Si3Trenn and Si3Silb: Using the SiSiSi Word Analysis System for Pre-Hyphenation and Syllable Counting in German Documents

Gabriele Kodydek and Martin Schönhacker

Institute of Computer Graphics and Algorithms,
Algorithms and Data Structures Group, Vienna University of Technology,
Favoritenstraße 9–11/186, A–1040 Vienna, Austria
{kodydek,schoenhacker}@ads.tuwien.ac.at
http://www.ads.tuwien.ac.at/

Abstract. We present two applications of a word analysis system for the German language: pre-hyphenation of documents in various formats, and counting the syllables of all words of a document. The *Si3Trenn* preprocessor provides pre-hyphenation for file formats allowing for soft hyphens (currently: plain text, LaTeX, RTF). It applies reliable, sense-conveying hyphenation (SiSiSi) to each word of the input text and inserts *soft hyphens* directly into the text. The resulting document can be processed as usual; soft hyphens will be used for hyphenation at the end of lines if appropriate. The *Si3Silb* syllable counter is a helpful tool for the statistical analysis of texts, e.g. in readability studies.

1 Introduction

German words require special treatment in text processing systems because of the possibility to form compound words as a combination of single words. We introduce tools for pre-hyphenation and syllable counting based on a word analysis system for German words. A recursive decomposition algorithm, following the rules for word flexion, derivation, and compound generation in the German language, splits compound words into single words, which are further decomposed into their smallest relevant parts (*atoms*, roughly corresponding to *morphems*).

Our pre-hyphenation and syllable counting applications are both up to date with the current ("new" since 1998) German orthography. In addition, the pre-hyphenation program can be set to process documents according to the previous ("old") orthography, which continues to be valid concurrently until 2005.

In the subsequent section we will give a short description of the hyphenation method *SiSiSi* (the German acronym for "*Si*chere *Si*nnentsprechende *Si*lbentrennung"). In Section 3 we will describe the tool for pre-hyphenation and present some practical applications and results. The description and some practical applications of the novel syllable counter follow in Section 4. Finally, we conclude by describing the possibilities and plans to extend the application towards HTML documents and platform independency in Section 5.

2 Word Analysis System

The SiSiSi method provides reliable and sense-conveying hyphenation of German words [1], see also [5]. The algorithm is necessarily based on *word analysis* because the classic pattern method, which is e.g. used in the TEX typesetting system [4, 7], does not apply well to German words. Unlike English, the German language allows for the construction of compound words of (almost) arbitrary length, which must be hyphenated according to the original single words. The pattern method usually causes problems at single word borders, and a similar problem applies to the simple rule-based standard hyphenation method which hyphenates according to the sequence of consonants and vowels, but completely ignores single word borders.

In its current version, the word analyzer uses a detailed classification of atoms, as well as a sizeable set of rules for word synthesis. This allows for very detailed word analysis and the elimination of many "nonsensical" words which would be accepted by a less detailed grammatical analyzer, see [11]. A recursive decomposition algorithm performs the word analysis, inserting hyphens at word and prefix borders. Possible additional hyphens in components consisting of stem(s) and suffix(es) or in polysyllabic prefixes are introduced by the simple rule-based algorithm.

If an input word can be decomposed into different sequences of atoms, and if these variants result in different sets of hyphenation points (e.g. *Wach=stu-be* vs. *Wachs=tu-be*), the algorithm takes no risks: Only hyphens which exist in all variants are considered to be "safe" and can be used in the final result (i.e. *Wachstu-be*); any other hyphens are "unsafe". Furthermore, hyphenation points at component borders ("=") are specially marked by the decomposition algorithm such that they can be prioritized at a later stage, see Section 3.

The syllable counter is based on the results of the hyphenation algorithm. It is also safe in the sense that it marks words for which the number of syllables cannot be determined accurately due to ambiguous results of the word analysis stage. Both tools have been thoroughly tested within a larger test system [9].

In principle, SiSiSi's decomposition algorithm is suitable for any language allowing for word composition or derivation. However, the language specific atom table, grammar and hyphenation rules need to be developed from scratch.

