Hyphenation in TEX — Quo Vadis?

Petr Sojka
Faculty of Informatics
Masaryk University
Burešova 20, 602 00 Brno
Czech Republic
Internet: sojka@muni.cz

Pavel Ševeček
Faculty of Informatics
Masaryk University
Burešova 20, 602 00 Brno
Czech Republic
Internet: pavel@muni.cz

Abstract

Significant progress has been made in the hyphenation ability of TEX since its first version in 1978. However, in practice, we still face problems in many languages such as Czech, German, Swedish etc. when trying to adopt local typesetting industry standards.

In this paper we discuss problems of hyphenation in multilingual documents in general, we show how we've made Czech and Slovak hyphenation patterns and we describe our results achieved using the program PATGEN for hyphenation pattern generation. We show that hyphenation of compound words may be partially solved even within the scope of TEX82. We discuss possible enhancements of the process of hyphenation pattern generation and describe features that might be reasonable to think about to be incorporated in Ω or another successor to TEX82.

Motivation

"Go forth and make masterpieces of hyphenation patterns ..."
(Haralambous, 1994)

Editors' and publishers' typographical requirements for camera-ready prepared documents are growing. To meet some of their requirements in TEX, especially when typesetting in narrow columns, one needs perfect hyphenation patterns in order to find almost all permissible hyphenation points.

When making Czech hyphenation patterns and typesetting multilingual documents we encountered some problems with achieving quality hyphenation and decent-looking documents with TEX. This work has led to our ideas about possible remedies and future extensions in a successor to TEX.

Our paper consists of three parts. In the first part we try to summarize the developments that have been made on the issue since TEX's birth.

In the second, we describe our attempts to create Czech and Slovak hyphenation patterns and summarize hints and suggestions for PATGEN users.

In the third part we discuss possible improvements that might take place in a TEX successor (Ω, ε-TEX or New Typesetting System ($\mathcal{NTS}$)).

1 The hyphenation story

Let's review the developments in hyphenation in TEX that have been made so far.

1.1 English

In TEX78 a rule-driven algorithm for English was built-in by Liang and Knuth. Their algorithm found 40 % of the allowable hyphens, with about 1 % error (Liang, 1981). Although authors claimed that these results are "quite good", Liang continued working on the generalization of the idea of rules expressed by hyphenating and inhibiting patterns. In his dissertation (Liang, 1983) he describes a method, which is used in TEX82, based on the generalization of the prefix, suffix and the vowel-consonant-consonant-vowel rules. He wrote (in WEB) the program PATGEN (Liang & Breitenlohner, 1991) to automate the process of pattern generation from a set of already hyphenated words. He started with the 1966 edition of Webster's Pocket Dictionary that

included hyphenated words and inflections (about 50 000 entries in total). In the early stages, testing the algorithm on a 115 000 word dictionary from the publisher, 10 000 errors in words not occurring in the pocket dictionary were found. "Most of these were specialized technical terms that we decided not to worry about, but a few hundred were embarrasing enough that we decided to add them to the word list." (Liang, 1983, p. 30). He reports the following figures: 89,3 % permissible hyphens found in the *input* word-list with 4447 patterns with 14 exceptions.

Liang's method is described by Knuth (1986*b*, Appendix H) and was later adopted in many programs such as `troff` (Emerson & Paulsell, 1987) and `Lout`, and in localizations of today's WYSIWYG DTP systems such as QuarkXPress, Ventura, etc. Although specialized dictionaries such as (Allen, 1990) by Oxford University Press separate possible word-division points into at least two categories (preferred and less recommended), we have not seen any program that incorporates the possibility of taking into account these classes of hyphenation points so far.

1.2 Those other languages

> *"... patterns are supposed to be prepared by experts who are paid well for their expertise."*
> *(Knuth, 1986b, p. 453, 8^{th} printing)*

The first version of TEX82 allowed only one set of patterns to be loaded at a time. Thus it was not possible to typeset multilingual documents with correct hyphenation in all languages and this limitation was quite unsatisfactory. Already in 1985, two attempts to solve the problem were made:

Multilingual TEX: Extensions, most of which afterwards Knuth adopted in TEX 3.x were suggested and implemented by Ferguson (1985). A new primitive `\language`[1] was introduced for switching between several sets of `\patterns` and hyphenation exceptions. A new `\charsubdef` primitive is no longer necessary in today's 8-bit TEX. Full details can be found in (Ferguson, 1988).

ISITEX: Barth & Nirschl (1985) presented an approach on achieving decent hyphenation in German texts under the name SITEX, or in its interactive version under the name ISITEX. Their method, (available as a change file for UNIXTEX from `eiunix.tuwien.ac.at`) has been used in Germany for years and is being improved (Barth & Steiner, 1992; Barth et al., 1993). This approach has been proposed for inclusion in $\mathcal{NTS}$ (NTS, 1992–).

SITEX (ISITEX for the interactive version) introduces a new primitive `\nebenpenalty` which allows differentiation between main (compound word boundaries) and secondary (word stem) hyphenation points.

