

COMPUTATIONAL MORPHOSYNTAX

Report on Research 1981—84

Edited by Fred KARLISSON

University of Helsinki
Department of General Linguistics
Hähtuskatu 11—13
SF-00100 HELSINKI 10
FINLAND

PUBLICATIONS
No. 13
1985

Fred Karlsson

AUTOMATIC HYPHENATION OF FINNISH

1. Introduction

In this paper, those structural aspects of Finnish graphemic structure will be discussed that are relevant for designing a high-quality, moderately sized, efficient and otherwise realistically implementable hyphenation algorithm. The meaning of the notions high quality, moderate size, efficiency, and implementability should be agreed upon. Here, only quality will be dealt with.

Optimally, a hyphenation algorithm should (a) be able to identify all potential hyphenation points (henceforth: HPS), and (b) do so errors. In accordance with standardized terminology, we talk about recall and precision. Recall defines the coverage of the model, precision its correctness rate.

Full recall means that all potential HPS of ordinary unedited running text were spotted. A potential HP is any syllable boundary that complies with certain morphological and typographical requirements. The former restrictions concern morphologically complex (compound or derived) words, especially those belonging to the following three groups. (Throughout this paper, erroneous hyphens are immediately prefixed by a star. Correct hyphens are indicated in the normal way.)

- compounds with a medial n or g immediately surrounded by vowels, e.g. känsäne*-dustaja, karsimil*-nealyje pro kansan-edustaja, karsimis-esitys.

- complex word-forms with a medial diphthong ending in -j-, e.g. isoil*-sä, viranom*-inen pro iso-ina, viennoma*-nen,

- compounds where a noninitial member starts with a consonant combination, e.g. Pikkuplanaetta, neekerlighetto pro pikku-planaetta, neekeri-ghetto.

Such HPs are unacceptable because one letter is dissociated from where it morphologically belongs, and because the hyphen runs counter to syllable structure.

The basic typographical requirement is that phonologically possible syllabifications such as a*-lue, i*-soisä, vältti*-o, korke*-ä are not potential HPs. Successful hyphenation must refer at least two letters to both lines (of which at least one must be a vowel, e.g. st*-ratostäkki is not allowed). Syllable boundaries of this type are not potential HPs and need not be included in recall assessments. Neither should they be suggested as HPs by a sophisticated algorithm.

The precision of a hyphenation algorithm making no errors is, of course, 100 %. There are many ways of estimating how close to perfection a given algorithm is. The strictest requirements are set when precision is determined in regard to unedited running text including i.a.:

- all inflectional and derivational forms of native words,
- compounds such as kansanedustaja, ydinenergia, voiklubi, pikkuplanaetta, kuikkustruuma, pieneläinklinikka, neekerlighetto,
- loan-words such as demonstratio, inflaatio, transplantaatio,
- foreign names such as Munsterheim, Whitbread, Zagortschik, München, Chotbzadeh,
- acronyms such as MSOY:llä, SKDL
- word forms containing non-letter characters such as killpa-ajo-ori, naimisissa/naimaton/jeski/eronnut

The most difficult hyphenation problems are encountered in compounds consisting of native and/or borrowed parts. These are not typical top items of frequency dictionaries but frequent enough in running text to pose a problem. Therefore, precision cannot be properly estimated on the basis of frequency dictio-

aries alone. Such estimates are bound to yield unrealistically good results.

Recall and precision are not independent of each other. By lowering recall standards in order to avoid certain types of errors it is possible to raise precision. But this is not optimal from the viewpoint of implementation realism. An intentionally lowered recall tends to leave fairly long letter sequences (e.g. longer than 7 letters) where no HP is spotted, cf. difficult instances such as astinosaisena, betoniseinämä, naisasiainainen. An hyphenation algorithm designed for use in real surroundings cannot leave such residues.

2. A survey of the rules

The notion "rule" here refers to any potential generalization concerning the locations of HPs. The basic Finnish hyphenation rule is the CV-rule (1).

- (1) There is an HP before any CV-sequence (i.e. sequence of consonant + vowel).

The recall of the CV-rule alone is 95,7 % and its precision 98,7 % when tested on a running text of 6,000 word-form tokens containing some 13,000 HPs. Plain CV-hyphenation thus works far better in Finnish than in e.g. English or Swedish. This is due to the high congruence between morphological and syllable boundaries in compounds. But disturbing errors at a rate of more than 1 % still remain, e.g. kansat*-nedustaja, limantia*-lallu, varap*-residentti. The lacking recall is manifested in vowel sequences such as maa-ilma, raa-oiisä, näyttely*-osasto.

The recall and precision of the CV-rule may be improved upon by postulating hyphenation rules for: (2) sequences of three or four vowels, (3) some sequences of two vowels, (4) some sequences of two or three consonants, (5) the consonants n and s in certain word-medial contexts, and (6) non-letter characters.