3 Pre-Hyphenation

The German hyphenation components of several popular text processing systems are based on patterns, simple rules or dictionaries, which can lead to serious hyphenation errors in compound words. Some other applications do not include any hyphenation algorithms. However, most applications allow for the use of soft hyphens. Pre-hyphenation by means of soft hyphens is a way to provide correct hyphenation points without directly interfering with text processors or changing file formats. Depending on font and display sizes, any word of a document could potentially require end-of-line hyphenation. Thus, in order to be inclusive, soft hyphens need to be inserted for all possible hyphenation points in a file.

We have developed a pre-hyphenator which can currently be used to process plain text, LaTeX, and RTF (Rich Text Format) documents. The most recent application uses SiSiSi to provide pre-hyphenated documents for a small PDA (Personal Digital Assistant) screen where hyphenation is essential to making efficient use of the limited display area even when the font size is changed, and where similar line lengths considerably improve a document's readability.

3.1 Features of the Si3Trenn Implementation

Si3Trenn is a 32-bit Windows application. Fig. 1 shows the process for producing a pre-hyphenated document *text.hyph.** from an input file *text.** using Si3Trenn. The resulting document can then be processed as usual, e.g. by converting it to a PDA reader format, or by compiling it using LaTeX. The files *text.log* and *text.mem* contain additional information about the pre-hyphenation process.

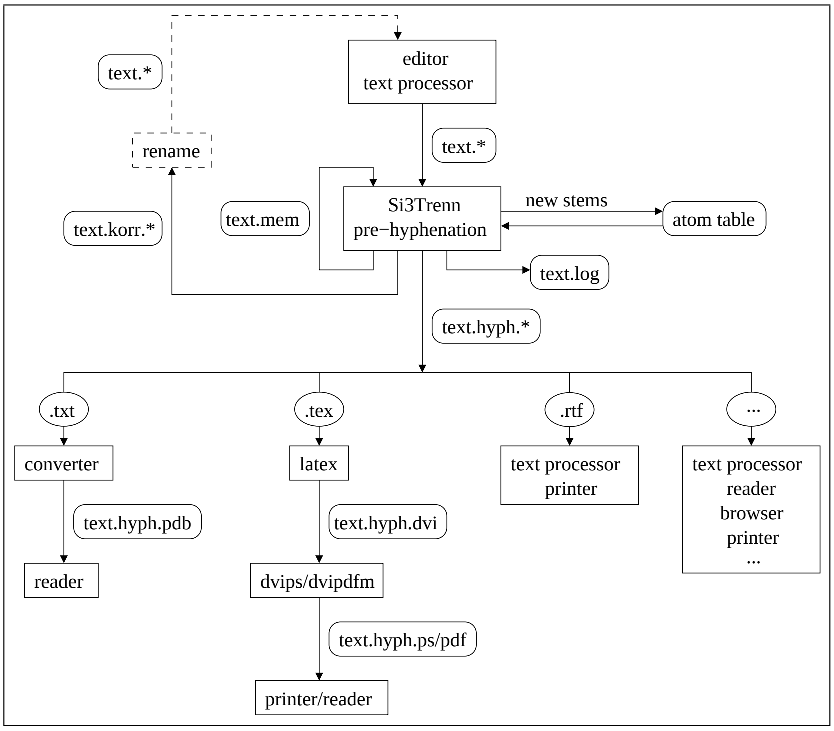
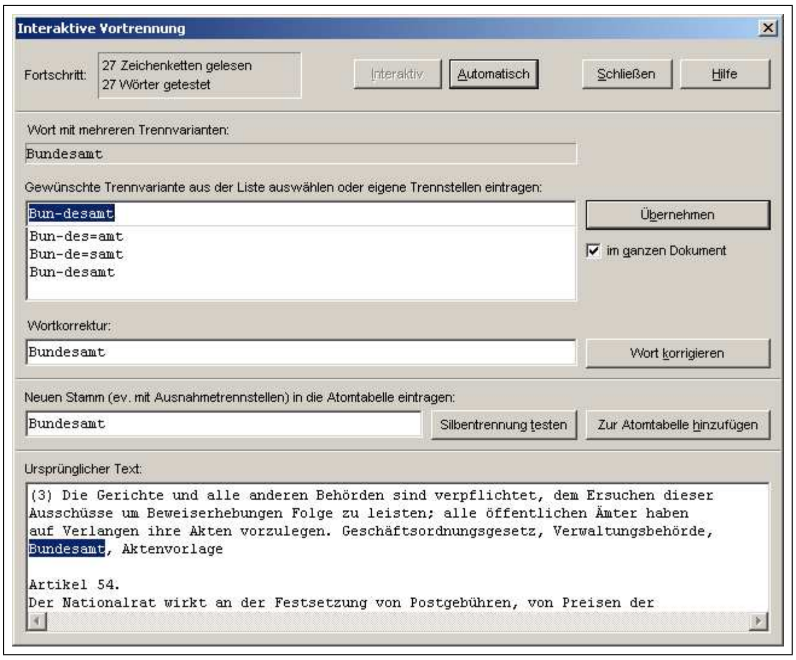