A new notation for hyphenation patterns is introduced and a hyphenation algorithm for German is hardwired into the program. The tables for the algorithm, file `sihyphen.tex` (60K) are written manually and can be simply edited and enriched. However, no provision for the generation of these patterns from a word-list (such as the `PATGEN` program) is offered.

During the last 15 years almost every year there appeared a paper in *TUGboat* reporting new patterns for some language (see table 1). Another couple of hyphenation patterns, fonts and preprocessors are available in ScholarTEX[2] (Haralambous, 1991).

Although Don Knuth introduced the new primitives `\language` and `\setlanguage` for switching between several sets of hyphenation patterns in TEX 3.0, there are indications that not all of the related problems have been solved and further investigations are necessary (Fanton, 1991).

Proposals on how to customize TEX for a new language were suggested by Partl (1990). New tools to simplify the generation of 8-bit (virtual) fonts were designed — `fontinst` (Jeffrey, 1993) and `accents` (Zlatuška, 1991). A macro package for simple language switching `babel` (Braams, 1991*b*; Braams, 1991*a*; Braams, 1993) was produced to simplify typesetting of multilingual documents. An international version of the `Makeindex` program was written (Schrod, 1991). The DC fonts (Ferguson, 1990; Haralambous, 1992*a*; Haralambous, 1993*a*), designed to permit hyphenation in many languages, are now being widely distributed, forced by the new LATEX wave. Compliance with the suggestions of the working group TWGMLC[3] (Haralambous, 1992*a*) could help too (naming conventions for hyphenation files, etc.). Multilingual document aspects of typesetting are being collected in the scope of LATEX3 project in (Gaulle, 1994), where a nice collection of language-related TEX primitives can be found, together with definitions of the terminology used.

[1] A rather misleading name, as it deals with only one particular feature of a language — hyphenation — which feature is of only limited interests to linguists.

[2] ScholarTEX is a registered trademark of Yannis Haralambous

[3] TEXnical Working Group on Multiple Language Coordination

Table 1: Hyphenation patterns for TEX with PATGEN statistics for various languages

language	trie	ops	done by	#patt	size	author (& reference)
BG (Bulgarian)	688	56	hand	263	1672	Ognyan Tonev/90
CA (Catalan)	661	11	hand	826	6136	Goncal Badenes, Francina Turon/91
CY (Welsh)	8552	143	PATGEN	6728	43162	Yannis Haralambous, (Haralambous, 1993*b*)
CZ_1 (Czech)	3676	90	hand	4479	25710	Ladislav Lhotka/91, (Lhotka, 1991)
CZ_2	5302	67	PATGEN	4196	23474	Pavel Ševeček/94, (Sojka & Ševeček, 1994)
DE_{min} (German)	6099	170	PATGEN	4066	25660	Norbert Schwarz/88
DE_{max}	9980	255	PATGEN	7007	45720	Norbert Schwarz/88
DE (v3.1)	8375	207	PATGEN	5719	39251	Norbert Schwarz, Bernd Raichle/94, (Schulze, 1984; Partl, 1988; Breitenlohner, 1988; Obermiller, 1991; Kopka, 1991)
DK (Danish)	1815	60	PATGEN	1145	6411	Frank Jensen/92
EL (Mod. Greek)	1278	23	hand	1616	8786	Yannis Haralambous/92
EO (Esperanto)	4895	143	PATGEN	4118	23224	Derk Ederveen/93
ES (Spanish)	1106	29	hand	578	4609	Francesc Carmona/93
ET (Estonian)	2054	45	PATGEN	1267	7976	Enn Saar/92
FI (Finnish)	583	27	hand	232	1342	Kauko Saarinen/92, (Saarinen, 1988)
FR (French)	1634	83	comb.	917	7240	Jacques Désarménien/92, (Jacques Désarménien, 1984)
Ancient Greek			hand			Yannis Haralambous/92, (Haralambous, 1992*b*)
HR (Croatian)	1471	46	hand	916	7250	Cvetana Krstev/93
HY (Armenian)						Yannis Haralambous (in ScholarTEX)
IS (Icelandic)	5477	145	PATGEN	4187	29919	Jorgen Pind/87
IT (Italian)	1327	15	hand	729	4255	Salvatore Filippone/92, (Canzii et al., 1984)
IT (Italian)	529	37	hand	210	2571	Claudio Beccari/93, (Beccari, 1992)
Latin			hand			Yannis Haralambous/92, (Haralambous, 1992*b*)
Modern Latin			hand			Claudio Beccari/92, (Beccari, 1992)
LT (Lithuanian)	2169	77	PATGEN	1546	9639	Vitautas Statulevicius & Yannis Haralambous/92
NL_1 (Dutch)	7824	124	PATGEN	6105	37997	CELEX/89
NL_2	10338	187	PATGEN	7928	50969	CELEX/89
NL_3	520	24	hand	326	1732	Peter Vanroose
NO (Norwegian)	3669	220	PATGEN	2371	15589	Ivar Aavatsmark/92
PL (Polish)	4954	194	hand	4053	28907	Hanna Kołodziejska/94, (Kołodziejska, 1987; Kołodziejska, 1988)
PT (Portuguese)	374	10	hand	126	534	Pedro J. de Rezende, (de Rezende, 1987)
RU (Russian)	4599	92	hand	4121	29272	Dimitri Vulis, (Vulis, 1989; Malyshev et al., 1991*a*; Malyshev et al., 1991*b*; Samarin & Urvantsev, 1991)
SK (Slovak)	3600	248	hand	2569	22628	Jana Chlebíkova/92
SK	7606	78	PATGEN	6137	35623	Pavel Ševeček/94, (Sojka & Ševeček, 1994)
SR (Serbian)	1475	40	hand	896	6890	Cvetana Krstev/89, (Krstev, 1991)
SV (Swedish)	5269	125	PATGEN	3733	23821	Jan Michael Rynning/91
TR (Turkish)	678	16	hand	1834	9580	Pierre A. MacKay/88, (MacKay, 1988)
UK (UK English)	10995	224	PATGEN	8527	54769	Dominik Wujastyk/93
US (US English)	6075	181	PATGEN	4447	27302	Frank Liang/82, (Liang, 1983)
US	6661	229	PATGEN	4810	30141	G.D.C. Kuiken/90, (Kuiken, 1990)

