Automatic non-standard hyphenation in
OpenOffice.org

LÁSZLÓ NÉMETH
nemeth (at) openoffice dot org

Abstract
The hyphenation algorithm of OpenOffice.org 2.0.2 is a generalization of \TeX’s hyphenation algorithm that allows automatic non-standard hyphenation by competing standard and non-standard hyphenation patterns. With the suggested integration of linguistic tools for compound decomposition and word sense disambiguation, this algorithm would be able to do also more precise non-standard and standard hyphenation for several languages.

Introduction
Standard hyphenation consists of splitting a word and including a hyphen at the end of the first part of the split word (unless the word already contained a hyphen or n-dash at the break). While standard hyphenation is widely applicable, several languages also use non-standard hyphenation.

Table 1 shows examples for non-standard hyphenation: character deletions and other changes at hyphenation break points in European writing systems. Some non-standard hyphenation can be handled easily by computer, like the mandatory middle dot deletion from Catalan digraph \textipa{ll}. But complex analysis is necessary for languages, like German, Hungarian, Swedish and Norwegian to recognize hyphenation points. For instance, the Swedish word form \textipa{glassko} has three different meanings, and can be hyphenated as \textipa{glas- sko} (glass shoe), \textipa{glass- ko} (ice cream cow) and in the non-standard way, \textipa{glass- sko} (ice cream shoe).

Such non-standard hyphenation plays an important role in good typesetting. Commercial DTP programs, even word processors, support automatic non-standard hyphenation, often by licensing third party libraries. The most important free alternatives, such as Apache FOP, GNU Tröff, KDE KOffice, OpenOffice.org, Scribus, and \TeX and its variants, do not support automatic non-standard hyphenation. \TeX has a hyphenation primitive, the \texttt{discretionary} command. There are \TeX macros in the Babel package for non-standard hyphenation, for instance, \texttt{\_1gem} for Catalan \textipa{H}, \texttt{\_ck} or \texttt{"ck} for German, \texttt{\_esz} for Hungarian, \texttt{\_w} for Polish, but there is no real automatic non-standard hyphenation in \TeX. Omega 2 has promising developments towards implementing sophisticated automatic non-standard hyphenation for German and other languages [4, 5].

The aim of the present project was to implement language-independent automatic non-standard hyphenation in OpenOffice.org. In this article we present our results, introduce old and new hyphenation algorithms, extension of the Hungarian hyphenation patterns and finally show the possibility of integrating compound word decomposition and word sense disambiguation to our algorithm.

Results
\TeX’s hyphenation is the de facto standard in the free software world, because the hyphenators of the free programs mentioned in the previous section are all based on Liang’s hyphenation algorithm from \TeX82 [9], and use the \TeX hyphenation patterns. Thus, we have developed an extension for OpenOffice.org’s ALT Linux LibHnj hyphenator to do automatic non-standard hyphenation. The result is based on a generalization of Liang’s original algorithm which also allows easy integration of special linguistic tools to handle compound word decomposition and word sense disambiguation in automatic hyphenation. The Hungarian hyphenation patterns were extended with non-standard hyphenation patterns.

The improved hyphenation library (without integrated linguistic tools) is part of the OpenOffice.org 2.0.2 with the extended Hungarian hyphenation patterns. Developers can download a standalone version