Fig. 1. Pre-hyphenation process for document text.* in different file formats (plain text .txt, LaTeX .tex, Rich Text Format .rtf, . . .) using the Si3Trenn tool (see text for a more detailed description).

During the pre-hyphenation process, which is exhaustively described in [2], the program may stop and request additional information from the user to re-



Fortschritt: 27 Zeichenketten gelesen
27 Wörter getestet

Wort mit mehreren Trennvarianten:
Bundesamt

Gewünschte Trennvariante aus der Liste auswählen oder eigene Trennstellen eintragen:
Bun-desamt

```
Bun-des=amt
Bun-de=samt
Bun-desamt
```

☑ im ganzen Dokument

Wortkorrektur:
Bundesamt

Neuen Stamm (ev. mit Ausnahmetrennstellen) in die Atomtabelle eintragen:
Bundesamt

Ursprünglicher Text:

```
(3) Die Gerichte und alle anderen Behörden sind verpflichtet, dem Ersuchen dieser
Ausschüsse um Beweiserhebungen Folge zu leisten; alle öffentlichen Ämter haben
auf Verlangen ihre Akten vorzulegen. Geschäftsordnungsgesetz, Verwaltungsbehörde,
Bundesamt, Aktenvorlage

Artikel 54.
Der Nationalrat wirkt an der Festsetzung von Postgebühren, von Preisen der
```

Fig. 2. Screenshot of the interactive pre-hyphenation dialog: progress (*Fortschritt*), ambiguous word (*Wort. . .*) and proposed hyphenation variants (*Gewünschte Trennvariante. . .*), word correction (*Wortkorrektur*), new stem (*Neuen Stamm. . .*), current context in original text (*Ursprünglicher Text*), and respective control buttons.

solve ambiguities or clarify unknown words, see Fig. 2. The user can (1) select from a list of proposed hyphenation variants, (2) provide the correct hyphenation for unknown words (e.g. proper names or foreign words), (3) correct a misspelled word (these corrections are stored in the file *text.korr.**, which should subsequently replace the original input file), or (4) add a new stem to the atom table to make a word recognizable to the system. User input according to (1) and (2) is stored in *text.mem* for automatic use in possible subsequent runs.

As mentioned above, hyphenation points are marked according to their priorities. The parts of a compound word are separated using *major hyphenation points* (“=”). *Minor hyphenation points* (“–”) are used within single words. Some minor hyphenation points are marked by “_” because their use is not desirable. The word *Wort=zer_le-gungs=ver_fah-ren* (*word decomposition method*) gives a good example for the use of these different types of hyphens.

An extensive start-up dialog provides a variety of parameters for the pre-hyphenation process, in particular the soft hyphen format and the level of interactivity. The user may also decide whether all possible hyphenation points are to be used. For books, it is preferable to use only major hyphenation points in compound words. When typesetting in narrow columns, all hyphenation points should be included.