1.3 Exception logs

"If any computer center decides to preload different exceptions from those in plain TEX (i.e., in the file HYPHEN.TEX), the changed exceptions should not under any circumstances be put into HYPHEN.TEX or PLAIN.TEX. All local changes should go into a separate file, so that TEX will still produce identical results on all machines. In fact, I recommend not preloading those changes, but rather assuming that individual users will have their own favorite collection of updates to the standard format files."
(Knuth, 1983)

The exception log and corrections for US English hyphenation have been reported several times – (e.g. Thulin, 1987; Beeton, 1989; Kuiken, 1990; Beeton, 1992), as shown in table 2. These listings are published in accordance with DEK's wish in (Knuth, 1983). Only words with *wrongly* placed hyphenation points are listed, not those where TEX finds only a subset of possible breakpoints.

Table 2: Growing number of exceptions for `hyphen.tex`

# of exceptions	where reported
14	(Liang, 1983)
24	(Beeton, 1984, *TUGboat* 5, no. 1)
88	(Beeton, 1985, *TUGboat* 6, no. 3)
127	(Beeton, 1986, *TUGboat* 7, no. 3)
129	(Thulin, 1987, *TUGboat* 8, no. 1)
501	(Beeton, 1989, *TUGboat* 10, no. 3)
543	(Beeton, 1992, *TUGboat* 13, no. 4)

This shows that significant care and effort is still needed and *is being gradually spent* on the checking of hyphenation points during proof-reading and that the standard US patterns are not sufficient to satisfy current needs. Additional sets of patterns (2 versions – `ushyphen.add` and `ushyphen.max`) have been generated by Kuiken (1990) to cover the exceptions by additional patterns and these add-on files are available on CTAN and other hosts, e.g., `ftp.cs.umb.edu`. *But*, after having added one of these files at the end of the `\patterns` command in `hyphen.tex`, in order to overcome huge exception lists that should be loaded with every document, one loses the compatibility between different installations and acts against Knuth's wishes.

1.4 The need to re-generate US English patterns

```
! TeX capacity exceeded, sorry
  [exception dictionary=307.]
```
DEK

So, to follow Knuth's rules, every document should start with loading the exception file – for this, one has to increase TEX82ś exception size (in words) from 307 to at least 607 (as is now usual in UNIXTEX, emTEX and other installations). However, this is barely sufficient for the current English exception file (remember one has to add words in all possible inflexions) but for flexive languages (such as Czech, where from one stem there are about 20 different suffices) it is unusable.

Maybe it is time to re-generate the patterns from a bigger (say, 200 000 entry) word-list once again from scratch?[4] Imagine the day when you will know that TEX will find 99.99 % of hyphens contained in your copy of Webster, so you will not have to go through a list of exceptions and a couple of dictionaries to check hyphenation points in your document! For backward compatibility one has to save every document together with the patterns and exceptions used anyway.[5]

2 Making Czech and Slovak hyphenation patterns with PATGEN

"A program should do one thing, and do it well."
Ken Thompson

The first Czech patterns were made in 1988 by Novák using `PATGEN` from a list of 170 000 word forms. Because of errors in his word-list, and only partially optimized `PATGEN` parameter settings, the patterns were good but not perfect.

The patterns weren't publicly available, so a second attempt was done by hand by Lhotka (1991) just as MacKay (1988) did for Turkish. Because of lots of exceptions to the 'rules', their usage was not quite comfortable either.

As Novák's list of words had been lately made public, we started compiling a bigger word-list from

[4] Otherwise in 2050 there will have to be an extra issue of *TUGboat* devoted to the publication of exceptions to `hyphen.tex`.