\footnote{German orthography before the spelling reform of 1996.}
<table>
<thead>
<tr>
<th>Language</th>
<th>Example</th>
<th>Hyphenation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalan</td>
<td>paraklel</td>
<td>paral-l el</td>
<td>digraph (l) represents long (geminated) (l)</td>
</tr>
<tr>
<td>Dutch</td>
<td>reëel</td>
<td>re- eel</td>
<td>diaeresis and hyphenation sign syllable breaks</td>
</tr>
<tr>
<td>English</td>
<td>eighteen</td>
<td>eight-teen</td>
<td>vowel lengthening with diminutive (-tje)</td>
</tr>
<tr>
<td>German</td>
<td>Zucker</td>
<td>Zücker</td>
<td>digraphs (ck) and (kk) represent long (k)</td>
</tr>
<tr>
<td>Greek</td>
<td>Μαίου</td>
<td>Μα- ίου</td>
<td>triple consonants at compound word boundary</td>
</tr>
<tr>
<td>Hungarian</td>
<td>asszonnyal</td>
<td>asz- szony- nyal</td>
<td>diaeresis and hyphenation sign syllable breaks</td>
</tr>
<tr>
<td>Norwegian</td>
<td>bussjåfør</td>
<td>buss- sjåfør</td>
<td>simplified double digraphs (long (sz) and (ny) phonemes)</td>
</tr>
<tr>
<td>Polish</td>
<td>kong-fu</td>
<td>kong- -fu</td>
<td>repeated hyphen at line begin</td>
</tr>
<tr>
<td>Swedish</td>
<td>tillåta</td>
<td>till- låta</td>
<td>triple consonants at compound word boundary</td>
</tr>
</tbody>
</table>

Table 1: Non-standard hyphenation in European languages

\[\text{algorithm}.\]

\[
\text{4l1g4}\\
\text{lgo3}\\
\text{lgo}\\
\text{2i th}\\
\text{4h1m}
\]

\[\text{--------------------}\]

\[\text{4 1 4 3 2 0 4 1}\\
\text{a l-g o-r i t-h-m}\]

Figure 1: \TeX\ hyphenation of ‘algorithm’

of this library with an example executable from the Linguocomponent project home [14].

Liang’s hyphenation algorithm

Franklin M. Liang’s hyphenation algorithm is based on competing hyphenation patterns. The patterns can give excellent compression for a hyphenation dictionary, and using these patterns the fast hyphenator algorithm can also correctly hyphenate unknown (non-dictionary) words most of the time. Liang’s work covers also the machine learning of the hyphenation patterns and exceptions by PatGen pattern generator.

The hyphenation patterns can allow and prohibit hyphenation breaks on multiple levels. Figure 1 shows the pattern matching on the word ‘algorithm’. The \TeX\ English hyphenation patterns \(4l1g4, lgo3, lgo, 2i th\) and \(4h1m\) match this word and determine its hyphenation. Only odd numbers mean hyphenation breaks. If two (or more) patterns have numbers in the same place, the highest number wins. The \text{alg-o-rith-m} hyphenation is bad, but the last one-letter hyphenation is suppressed by \TeX, so we end up with the correct \text{al-go-rithm}.

One of the most notable features of this pattern-based hyphenation is the human-readable format of the knowledge database, in contrast to an equivalent finite state machine or a similarly good artificial neural network. This format is good for manual checking and corrections.

Missing features

In \TeX’s automatic hyphenation the most wanted features are non-standard hyphenation, compound word analysis, word sense disambiguation and taboo word filtering [12, 13].

Sojka’s non-standard hyphenation extension

In [12] Petr Sojka suggests a non-standard hyphenation extension for Liang’s algorithm. His algorithm first searches all hyphenation points of a word using Liang’s algorithm, and then matches patterns from a non-standard hyphenation table at valid hyphenation points, replaces the matching pattern with a special character, and rechecks the hyphenation of the new word at this special character with Liang’s algorithm. The non-standard hyphenation point will be chosen if the second hyphenation is successful. Using a \(ck \rightarrow %\) (\(k\- k\)) pattern data from a non-standard hyphenation table, the German word \text{Zucker} will be \text{Zu%er} after

\[\text{2Liang’s hyphenation algorithm and its compact implementation using packed trie data structure was perfect twenty-five years ago for English and for computers with less than a few MB RAM. Nowadays internationalization (handling multiple languages) is a standard in software industry and free software development. Modern personal computers have much more memory and speed to enable using additional special linguistic tools in hyphenation.}\]
the replacement, and the pattern zu%1er permits non-
standard hyphenation with k k (Zuk- ker).