3.2 Practical Applications of Si3Trenn and Experimental Results

Experimental results are given for various types of text in these formats (tests were conducted on a Pentium 4, 1.4 GHz, 512 MB, using Windows 2000 Pro):

1. lit (.txt): a plain-text literary document ("Die Verwandlung"[1] by F. Kafka)
2. leg (.txt): a plain-text legal document (the Austrian Constitution[2])
3. tech (.tex): a scientific text in the field of computer science in LaTeX[6]
4. econ (.rtf): a scientific text in the field of economy in RTF [8]

	lit (.txt)		leg (.txt)		tech (.tex)		econ (.rtf)	
total	19 148	100%	30 796	100%	25 743	100%	23 621	100%
short	7 111	37.14%	10 466	33.98%	7 690	29.87%	7 884	33.38%
unamb	11 062	57.77%	17 798	57.79%	16 410	63.75%	13 130	55.59%
amb_1	339	1.77%	1 374	4.46%	818	3.18%	1 130	4.78%
(uniq)	(219)		(409)		(272)		(336)	
amb_2	228	1.19%	940	3.05%	460	1.79%	449	1.90%
(uniq)	(162)		(366)		(203)		(231)	
unknw	408	2.13%	218	0.71%	365	1.42%	1 028	4.35%
(uniq)	(53)		(65)		(91)		(379)	
w/sec	1 835.0		857.3		1 093.4		604.5	

Fig. 3. Results of pre-hyphenation of selected documents (see text): number of analyzed words (total), words with <4 letters (short), unambiguous words (unamb), ambiguous words without/with unsafe hyphens (amb_1/amb_2), unknown words (unknw), respective numbers of unique occurrences (uniq), respective percentages, and words per second (w/sec).

Fig. 3 shows that about one third of the words in all document types are too short to be considered for hyphenation, while most remaining words are recognized and can be hyphenated without any ambiguity. Note that 1. (lit) tends to use shorter words, which results in a lower percentage of ambiguities. The use of proper names and specialized scientific terms leads to an increased rate of unknown words which is most noticeable in 1. (lit) and 4. (econ).

As expected, Fig. 4 indicates a reduction in the number of target format pages for all types of documents, while only requiring a moderate amount of additional space in the pre-hyphenated target files. Fig. 5 shows how pre-hyphenation can substantially improve the readability of a document on a limited display area.

4 Counting Syllables

In order to analyze the content and readability of various texts such as business reports [3], *readability formulas* are applied. They arithmetically compile statistical data about words and sentences into key figures. Many of those metrics,

[1] http://www.gutenberg2000.de/kafka/verwandl/verwa001.htm

[2] http://www.parlament.gv.at/pd/gesetze/b-vg/

	lit (.txt → .pdb)			leg (.txt → .pdb)		
	unhyph	hyph	Δ [%]	unhyph	hyph	Δ [%]
source [Bytes]	121 404	145 522	+19.87	268 912	348 656	+29.65
target [Bytes]	63 969	69 345	+8.40	119 396	131 875	+10.45
pages (PDA)	278	268	-3.60	719	672	-5.15
pages (Win)	79	77	-2.53	201	194	-3.48

	tech (.tex → .ps)			econ (.rtf → .doc)		
	unhyph	hyph	Δ [%]	unhyph	hyph	Δ [%]
source [Bytes]	269 727	345 423	+28.06	1 635 657	1 688 483	+3.23
target [Bytes]	1 837 366	1 835 699	-0.001	978 944	1 024 512	+4.65
pages (A4)	113	112	-0.88	99	97	-2.02

Fig. 4. Effects of applying pre-hyphenation to selected documents: size of the respective source and target files and number of pages in target format (PDA: PalmReader for PDA, Win: PalmReader for Windows in default window size, A4: print output on A4 size paper), in unhyphenated (unhyph) and fully pre-hyphenated (hyph) versions, and relative changes (Δ) with respect to unhyphenated versions.

such as the Flesch or Fog index, rely explicitly or implicitly on the average number of syllables per word. Si3Silb provides a convenient way of determining this value for German texts.