[5] A search on `CTAN` via `quote site index` command shows 5 files of *different* lengths with the name `hyphen.tex`. (And Knuth and Liang's `hyphen.tex` can be found there under four different names – `hyphen.tex`, `ushyph1.tex`, `ushyphen.std`, `ushyphen.tex` – which leads to the total confusion!)

various sources using the old patterns for bootstrapping. We've learned a lot from the experience described by Rynning (1991) and Haralambous (1993*b*) and in a tutorial (Haralambous, 1994).

2.1 Czech hyphenation rules

Czech hyphenation rules are described in (Zdeněk Hlavsa et al, 1993, p. 56–57) and in a special book (Haller, 1956) where a list of exceptions was published. Briefly, we have syllable hyphenation with 'etymological' exceptions. Hyphenation is preferred between a prefix and the stem, and on the boundary of compound words. Things become complicated when:

1. The word evolved in such a way that although historically it was built from a prefix plus the stem of another word, today it is perceived as a new word stem. As an example may serve the word `ro-zu-mět` – "to understand" (syllable division) against `roz-u-mět` (`roz` is the prefix and `umět` means "to know").
2. There is no agreement on word hyphenation – e.g., the *current* rules for word `sestra` – "sister" allow one to hyphenate `se-stra`, `ses-tra` and `sest-ra`.
3. Word stem hyphenation points change when a suffix is added – e.g., `hrad` – "castle" can't be hyphenated, but with a suffix could – `hra-du`.
4. Compound words e.g. `tři-a-třiceti-letý` – "33 years old" are taken into account. Czech has a lot of compound words, but not to the extent that German has.
5. The hyphenation of a word depends on the semantics: `nar-val` and `na-rval`.

These rules make it hard to create patterns that describe all these exceptions and exceptions to exceptions. As we had handy a word-list with lists of allowable prefixes and suffixes, together with preliminary patterns to hyphenate word stems for bootstrapping, we decided to generate a hyphenated list of Czech words for `PATGEN`.

2.2 Stratified sampling

> *"A large body of information can be comprehended reasonably well by studying more or less random portions of the data. The technical term for this approach is stratified sampling."*
> *(Knuth, 1991, p. 3)*

Czech is a very flexive language; on average 20–30 inflexions can be derived from one word stem by changing the suffix added and one can multiply it almost twice, as negation can be created from many words (adjectives, verbs) by prefixing `ne`. Thus from a 170 000 stem word-list about 5 000 000 inflexions may be generated. Generating patterns from such a list would be very impractical. Because the suffixes are often the same or similar, we generated a word-list by means of the following rules:

1. We add only every 7^{th} (actually 17^{th} worked as well) derived word form from the full list to the `PATGEN` input list, with exceptions that:
2. every stem must be accompanied by at least one derived form, and
3. every derived form with overlapping prefixes has to be present in the `PATGEN` input list as well, and
4. only one word with prefixes `ne` (by which one can create negation to almost every word) and `nej` (by which one creates superlatives) is included, and
5. the hand-made list of exceptions from (Haller, 1956) (about 10 000 words) and other sources are always included.

With this procedure we have 372 562 Czech words to work with `PATGEN`. We used the same approach for Slovak. The results are in table 3.

Table 3: `PATGEN` statistics for the Czech and Slovak languages

# of words	# of hyphenation points		
	Correct	Wrong	Missed
Czech			
372562	1019686 (98.26 %)	39 (0.01 %)	18086 (1.74 %)
Slovak			
333139	1025450 (98.53 %)	34 (0.01 %)	15273 (1.47 %)

Samples of `PATGEN` statistics are presented in tables 4, 5 and 6. These tables show that by twiddling with `PATGEN` parameters one may generate various versions of patterns. Usually the size of patterns and % of bad hyphenations are the minimization criteria, but maximization of % of good (found) hyphenations and other strategies might be chosen.

2.3 Compound words

> *"Hints for hyphenation are most often needed at the word boundaries of compound words."*
> *(Saarinen, 1988, p. 191)*

As an experiment we took our (rather huge) word-list of Czech words in which there was marked hyphenation only on prefix and compound word boundaries.

Table 4: Standard Czech hyphenation with Liang's parameters for English

level	length	param	% correct	% wrong	# patterns	size
1	2–3	1 2 20	96.95	14.97	+ 855	
2	3–4	2 1 8	94.33	0.47	+1706	
3	4–5	1 4 7	98.28	0.56	+1033	
4	5–6	3 2 1	98.22	0.01	+2028	32 kB

Table 5: Standard Czech hyphenation with improved (size optimized) strategy (cf. table 3)

level	length	param	% correct	% wrong	# patterns	size
1	1–3	1 2 20	97.41	23.23	+ 605	
2	2–4	2 1 8	85.98	0.31	+ 904	
3	3–5	1 4 7	98.40	0.78	+1267	
4	4–6	3 2 1	98.26	0.01	+1665	23 kB

