Testing a Word Analysis System for Reliable and Sense-Conveying Hyphenation and Other Applications

Martin Schönhacker*, Gabriele Kodydek

Institute of Computer Graphics, Algorithms and Data Structures Group,
Vienna University of Technology, Favoritenstraße 9–11/186, A–1040 Vienna, Austria
*currently: Franklin & Marshall College, Lancaster, PA 17604–3003, USA
m_schoenhack@acad.fandm.edu, {schoenhacker,kodydek}@apm.tuwien.ac.at
http://www.apm.tuwien.ac.at/

Abstract. In this article, we present a test environment for a word analysis system that is used for reliable and sense-conveying hyphenation of German words. A crucial task is the hyphenation of compound words, a huge set of those can readily be formed from existing words. Due to this fact, testing and checking all existing words for correct hyphenation is infeasible. Therefore we have developed special test methods for large text files which filter the few problematic cases from the complete set of analyzed words. These methods include detecting unknown or ambiguous words, comparing the output of different versions of the word analysis system, and choosing dubious words according to other special criteria. The test system is also suited for testing other applications that are based on word analysis, such as full text search.

1 Introduction

We propose methods for testing a word analysis system for reliable and sense-conveying hyphenation of German words. A particular goal of the presented test environment was to enable especially the comparison of the results of different versions of such a word analysis system.

In the next section we will give a short description of the hyphenation method SiSiSi (the German acronym for "Sichere Sinnentsprechende Silbentrennung"). In Sect. 3 we will describe the test system. The results of empirical tests follow in Sect. 4. Finally, we conclude by describing the possibilities to incorporate other applications into the presented test environment in Sect. 5.

2 SiSiSi: Reliable and Sense-Conveying Hyphenation

The SiSiSi method provides reliable and sense-conveying hyphenation of German words [1], see also [3]. Since the most difficult problem in this context is the hyphenation of (long) compound words, the application is based on word analysis: Compound words are split into single words, which are in turn decomposed into atomic parts (= atoms) according to the rules for word formation in German. With respect to their use in word synthesis, atoms can coarsely be classified into prefixes, stems, and suffixes. In practice, it turns out to be advisable to use a more detailed classification of atoms, which in particular allows for a limitation of the sets of permissible suffixes for certain classes of words [6, 7]. The word analysis system consists of two basic building blocks: the atom table, which contains all the atomic components of words, and a recursive decomposition algorithm, which actually performs the word analysis and splits words into atoms according to grammar rules. After analysis, hyphenations between consecutive single words and between prefixes and stems have already been identified. Thereafter, an algorithm based on the sequence of consonants and vowels is used to find any hyphens in the identified stem-suffix parts and in polysyllabic prefixes such as *unter* (under).

The result of hyphenation for every given input word is a set of hyphenation variants for that word. Points of hyphenation between single words of a compound are marked as major hyphenation points (”Haupttrennstellen”, displayed as ”=”), those within single words as minor hyphenation points (”Nebentrennstellen”, represented by ”-”). Some minor hyphenation points such as those following closely after a major hyphenation point are marked by ”_” because it is less advisable to use these hyphens for end-of-line divisions. The word *Wort=zer_le-gungs=ver_fah-ren* (word decomposition method) gives a good example for the usage of the different types of hyphens.

Different variants of hyphenation for a particular word often result from the fact that there are several grammatically correct decompositions for that word, although not all of them are necessarily meaningful, e.g. *Per-so-nal=man_gel* (manpower shortage) and *Per-son=alm=an_gel* (formed from the components ”person”, ”mountain pasture”, and ”fishing rod”), where the latter obviously does not make much sense. Only hyphens contained in all hyphenation variants are considered to be reliable and are used in the final result.

3 Test System

A considerable amount of testing is needed to ensure that no correctly spelled word will receive an incorrect hyphenation, and orthographically and grammatically illegal words will be recognized as such. Because SiSiSi looks for all allowed decompositions of a word, the following holds: If the atom table contains all atomic components of the German language and the implemented word formation rules do not exclude any correct word formation, then for any correctly spelled word none of the found *reliable hyphenation points* (i.e. occurring in all decompositions) is wrong. This immediately results from the fact that the set of found decompositions contains the correct decomposition.

Ideally, any existing word should be tested and checked for. However, the possibility to form an infinite number of new compound words make this approach infeasible. Therefore, selective tests are used which present to the user only words matching certain filter criteria.