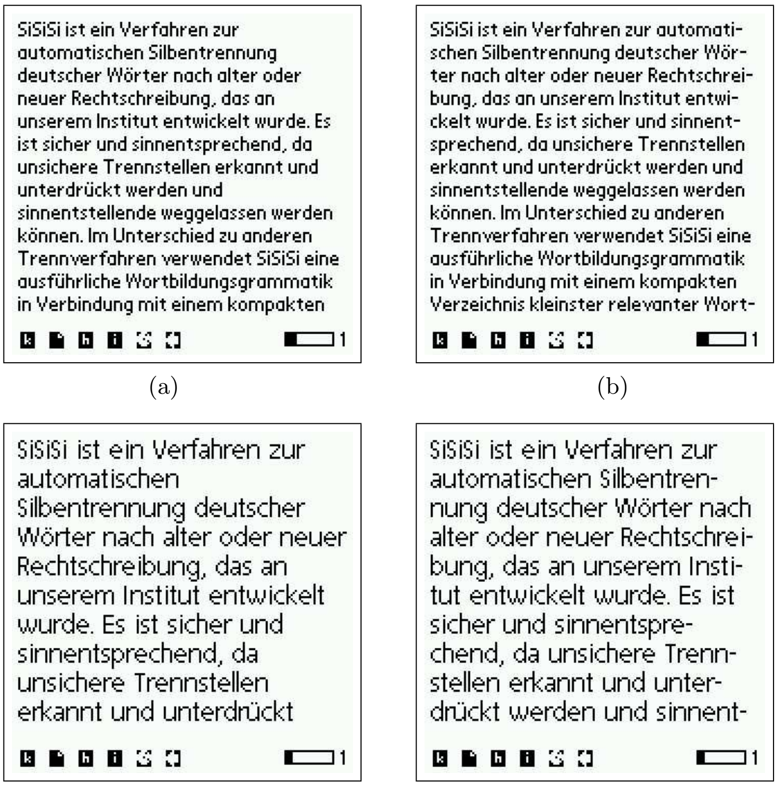
4.1 Features of the Si3Silb Implementation

Si3Silb is a 32-bit Windows application. It accepts a plain-text input file and generates a list of all words along with the respective syllable count. If the analysis fails because one of a word's atoms is unknown, a syllable count of 0 is returned. In the (very rare) case that two different compositions of one word lead to a differing number of syllables, the count is given as -1. Single-letter words are ignored.

4.2 Practical Applications of Si3Silb and Experimental Results

Si3Silb has been applied in a readability analysis of German business reports [8] and is currently being used for German documents in a statistical comparison of poetry in different languages.

The table in Fig. 6 summarizes the results of syllable counting experiments for different types of text (documents 1. and 2. as listed in Section 3.2). Note that the legal text contains an average of 2.4 syllables per word, as opposed to just 1.7 in the literary sample. This explains the significantly lower word processing rate for the legal text, as longer words potentially generate more variants during analysis. The number of unknown words is slightly higher than in Fig. 3 because two- and three-letter words are included.



SiSiSi ist ein Verfahren zur
automatischen Silbentrennung
deutscher Wörter nach alter oder
neuer Rechtschreibung, das an
unserem Institut entwickelt wurde. Es
ist sicher und sinnentsprechend, da
unsichere Trennstellen erkannt und
unterdrückt werden und
sinnentstellende weggelassen werden
können. Im Unterschied zu anderen
Trennverfahren verwendet SiSiSi eine
ausführliche Wortbildungsgrammatik
in Verbindung mit einem kompakten

(a)

SiSiSi ist ein Verfahren zur automati-
schen Silbentrennung deutscher Wör-
ter nach alter oder neuer Rechtschrei-
bung, das an unserem Institut entwi-
ckelt wurde. Es ist sicher und sinnent-
sprechend, da unsichere Trennstellen
erkannt und unterdrückt werden und
sinnentstellende weggelassen werden
können. Im Unterschied zu anderen
Trennverfahren verwendet SiSiSi eine
ausführliche Wortbildungsgrammatik
in Verbindung mit einem kompakten
Verzeichnis kleinster relevanter Wort-

(b)

SiSiSi ist ein Verfahren zur
automatischen
Silbentrennung deutscher
Wörter nach alter oder neuer
Rechtschreibung, das an
unserem Institut entwickelt
wurde. Es ist sicher und
sinnentsprechend, da
unsichere Trennstellen
erkannt und unterdrückt

(c)