Table 6: Standard Czech hyphenation with improved (% of correct optimized) strategy

level	length	param	% correct	% wrong	# patterns	size
1	1–3	1 5 1	95.43	6.84	+2261	
2	1–3	1 5 1	95.84	1.17	+1051	
3	2–5	1 3 1	99.69	1.24	+3255	
4	2–5	1 3 1	99.63	0.09	+1672	40 kB

Table 7: Czech hyphenation of composed words with Liang's parameters but allowing 1-length patterns in level 1

level	length	param	% correct	% wrong	# patterns	size
1	1–3	1 2 20	72.97	14.32	+ 300	
2	2–4	2 1 8	69.32	3.09	+ 450	
3	3–5	1 4 7	84.09	4.02	+ 870	
4	4–6	3 2 1	82.61	0.33	+2625	25 kB

Table 8: Czech hyphenation of composed words with slightly modified parameters (% of correct slightly optimized)

level	length	param	% correct	% wrong	# patterns	size
1	1–3	1 2 20	72.97	14.32	+ 300	
2	2–4	2 1 8	69.32	3.09	+ 450	
3	3–5	1 4 3	90.82	4.24	+3014	
4	4–6	3 2 1	89.07	0.36	+2770	40 kB

Table 9: Czech hyphenation of composed words with another parameters (% of correct optimized, but % of wrong and size increased)

level	length	param	% correct	% wrong	# patterns	size
1	1–3	1 5 1	64.35	5.34	+1415	
2	2–4	1 5 1	67.10	1.88	+1261	
3	3–5	1 3 1	97.94	5.39	+8239	
4	4–6	1 3 1	97.91	1.14	+2882	84 kB

The PATGEN program was able to produce hyphenation patterns for this list succesfully. The number of patterns was rather large, but feasible (25–84 kB, depending on parameters). From a 380 698 item word-list the patterns found 307 470 of the hyphenation points[6] correctly, 5 040 points were hyphenated wrongly (exceptions), and 4 680 hyphenation points were missing.

To test the possibility of creating patterns for compound words in detail, we generated a word-list of more than 100 000 words with 101 687 hyphenation points marked. The list included both compound words and simple ones too.

The results of some of the runs are shown in tables 7, 8 and 9.

2.4 Some other numbers

Just for fun we've tried patterns for different languages on our Czech PATGEN input word-list—see table 10. There are interesting speculations about these numbers—e.g., trying Slovak patterns on the Czech word-list, one finds more than 90 % of hyphenation points. On the contrary, probably because of non-syllabic principles and different rules for pronunciation, UK English rules are totally different — only 19 % of Czech words are hyphenated correctly by UK patterns. Surprisingly, Swedish, Finnish and Dutch (NE_3) patterns make fewer wrong hyphenations than the Czech old hyphenation patterns. The difference between Dutch patterns made by hand (NE_3) based on the syllabic principle) and those made by PATGEN (NE_1, NE_2) may by caused by the fact that general syllable hyphenation is relatively good for languages in which the hyphenation is based on syllabic principles. Having hyphenated word lists of different languages, it might be interesting to measure the 'syllabic principles of hyphenation' of different languages on an universal syllable hyphenation.

As hyphenation in most languages is based on syllabic principles, it is worth trying to create universal syllabic hyphenation and only learn the difference (exceptions) from this universal hyphenation. Let's try to summarize what we think that should be done in the future.

[6] Some of these points might be wrong, as the database we used is only preliminary. Due to our experience with the standard hyphenation list, after correction of errors (wrongly marked hyphenation points, typos) PATGEN can generalize *substantially* better and the size of the list of patterns is reduced significantly.

Table 10: Patgen-like statistics for using various language patterns on Czech hyphenated word-list

Language	Correct	Wrong	Missed
CZ (Sev)	98.26 %	0.01 %	1.74 %
NE_3	57.38 %	4.11 %	42.62 %
SV	57.10 %	5.32 %	42.90 %
FI	52.67 %	5.40 %	47.32 %
CZ (Lho)	93.39 %	5.89 %	6.61 %
SK	90.77 %	7.28 %	9.23 %
US	31.84 %	9.58 %	68.16 %
IT	49.27 %	9.88 %	50.73 %
NO	51.61 %	11.32 %	48.39 %
FR	59.07 %	11.54 %	40.93 %
NE_1	59.14 %	11.59 %	41.86 %
NE_2	58.80 %	11.99 %	41.20 %
UK	18.84 %	12.19 %	81.16 %
DE_{min}	58.62 %	12.50 %	41.38 %
DE_{max}	58.56 %	12.70 %	41.44 %
PL*	69.00 %	12.96 %	31.00 %
PL	68.06 %	13.12 %	31.94 %
DE (v.3.1)	58.84 %	13.86 %	41.16 %

* with transformed patterns — accented letters substituted by non-accented ones

3 Future

"I hope TEX82 will remain stable
at least until I finish Volume 7
of The Art of Computer Programming."
(Knuth, 1989, p. 625)

3.1 Possible extensions in a successor to TEX

"Good typography therefore is a silent art;
not its presence but rather
its absence is noticeable"
(Mittelbach & Rowley, 1992b)

It seems feasible to incorporate either SITEX (Barth et al., 1993) changes or separate compound word hyphenation patterns in ε-TEX.