Problems
It’s possible to use the pattern generator on a pre-
pared input dictionary for Sojka’s algorithm, but then
we lose the human-readable format of hyphenation
patterns. The biggest problem is to use competing
patterns on multiple levels. That is why instead of us-
ing difficult redundant patterns with special hyphen-
ation characters, Sojka suggests global parameters (left
and right non-standard hyphenation penalties) to for-
bid standard hyphenations near the non-standard hy-
phenation points. But German, Hungarian, Norwe-
gian and Swedish non-standard hyphenation need true
competing patterns.

OpenOffice.org’s extension
To keep the flexibility of Liang’s algorithm, Open-
Office.org augments the original hyphenation patterns
with extended patterns defining non-standard hyphen-
ation points as subregions and replacements of the sub-
regions. To keep the clear syntax, a non-standard hy-
phenation pattern is denoted as a plain hyphenation
pattern and a record separated by a slash.

For example, the pattern zuc1ker/k=k,3,2 re-
presents the hyphenation of Zucker. This means the
non-standard hyphenation subregion will be replaced
by k=k, where the = indicates the break point with
a hyphen. The subregion begins at the 3rd character,
and contains 2 characters (ck).

Table 2 shows possible hyphenation patterns for
Table 1. The dots in the patterns match the word
boundaries. The first dot doesn’t affect the character
positions in the non-standard hyphenation subranges:
. zuc1ker/k=k,3,2. Figure 2 shows the result of ap-
plying multiple non-standard pattern matching.

Rules
A single subregion must contain exactly one hyphen-
ation point (indicated by an odd number in Liang’s
syntax). There may also be explicit non-breakable
points (indicated by even numbers) in the subregion,
and any breakable or non-breakable points out of the
subregion.

A standard and a non-standard hyphenation pat-
tern matching the same hyphenation point must not
be on the same hyphenation level. For instance,
c1 and zuc1ker/k=k,3,2 are invalid, while c1 and
zuc3ker/k=k,3,2 are valid extended hyphenation
patterns.

Unicode character encoding
Unicode is the basis for internationalization. Thanks
to the unambiguous start positions of the multibyte-
characters, Liang’s algorithm works perfectly with the
UTF-8 Unicode encoding. Subregion parameters of
non-standard hyphenation patterns use Unicode char-
acter (not byte) positions and lengths.

Changing hyphen
Missing or alternative hyphenation marks are handled
by using underline characters instead of equal signs in
our non-standard hyphenation patterns, where under-
line character indicates only the break point, without
an implied hyphen. For example, using the underline
with an explicit hyphen, k- k and k=k are equivalent

3Not only for exotic writing systems. Affix-rich languages can
differentiate between characters in a word. For exam-
example, Neoimir (about Nexø in Hungarian) contains special characters
from Latin-1 and Latin-2 character tables.
Automatic non-standard hyphenation in OpenOffice.org

Table 3: Hungarian hyphenation examples with ambiguous ggy and ssz patterns

<table>
<thead>
<tr>
<th>Example</th>
<th>Hyphenation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>meggy</td>
<td>meggy</td>
<td>noun with long phoneme gy (sour cherry)</td>
</tr>
<tr>
<td>meggyez</td>
<td>meggy- gyez</td>
<td>derived verb (to do something with sour cherry)</td>
</tr>
<tr>
<td>meggyúz</td>
<td>meggy- íz</td>
<td>compound (sour cherry jam)</td>
</tr>
<tr>
<td>meggyőz</td>
<td>meg- győz</td>
<td>verb prefix meg- + verb győz (persuade)</td>
</tr>
<tr>
<td>esszék</td>
<td>esz- szék</td>
<td>long phoneme sz (essay)</td>
</tr>
<tr>
<td>Jameszék</td>
<td>James- szé</td>
<td>noun James + suffix -szé (to become) James</td>
</tr>
<tr>
<td>sameszék</td>
<td>samesz- szé</td>
<td>noun samesz + suffix -szé (to become) verger</td>
</tr>
<tr>
<td>vizeszék</td>
<td>vizes- szék</td>
<td>compound (special chair (szék) in Hungarian folklore)</td>
</tr>
<tr>
<td>rekekessék</td>
<td>rekesz- szék</td>
<td>verb rekeszt + suffix -jék (third-person plural obstruct!)</td>
</tr>
<tr>
<td>kirekekessék</td>
<td>kirekesz- szék</td>
<td>prefix ki + verb rekeszt + suffix -jék (third-person plural exclude!)</td>
</tr>
</tbody>
</table>