Hyphenation vectors:

`+ wach=stu_ben`	vector for the meaning "police offices"
`+ wachs=tu_ben`	vector for the meaning "tubes of wax"
`= wachstu_ben`	combined vector, containing only reliable hyphenation points

Sequence of atoms:

`+ wach·stube·n`
`+ wachs·tube·n`

Sequence of attributes (when using simple grammar rules as in UrSi):

`+ wach=stuben`	which corresponds to:
`+ S...=S....E`	stem *wach* + stem *stube* + suffix *n*
`+ wachs=tuben`	which corresponds to:
`+ S....=S...E`	stem *wachs* + stem *tube* + suffix *n*

Fig. 1. Example for the various representations of a result

A test system has been implemented which tests word analysis and hyphenation. The system incorporates three different versions of SiSiSi. Two of these versions use a very detailed classification of atoms and a grammar derived from German word formation rules. The first one (denoted ReSi) adheres to the new rules, which have come into effect in 1998 as a result of the reform of German orthography; the other one (HelSi) adheres to the old rules which continue to be in effect until 2005. For purposes of comparison, a primitive version (UrSi) using an extremely simple word grammar according to the old rules for German orthography has also been included; however, this version allows a considerable number of meaningless compounds, such as attaching the same suffix repeatedly to a stem.

The test system presented here allows for easy testing of hyphenation based on word analysis. Both isolated words and entire text files in various formats (that is with different representations of the German umlauts ä, ö, ü and other special characters such as ß or é) can be tested. The results for hyphenating a given word are being presented in a structured format on screen, and can also be logged to an output file. On screen different views are used in order to get an insight into the functionality of the algorithm.

As an example, we use the compound word *wachstuben*. It can be formed in two different ways that are both legal and meaningful, but render different ways of hyphenation: either from the words *wach* (to watch) and *stuben* (rooms), meaning "police office", or from *wachs* (wax) and *tuben* (tubes). Notice that *stuben* is formed using the stem *stube* plus a plural suffix *n*; *tuben* is formed analogously. The views for this example are given in Fig. 1.

The view called *hyphenation vectors* shows all distinct ways for hyphenating the word plus the so-called combined hyphenation vector at the bottom, which is the result of combining all hyphenation variants. It contains only the reliable hyphenation points for the given word, which are those hyphens that occur in all variants. A second view called *sequence of atoms* shows all the constituents of each distinct decomposition, separated by a dot. The most details are presented in the view called *sequence of attributes*, which shows for each decomposition all atoms along with their respective function.

3.1 Special Methods

The test system allows for immediate comparison of the results of all three versions of SiSiSi, as well as for intense testing of any of them.

Filtering *unknown words* from the large amount of analyzed words in the test file is probably the most obvious way of finding problematic words. We call a word *unknown*, if the system is not able to find a legal decomposition for it and therefore cannot provide proper hyphenation for it. Mostly such a word is not spelled correctly. Otherwise an error of our system (an incorrect rule or a missing atom) has been found. As a third possibility, the tested word might be a biographical or geographical name, an abbreviation, or an uncommon (possibly foreign) word.

Additionally, the test environment contains a function for detecting all *ambiguous words*, that is all words which have at least two distinct hyphenation vectors because they can be broken up in more than one way. Undesired ambiguities, such as in the case where one of the decompositions is absolutely meaningless, can be eliminated by adding the desired compound word (without suffixes) with an appropriate major hyphenation point to the atom table since atomic parts are not further decomposed by the algorithm.

When comparing two or all three of the versions, the output contains all words for which the considered variants do not find identical hyphenation vectors. The comparison of ReSi and HelSi is mainly used to detect words that were affected by the reform of the German orthography, whereas HelSi and UrSi are compared to examine the impact of the more sophisticated classification of atoms and grammar rules.

The rules of orthography permit the usage of hyphens within a compound word in order to disambiguate it or to increase its readability, such as in *Wach–Stube* (police office) or *Tee–Ei* (tea ball). The test system accepts an input word with such hyphens (–) between its constituent single words or with indicators for optional hyphenation points (˜) and only provides those hyphenation variants which comply with these predefined hyphens. This feature allows using correctly hyphenated words as input for checking whether the correct hyphenation variant is found by SiSiSi. It is therefore useful for identifying words for which SiSiSi provides no or only incorrect hyphenation, e.g. due to missing atoms.

The test system includes a function for restricting the length of the words to be analyzed to any interval. For example, the number of test cases can be reduced by ignoring words with a length less than four letters, since it is rather unlikely that wrong hyphenations for such short words are produced (according to the old rules they remain undivided; according to the new rules a hyphen can be set after a single initial vowel in polysyllabic words with at least three letters, but we mark this as an undesirable hyphenation point).

For the two versions based on detailed grammar rules (i.e. ReSi and HelSi), more selective tests using elaborate filter criteria have been implemented. These criteria have been chosen with care to exhibit the few potential problem cases. For instance, it is possible to select words which contain atoms with varied functionality and can therefore be problematic (e.g. *ende* can be used as a stem or as a suffix: *Spiel=en_de* (the end of the game) as well as *Spie-len-de* (players).