These experiments, discussed in section 2.3 show that, even with the current TEX, only doubling the patterns for a language with compounds might allow, e.g., switching between standard hyphenation in narrow columns and compound-word-only hyphenation in wide columns.

With a simple change in the program, one may achieve additional flexibility in hyphenation:

New registers \leftcompoundhyphenmin and \rightcompoundhyphenmin may be helpful for filtering unneeded hyphenation near compound word borders and \compoundwordhyphenpenalty might set a penalty (usually much lower than

\hyphenpenalty) for breaks on compound word boundaries. In this case \compoundwordchar character (i.e., the compound work mark in the DC fonts) could be *automatically* inserted there to prevent ligatures going over a compound word boundary.

Another minor addition may be added too, e.g., ε-TEX: already in MLTEX there was implemented a flag \dischyph indicating whether or not to hyphenate words with discretionaries (i.e. embedded hyphena) or not. As an example may serve citation (AMS, 1993) in this paper, where we had to insert discretionaries by hand in the compound word "Author-Prepared" to achieve the limits on underfull boxes set by the editor. With setting \dischyph=1 this wouldn't be necessary.

3.2 Pattern generalization

Apart from PATGEN extensions according to character clustering, which are orthogonal, we are thinking of the following generalization. Currently, there are only 2 classes of inter-letter state: an odd or even number that carries information whether to hyphenate or not. The natural generalization would be to have n classes. Inter-letter numbers in patterns would code these classes in such a way that number m between letters will mean that this position belongs to the class number $m \pmod n$ (when numbering classes from 0). The case $n = 2$ is the current situation, so \pattern[2] might mean classical Liang's patterns. Another class might be prefix boundary, compound word boundary or whatever else might possibly be useful for the hyphenation algorithm to be aware of the word (discretionary being another possibility).

An application for English is straightforward too. Our approach will allow one to distingush "preferred" and "less recommended" classes of hyphenation points as published in (Allen, 1990).

In German, one may make other classes (and patterns), e.g. classes for different discretionary breaks.

3.3 Possible extensions in a successor to TEX.

"Please correct if you have a hyphenated word at the bottom of a right-hand page."
(AMS, 1993)

A possible direction was shown by Plaice (1993) and in (Haralambous & Plaice, 1994; Plaice, 1994). With suggested clustering of letters and enriched PATGEN (Liang & Breitenlohner, 1991) one could achieve context-dependent discretionaries and thus solve the c-k → k-k-like problems in German.

Taylor (1992, p. 249) mentions a possible definition of \brokenpenalty = \ifrecto 500\else 200\fi. If the output routine could communicate with the parameter-breaking algorithm, word breaks crossing page boundaries could be eliminated.

Conclusions

"Therefore it still is not the right moment to manufacture TEX on a chip."
(Knuth, 1989, p. 641)

In our survey we presented an overview on the topic of hyphenation in TEX and our results based on experience with Czech and Slovak. We conclude that the current possibilities of TEX are far from perfect and might be improved either in the scope of TEX82 (creation of better hyphenation patterns for various languages by PATGEN), ε-TEX (e.g. duplication of hyphenation mechanism for compound words), or Ω or $\mathcal{NTS}$ (special capabilities for context-dependent discretionaries).

Acknowledgement

The presentation of this work has been made possible due to the support of Czech Grant Agency (grant Nr. 201/93/1269). We would like to thank our referee, Yannis Haralambous, Libor Škarvada and Jiří Zlatuška for comments and useful suggestions on how to improve this paper.

References

Allen, R. E. (1990), *The Oxford Spelling Dictionary*, Vol. II of *The Oxford Library of English Usage*, Oxford University Press.

AMS (1993), 'AMS–Instructions for Author-Prepared Books'.

Barth, W. & Nirschl, H. (1985), Implementierung eines Verfahrens für die Silbentrennung, Technical Report Bericht Nr. 26, Institut für Praktische Informatik.

Barth, W. & Steiner, H. (1992), 'Deutsche Silbentrennung für TEX 3.1 (German hyphenation for TEX 3.1)', *Die TEXnische Komödie*. Journal of DANTE (Deutschsprachige Anwendervereinigung TEX e.V.); Group of German-speaking TEX Users..

Barth, W., Steiner, H. & Herbeck, H. (1993), 'ISITEX Interaktive Silbentrennung für die deutsche Sprache untex TEX 3.14 und 3.141 unter UNIX (Interactive hyphenation for German and TEX 3.14 and 3.141 under UNIX)', electronic documentation of ISITEX from eiunix.tuwien.ac.at.

Beccari, C. (1992), 'Computer Aided Hyphenation for Italian and Modern Latin', *TUGBoat* **13**(1), 23–33.