SiSiSi ist ein Verfahren zur
automatischen Silbentren-
nung deutscher Wörter nach
alter oder neuer Rechtschrei-
bung, das an unserem Insti-
tut entwickelt wurde. Es ist
sicher und sinnentspre-
chend, da unsichere Trenn-
stellen erkannt und unter-
drückt werden und sinnent-

(d)

Fig. 5. The same text as displayed by PalmReader on a PDA screen using the small font (a) without and (b) with pre-hyphenation, and the large font (c) without and (d) with pre-hyphenation, respectively. Note that (b) and (d) show the identical pre-hyphenated file, but different soft hyphens are being used for actual end-of-line hyphenation.

	lit (.txt)		leg (.txt)	
total	19 148	100%	30 825	100%
unamb	18 490	96.56%	29 748	96.51%
amb	6	0.03%	25	0.08%
unknw	652	3.41%	1 052	3.41%

	lit (.txt)	leg (.txt)
let/w	5.13	7.23
syl/w	1.69	2.40
w/sec	1 805.6	890.4

Fig. 6. Results of syllable counting: number of analyzed words excluding single-letter words (total), words with an unambiguous/ambiguous (unamb/amb) number of syllables, unknown words (unknw), average number of letters per word (let/w), average number of syllables per word (syl/w), and words per second (w/sec).

5 Conclusion and Future Work

Pre-hyphenation can generally be applied to documents in any format providing for the use of soft hyphens. Si3Trenn provides fast and convenient pre-hyphenation for German documents prepared for the type setting system LaTeX, the text processor *Microsoft Word* (using the RTF format), and PDA readers such as *PalmReader*. Planned additions include a parser for the HTML file format. Currently, the pre-hyphenator is being reimplemented in Java to provide a platform independent solution.

Our experiments on large files show that the use of our pre-hyphenation program considerably improves the quality of typesetting and the on-screen readability of various types of documents, while only requiring a moderate amount of additional space in the pre-hyphenated files.

The syllable counter Si3Silb has shown to be a fast and practical tool for the statistical analysis of German texts, e.g. in readability studies.

References

1. Barth, W., Nirschl, H.: Sichere sinnentsprechende Silbentrennung für die deutsche Sprache. Angewandte Informatik 4 (1985) 152–159
2. Barth, W.: Ein schönes Schriftbild erzeugen mit der Sicheren Sinnentsprechenden Silbentrennung SiSiSi. Institute of Computer Graphics and Algorithms, Vienna University of Technology (2002) `http://www.ads.tuwien.ac.at/research/SiSiSi/Si3Anleitung.hyph.pdf`
3. Jones, M.J. and Shoemaker, P.A.: Accounting Narratives: A Review of Empirical Studies of Content and Readability. In: Journal of Accounting Literature, Vol. 13 (1994) 142–184
4. Knuth, D.E.: TeX: The Program. Addison Wesley, Reading, Massachusetts (1986)
5. Kodydek, G.: A Word Analysis System for German Hyphenation, Full Text Search, and Spell Checking, with Regard to the Latest Reform of German Orthography. In: Proceedings of TSD 2000, LNAI 1902, Springer-Verlag, Berlin (2000) 39–44
6. Kodydek, G.: Automatische Wortanalyse für die deutsche Sprache. PhD thesis, Vienna University (2001)
7. Liang F.M.: Word Hy-phen-a-tion by Com-put-er. PhD thesis, Department of Computer Science, Stanford University (1983)
8. Schittko L.: Der Geschäftsbericht als Kommunikationsinstrument – Eine empirische Analyse der Kapitalmarktreaktion auf die Semiotik des Geschäftsberichts. Diploma thesis, University of Essen (2001)
9. Schönhacker M., Kodydek G.: Testing a Word Analysis System for Reliable and Sense-Conveying Hyphenation and Other Applications. In: Proceedings of TSD 2000, LNAI 1902, Springer-Verlag, Berlin (2000) 127–132
10. Sojka P.: Notes on Compound Word Hyphenation in TeX. In: TUGboat, Vol. 16, No. 3 (1995) 290–296
11. Steiner, H., Barth, W.: Sichere sinnentsprechende Silbentrennung mit Berücksichtigung der deutschen Wortbildungsgrammatik. Tagungsband Konvens'94, ed. H. Trost, Vienna (1994) 330–340