Extending Hungarian hyphenation patterns

The Hungarian language uses simplified forms to represent its double digraph and trigraph consonants (sz+sz→ssz, dzs+dzs→ddzs, etc.), but hyphenation undoes the simplification (sz- sz, dzs- dzs). Some ambiguity results from this non-standard hyphenation in Hungarian, caused by rich compounding and affixation, see Table 3.

Manual extension of the Huhyphn Hungarian hyphenation patterns based on Hungarian vocabularies and morphology has been accomplished, and the result contains over two thousand non-standard hyphenation patterns. For example, Figure 3 shows the competing patterns matching the word esszé (essay).

The Huhyphn distribution consists over 63 thousand hyphenation patterns generated from a 2.5 million word hyphenation dictionary by PatGen [10].

Linguistic tools for better hyphenation

Pattern-based hyphenation doesn't work well on languages with an unlimited number of compound words [7]. Compound word decomposition by patterns results in an enormous number of hyphenation patterns in the Huhyphn distribution. However, within a few minutes, an expert could be able to find a dozen badly hyphenated compound words in Magyar webkorpusz, a Hungarian gigaword corpus with 21 million word forms. We need more sophisticated compound word decomposition methods, like SISiSi [1, 7, 8]. OpenOffice.org’s Hunspell spell checker also has morphological analyzer capability to decompose compound words. We suggest a simple method and formalism to integrate these tools with the pattern-based hyphenation algorithm cannot decompose compounds from three or more dictionary words.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Non-standard hyphenation of esszé}
\end{figure}
Figure 4: Hyphenation by decomposition

Dots within patterns

Dots denote word boundaries in Liang’s algorithm. Extending this formalism, let us also allow dots to denote the word boundaries within compounds. The compound word decomposition makes only a boundary annotation with dots, and we can hyphenate the decomposed word by dotted hyphenation patterns.

For instance, the Swedish word glassko would be glas.sko or glass.ko after compound word decomposition, and can be hyphenated with the pattern .7 as in Figure 4.

Double dots

We denote non-standard compounding by double dots, as in glas.sko. This annotated word can then be hyphenated with a non-standard hyphenation pattern, such as s.8.9s/ss=s,1,4 in our example.

The annotation is removed from the output of the hyphenation algorithm, as in the three possible annotated and hyphenated forms of glassko in Figure 4. With a suitable word sense disambiguation, the pattern based hyphenator is given exactly one of them. (Without word sense disambiguation, glassko is not annotated and hyphenated).

Conclusion

The new version of OpenOffice.org contains state-of-the-art Hungarian hyphenation, solving the problem of automatic non-standard hyphenation in a generalized way. The extended version of Liang’s hyphenation algorithm is suitable for other languages. With the suggested formalism and minimal extension, the algorithm can also be integrated with sophisticated linguistic tools to handle compound word decomposition and word sense disambiguation in automatic hyphenation.

Acknowledgments

OpenOffice.org’s improved hyphenation was demonstrated at the 3rd Conference on Hungarian Computational Linguistics in 2005, Szeged, Hungary, thanks to Media Education and Research Center (MOKK), University of Technology and Economics, Budapest. I would like to thank Stefan Baltzer and other developers at Sun Microsystems for testing and quick integration of the extended ALT Linux LibHnj library into OpenOffice.org. Péter Szabó for his helpful comments and Yannis Haralambous for his information about Dutch and Greek non-standard hyphenation. Also thanks to Marcin Miłkowski and the authors of various Internet sources [2, 3, 11] for the Polish and other examples. Special thanks for the editorial work of TUGboat.

References


[8] Gabriele Kodydek and Martin Schönhacker. Si3trewn and Si3Silb: Using the SiSiSi word analysis system for pre-hyphenation and syllable counting in German documents.