Beeton, B. (1984), 'Hyphenation exception log', *TUGBoat* **5**(1), 15.

Beeton, B. (1985), 'Hyphenation exception log', *TUGBoat* **6**(3), 121.

Beeton, B. (1986), 'Hyphenation exception log', *TUGBoat* **7**(3), 145–146.

Beeton, B. (1989), 'Hyphenation exception log', *TUGBoat* **10**(3), 336–341.

Beeton, B. (1992), 'Hyphenation exception log', *TUGBoat* **13**(4), 452–457.

Braams, J. (1991*a*), 'Babel, a multilingual style-option system for use with LATEX's standard document styles', *TUGBoat* **12**(2), 291–301.

Braams, J. (1991*b*), 'Babel, a multilingual style-option system', *Cahiers GUTenberg* **10-11**, 71–72.

Braams, J. (1993), 'An update on the babel system', *TUGBoat* **14**(1), 60–62.

Breitenlohner, P. (1988), 'German TEX, a next step', *TUGBoat* **9**(2), 183–185.

Canzii, G., Genolini, F. & Lucarella, D. (1984), 'Hyphenation of Italian words', *TUGBoat* **5**(1), 14.

de Rezende, P. (1987), 'Portuguese hyphenation table for TEX', *TUGBoat* **8**(2), 102.

DUDEN (1991), *Duden Band 1 — Rechtschreibung der deutschen Sprache*, 20., neu bearbeitete und erweiterte Auflage edn, Dudenverlag.

Emerson, S. L. & Paulsell, K. (1987), *troff Typesetting for UNIX™ Systems*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Fanton, M. (1991), 'TEX: les limites du multilinguisme', *Cahiers GUTenberg* **10-11**, 73–80.

Ferguson, M. J. (1985), 'A multilingual TÊX', *TUGBoat* **6**(2), 57–58.

Ferguson, M. J. (1988), TEX is Multilingual, *in* Thiele (1988), pp. 179–189.

Ferguson, M. J. (1990), 'Fontes latines européennes et TEX 3.0', *Cahiers GUTenberg* **7**, 29–32.

Gaulle, B. (1994), 'Requirements in multilingual environments', in electronic form (version 1.02) on CTAN as file `vt15d02.tex`.

Goossens, M., ed. (1994), *Proceedings of the TEX Users Group* 15th *Annual Meeting, Santa Barbara, 1994*, TEX Users Group, Providence, U.S.A.

Haller, J. (1956), *Jak se dělí slova (How the words get hyphenated)*, SPN Praha.

Haralambous, Y. (1991), 'ScholarTEX', *Cahiers GUTenberg* **10-11**, 69–70.

Haralambous, Y. (1992*a*), 'TEX Conventions Concerning Languages', *TEX and TUG NEWS* **1**(4), 3–10.

Haralambous, Y. (1992*b*), 'Hyphenation patterns for ancient Greek and Latin', *TUGBoat* **13**(4), 457–469.

Haralambous, Y. (1993*a*), 'DC fonts — questions and answers', *TEX and TUG NEWS* **2**(1), 10–12.

Haralambous, Y. (1993*b*), 'Using `PATGEN` to Create Welsh Patterns', submitted to *TUGboat*.

Haralambous, Y. (1994), 'A small tutorial on the multilingual features of Patgen2', in electronic form, available from CTAN as `info/patgen2.tutorial`.

Haralambous, Y. & Plaice, J. (1994), First applications of Ω: Greek, Arabic, Khmer, Poetica, ISO 10646/UNICODE, etc., *in* Goossens (1994), pp. 256–264.

Jacques Désarménien (1984), 'How to run TEX in a French environment: Hyphenation, fonts, typography', *TUGBoat* **5**(2), 91.

Jeffrey, A. (1993), 'A Postscript font installation package written in TEX', *TUGBoat* **14**(3), 285–292.

Knuth, D. (1983), 'A note on hyphenation', *TUGBoat* **4**(2), 64.

Knuth, D. E. (1986*b*), *The TEXbook*, Vol. A of *Computers and Typesetting*, Addison-Wesley, Reading, MA, USA.

Knuth, D. E. (1986*a*), *TEX: The Program*, Vol. B of *Computers and Typesetting*, Addison-Wesley, Reading, MA, USA.

Knuth, D. E. (1988), The Errors of TEX, Technical Report STAN-CS-88-1223, Stanford University, Department of Computer Science.

Knuth, D. E. (1989), 'The Errors of TEX', *Software—Practice and Experience* **19**(7), 607–681. This is an updated version of (Knuth, 1988).

Knuth, D. E. (1991), *3:16 Bible texts illuminated*, A-R Editions, Inc.

Kołodziejska, H. (1987), Dzielenie wyrazów polskich w systemie TEX, Technical Report 165, Sprawozdania Instytutu Informatyki Uniwersytetu Warszawskiego.

Kołodziejska, H. (1988), 'Le traitement des textes polonais avec le logiciel TEX', *Cahiers GUTenberg* (0), 3–10.