Two filter criteria allow the specification of a set of composition rules and derivation rules respectively in order to examine the validity and area of application of these rules for word composition and derivation. Another criterion allows filtering words which contain atoms of specific atom classes. This is useful for examining words which contain atoms of specified stem or suffix classes. Furthermore, a filter can be used to discard words which only contain a single atom since these words are either monosyllabic and therefore not hyphenated at all, or they are hyphenated according to the consonants-vowel-rules. Therefore no problems are expected for such words.

A considerable amount of testing conducted using this test system played a vital role in improving the word analysis system, e.g. by uncovering stem classification errors or missing stems (often naturalized foreign-language words), which could subsequently be corrected. The next section shows a short extract from the empirical tests we conducted using the presented test system.

4 Experimental Results

For our tests, we have tried to find large text files that cover all stems and preferably all types of word forms of the German language. Unfortunately, most dictionaries were not available to us in an appropriate form for our test runs or they did not contain a sufficient variety of word forms.

We used a test file with more than 209,000 different word forms[1] to look for unknown words using HelSi. Initially, 12.5% of the words were unknown. These words were further examined and used for updating the atom table. Note that the test file contains a large number of inflected word forms, derivations, and compound words. Therefore by adding a single atom, a large number of unknown words could be eliminated. Among the unknown words were biographical and geographical names and a number of word forms using these stems. The number of words reported as unknown is slightly biased due to some misspelled or abbreviated words in the test file, which our system correctly failed to recognize.

Furthermore, we have compared the ReSi and HelSi versions using a list provided in [5][2] which contains words and phrases which are affected by the reform of German orthography. From this list, two test files of about 2,000 words each were created with respect to the new and old orthography. The comparison showed the expected differences: ReSi and HelSi both correctly analyzed the words according to the corresponding spelling rules and recognized words spelled according to the respective other orthography rules. Furthermore, the results

[1] available via `http://www.sibiller.de/anagramme/`
[2] also available online via `http://www.duden.de/`

correctly pointed out words with letter combinations such as *st* and *ck* to which different hyphenation rules apply.

We plan to perform further tests using another set of test files that was created by exporting word forms from the Morphy morphological analyzer [4].

5 Outlook

Other applications based on word analysis can easily be incorporated into the test system, and can then also be tested as described. One example is sense-conveying full text search, which uses the word analysis algorithm to return a set of atoms which determine the meaning of a given input word [2]. The meaning is determined by the components of compound words and by prefixes and stems in simple words. For example, the word *Textverarbeitungssysteme* (text processing systems) contains {*text, verarbeitung, system; verarbeit*} (text, processing, system; process) as meaningful components. If there are several variants for forming a word from atoms, the meaning of that word may turn out to be indeterminable without context. The result will then contain several sets of atoms which provide meaning, e.g. *Baumast* → {*baum, ast*} (tree, branch), {*bau, mast*} (build, mast). For sense-conveying full text search, testing is necessary to detect words that cannot be broken up into their meaningful atoms in the correct manner. A function which returns the atoms providing meaning has already been included in the test system for all three versions. There are further plans to incorporate similar functions for spell checking, as well as for checking the proper capitalization of words, as all nouns are being capitalized in German.

References

1. Barth, W., Nirschl, H.: Sichere sinnentsprechende Silbentrennung für die deutsche Sprache. Angewandte Informatik 4 (1985) 152–159
2. Barth, W.: Volltextsuche mit sinnentsprechender Wortzerlegung. Wirtschaftsinformatik, vol. 32, no. 5 (1990) 467–471
3. Kodydek, G.: A Word Analysis System for German Hyphenation, Full Text Search, and Spell Checking, with Regard to the Latest Reform of German Orthography. To appear in Proc. of the Third Int. Workshop on Text, Speech and Dialogue, Brno, Czech Republic (2000)
4. Lezius, W., Rapp, R., Wettler, M.: A Freely Available Morphological Analyzer, Disambiguator, and Context Sensitive Lemmatizer for German. In: Proceedings of the COLING-ACL 1998, Montreal, Canada (1998)
5. Sitta, H., Gallmann, P.: Duden, Informationen zur neuen deutschen Rechtschreibung, ed. Dudenredaktion, Dudenverlag, Mannheim (1996)
6. Steiner, H.: Automatische Silbentrennung durch Wortbildungsanalyse. PhD thesis, Institute of Computer Graphics, Vienna University of Technology (1995)
7. Steiner, H., Barth, W.: Sichere sinnentsprechende Silbentrennung mit Berücksichtigung der deutschen Wortbildungsgrammatik. Tagungsband Konvens'94, ed. H. Trost, Vienna (1994) 330–340

This project was in part supported by *Hochschuljubiläumsstiftung der Stadt Wien* under the grant number H-75/99.