B is significant though perhaps not as drastic as one might have expected. If we use the formula $x = \frac{\text{TESTA} - \text{TESTB}}{\text{TESTA}}$ in order to quantify for TEST A the parse time consumed by analysing the inflected word forms, we obtain the figures shown in Figure 4. Thus, in parsing sentence (2) with the record parser, for instance, morphotactic analysis consumes more than one fourth (0.28) of the overall parse time. This is significant enough to induce some thought about how to reduce non-determinism residing in the domain of morphological analysis.

	basic	record
	algorithm	algorithm
sentence (1)	0.45	0.38
sentence (2)	0.18	0.28

Figure 4: Factor indicative of the parse time consumed by morphotactic analysis in TEST A relative to overall parse time

3 Options of Reducing Non-Determinism of Morphological Analysis in ALEP

Non-determinism residing in the domain of morphological analysis can be reduced either at the level of linguistic description or at the level of processing functionality.

At the level of linguistic description various strategies can be pursued in order to reduce multiple morpheme entries. As for multiple affix entries, one may consider the option of moving from an affix-based to an ending-based approach to (inflectional) morphology, thus reducing the number of multiple affix en-

tries by treating sequences of infixes and suffixes as single units.¹ In accounting for German inflectional morphology, for instance, this approach would eliminate the entry for the preterite infix ‘t’, as well as for the preterite agreement suffix ‘e’, since both affixes in combination would be treated as one ending. Thus, the verb form ‘kaufte’ would be segmented as ‘kauf+te’ rather than as ‘kauf+t+e’.

A second option is to collapse distinct entries for the same affix string into a single entry by means of a boolean disjunctive encoding of the distinguishing dimension of information. The German LS-GRAM grammar, for instance, encodes verb form information in terms of disjoint types. By consequence, distinct entries must be provided for respective affixes, e.g. ‘(e)n’ marking the infinitive verb form, respectively the finite 1st and 3rd person plural verb form. By merging verb form and agreement information into one boolean feature, the two readings of the affix ‘(e)n’ could be collapsed into a single entry at the cost, however, of losing a structured account of the two dimensions of information.

A third option, realized in the Danish LS-GRAM module (cf. (Music & Navarretta (1996))), is to completely move morphotactic analysis from the parsing stage to the preceding stages of processing, that is, morphographemic analysis and the lift operation which converts output structures obtained from texthandling and morphographemic analysis to data structures suitable for parsing. The core idea, here, is to featurize inflectional suffixes by means of distinct two-level rules which delete the suffix while assigning the preceding stem a distinctive feature (cf. Figure 5). This information is then accessed during the lift operation in order to assign agree-

¹This approach was adopted in several LS-GRAM modules, e.g. the Danish module; cf. (Music & Navarretta (1996)).

ment or other information in accordance with the word class and the distinctive feature assigned during morphographemic analysis.

```
t1m_rule(
suffix_en,
[X] [e,n,=] [] => [] [+ ] [],
      [
      infl      en
      grf      tgrf [fuge 'n']
      contin   '-en'
      seq      last
      ]
tstem [
[X in alphabet, Y in alphabet] ).
```

Figure 5: *Two-level based featurization of suffixes*

Irrespective of such options of reducing morphological non-determinism at the level of linguistic description, solutions should be sought of meeting with non-determinism at the level of processing functionality. An obvious solution at hand is to enhance the ALEP two-level formalism and underlying algorithm such that morphotactic constraints may be dealt with during morphographemic analysis already. This would require the following amendments:

(i) The diacritic symbol representing lexical morpheme boundaries is rendered ambiguous wrt. whether it represents the beginning or the end of a morpheme string (since lexical filters are currently always interpreted wrt. the string occurring to the left of the diacritic morpheme boundary symbol, this symbol actually functions as a morpheme *end* boundary marker).

(ii) Lexical filters may be expressed both wrt. the string occurring to the left and wrt. the string occurring to the right of a morpheme boundary symbol figuring as the righthand side of a two level description; whether a lexical filter is to be interpreted wrt. to the string to the left or right of a morpheme boundary symbol will be indicated by embedding lexical filters in a respective term.

On the basis of these amendments, basic morphotactic constraints (e.g. selection of a stem by its affix) could be directly encoded in two-level rules as illustrated here:

```
t1m_rule(
segment_stem_affix,
[X] [ ] [Y] => [] [+ ] [],
left( [ ] stem [last no] ),
right( [ ] affix [selects [ ] ] ),
[X in alphabet, Y in alphabet] ).
```

Figure 6: *Two-level rule encoding morphotactic constraints*

An alternative (and perhaps less demanding) solution would be to leave the ALEP two-level machinery as is, but to provide feature inheritance mechanisms operating on output structures obtained from morphographemic analysis, and enforcing a sharing of some feature value (e.g. a part-of-speech tag) for the segments constitutive of a word form. Respective mechanisms are already available (though not yet sufficient) as part of the ALEP texthandling component.

For both of the proposed solutions, benchmark tests would have to prove that the respective enhancements do not neutralize the envisaged gain of efficiency due to the additional processing effort being required.

References

- Music, B. & Navarretta, C. (1996) 'Documentation of the Danish Lingware' (LRE 61029, Deliverable E-D8-DK, CST, Copenhagen (<http://www.iai.uni-sb.de/LS-GRAM>)).
- Schmidt, P., Rieder, S., Theofilidis, A., Declerck, T. (1996) 'Final Documentation of the German LS-GRAM Lingware' (LRE 61029, Deliverable DC-WP6e (German)), IAI, Saarbrücken (<http://www.iai.uni-sb.de/LS-GRAM>)).
- Trost, H. (1990) 'The Application of Two-Level Morphology to Non-Concatenative German Morphology', in: *COLING-90*, 371-376.