Kopka, H. (1991), *LaTeX—Erweiterungsmöglichkeiten mit einer Einfürung in* METAFONT, second edn, Addison-Wesley Verlag, Bonn, Germany.

Krstev, C. (1991), 'Serbo-Croatian hyphenation: a TeX point of view', *TUGBoat* **12**(2), 215–223.

Kuiken, G. (1990), 'Additional Hyphenation Patterns', *TUGBoat* **11**(1), 24–25.

Lhotka, L. (1991), 'České dělení pro TeX (Czech hyphenation for TeX)', *CSTUG bulletin* (4), 8–9.

Liang, F. & Breitenlohner, P. (1991), 'PATtern GENeration program for the TeX82 hyphenator', electronic documentation of `PATGEN` program version 2.0 from UNIXTeX distribution at `ftp.cs.umb.edu`.

Liang, F. M. (1981), 'TeX and hyphenation', *TUGBoat* **2**(2), 19.

Liang, F. M. (1983), Word hy-phen-a-tion by com-pu-ter, Technical Report STAN-CS-83-977, Stanford University.

MacKay, P. A. (1988), 'Turkish Hyphenations for TeX', *TUGBoat* **9**(1), 12–14.

Malyshev, B., Samarin, A. & Vulis, D. (1991*a*), 'Russian TeX', *Cahiers GUTenberg* **10-11**, 1–6.

Malyshev, B., Samarin, A. & Vulis, D. (1991*b*), 'Russian TeX', *TUGBoat* **12**(2), 212–214.

Mittelbach, F. & Rowley, C. (1992*a*), The future of high quality typesetting: structure and design, *in* Zlatuška (1992), p. 255.

Mittelbach, F. & Rowley, C. (1992*b*), The pursuit of quality — How can automated typesetting achieve the highest standards of craft typography?, *in* C. Vanoirbeek & G. Coray, eds, 'Proceedings of the International Conference on Electronic Publishing, Document Manipulation & Typography, Lausanne, Switzerland, 1992', Cambridge University Press, New York, pp. 261–273.

NTS (1992–), 'New Typesetting System discussion list'. This is an electronic list devoted to discussions about TeX's successor. To subscribe, send a request with the text `subscribe nts-l` to `listserv@vm.urz.uni-heidelberg.de`.

Obermiller, W. (1991), 'TeX in Germany', *TUGBoat* **12**(2), 211–212.

Partl, H. (1988), 'German TeX', *TUGBoat* **9**(1), 70–72.

Partl, H. (1990), How to make TeX and LaTeX international, *in* J. Nadrchal, ed., 'Man-Machine Interface in the Scientific Environment. Proceedings of the 8th European Summer School on Computing Techniques in Physics. Skalský Dvůr, Czechoslovakia, 19–28 September 1989', Vol. 61 of *Computer Physics Communications*, European Summer Schools on Computing Techniques in Physics, North-Holland Publishing Company; Elsevier Science Publishers B. V., pp. 190–200.

Plaice, J. (1993), 'Language-Dependent Ligatures', *TUGBoat* **14**(3), 271–274.

Plaice, J. (1994), Progress in the Omega Project, *in* Goossens (1994), pp. 190–193.

Rynning, J. M. (1991), 'Swedish Hyphenation for TeX', received in electronic form from author via email `jmr@nada.kth.se`.

Saarinen, K. (1988), Experiences with TeX in Finland, *in* Thiele (1988), pp. 189–194.

Samarin, A. & Urvantsev, A. (1991), 'CyrTUG, le monde TeX en cyrillique', *Cahiers GUTenberg* **12**, 71–74.

Schrod, J. (1991), 'An International Version of MakeIndex', *Cahiers GUTenberg* **10-11**, 81–90.

Schulze, B. (1984), 'German hyphenation and *Umlaut*s in TeX', *TUGBoat* **5**(2), 103.

Sojka, P. & Ševeček, P. (1994), Hyphenation in TeX — Quo Vadis?, *in* W. Bzyl & T. Przechlewski, eds, 'Proceedings of the 9th European TeX Conference, Gdańsk, 1994', pp. 59–68

Taylor, P. (1992), The Future of TeX, *in* Zlatuška (1992), pp. 235–254.

Thiele, C., ed. (1988), *Proceedings of the TeX Users Group 9th Annual Meeting, Montréal, 1988*, TeX Users Group, Providence, U.S.A.

Thulin, A. (1987), 'More hyphenation exceptions', *TUGBoat* **8**(1), 76.

Vulis, D. (1989), 'Notes on Russian TeX', *TUGBoat* **10**(3), 332–336.

Zdeněk Hlavsa et al (1993), *Pravidla českého pravopisu (The rules of the Czech spelling)*, Academia Praha.

Zlatuška, J. (1991), 'Automatic generation of virtual fonts with accented letters for TeX', *Cahiers GUTenberg* **10-11**, 57–68.

Zlatuška, J., ed. (1992), *Proceedings of the 7th European TeX Conference, Prague, 1992*, Masarykova Universita Brno.