3. The CV-rule

The CV-rule (1) and its central typographical requirements were stated above. Some more pervasive restrictions on this rule should now be considered. First, proper CV-hyphenation is dependent upon the nature of an eventually preceding VV-string. After VV-strings, CV-hyphenation is possible only if the second vowel of VV is i. Thus, one can hyphenate ajankohdai-nen, liipäi-sen, väpöi-lustä, haltioi-tumaan with a very small error risk (isoi*-säälä). Without this restriction, numerous errors would be produced: läpiai*-jokielto, parioi*-velliä, olemassao*-loon, puoliu*-neen etc.

Second, the CV-rule must be made subordinate to certain special rules concerning medial CC-combinations in compounds. Otherwise, errors such as varap*-residentti, korkop*-rosentti, yök*-lubi, nakkist*-roganolf are bound to occur.

Decisions have to be made for how to hyphenate certain letter strings occurring only in loans. Are e.g. hyphenations such as Ric-hard, Manc-hester acceptable?

4. Rules for three- and four-vowel strings

In most VVV(V)-strings there is a unique HP unnoticed by the CV-rule. The vowel rules below are equally valid for simplex and compound words. These rules are presupposed not to apply to vowel strings where the loss of a k due to consonant gradation has been marked by an apostrophe (e.g. lää'an, poi'ille, liu'un), nor to vowel strings in compounds where the hiatus between identical vowels is hyphenated by another convention (e.g. etäma-a-lueen, maanantai-illa).

- (2a) If there are two identical adjacent vowels in a three- or four-vowel string, there is an HP between these and any neighbouring vowel.

The relevant types are (V = vowel):

V-aa	aa-V
V-äa	äa-V
V-ee	ee-V
V-öo	öö-V
V-oo	oo-V
V-ii	ii-V
V-yy	yy-V
V-uu	uu-V

The following are examples of the application of rule (2a):
 raa-oissa, raa-uitte, perjantai-aamu, sunnuntai-uisinta, raa-empi, alue-uutiset, radio-uutinen, maa-ikkii, ka-ottinen, mää-ynä, korke-aa, veike-ään, kull-aa, valti-oon, yhti-öön, mini-äänsä, aino-aa, halu-aan, pesy-een, käry-ää, näkö-ään, po-ettinen, kofe-ini, muse-oon, eläke-uudistus, puolu-essaa, di-reetti, heroo-inen, hero-inni, poro-uuni, maa-ilma, raa-unta, lili-an, lili-että, ruu-assa, puu-istutus, mää-llte, noo-an, pää-elinkeino.

The following rule (2b), which is nothing but a list of instances, completes the hyphenation of three- and four-vowel strings. The most frequently occurring strings have been pre-fixed by a plus sign.

- (2b) In the following three- and four-vowel strings the HP is where the hyphen indicates.

STRING	EXAMPLE
AI-OI	ai-oin
AI-UI	perjantai-uinti
AD-OI	tau-oissa
EU-OI	leu-oissa
IE-OI	lie-oissa
IE-UI	kie-uitte
OI-OI	loi-olimme
OI-UI	joi-uissa

UO-UI	huo-uimme
YÖ-YI	hyö-yissä
A-EI	la-eissa
A-OI	va-oissa
E-AI	halke-aisi
E-OI	hope-oida
E-ÄI	repe-äisi
E-ÖI	lipe-öidä
+ I-AI	kuti-aisi
I-AU	farmari-auto
I-EU	keski-eurooppalainen
+ I-OI	asi-olta
+ I-ÄI	selvi-äisi
+ I-ÖI	yhti-öitä
+ O-AI	koho-aisi
+ U-AI	halu-aisi
+ U-EI	puolu-eissa
U-OI	salppu-olta
Y-EI	pesy-eitä
Y-ÄI	käry-äisi
Ä-YI	nä-yissä
Ä-ÖI	nä-öissä
AI-A	tai-an
AI-E	perjantai-esitys
AI-O	ai-on
AI-U	kai-un
AU-A	kau-an
AU-E	hau-en
EU-A	leu-an
IE-A	lie-an
IE-O	lie-on
IE-U	kie-utte
IE-Y	tie-yhdistys
IO-A	valio-akku
IO-E	valtio-elin
IO-E	yhtiö-esitys

IÖ-Ä	kiu-as
IU-A	liu-eta
IU-E	liu-ote
IU-O	joi-ata
OI-A	oi-eta
OI-E	oi-omme
OI-O	joi-un
OI-U	lou-elta
OU-E	lou-on
OU-O	puolue-ansiot
UE-A	hui-eta
UI-E	ruo-an
UO-A	ruo-issa
UO-I	huo-utte
UO-U	
YE-Ä	työ-eläke
YÖ-E	säi-että
ÄI-E	

The list contains some strings where the HP location is, in fact, ambiguous: EAI, EÄI, IAI, IAU, IOI, IÄI, IÖI, UAI, UEI, YEI. Most of these, in particular IOI and IÖI, are so frequent that recall considerations argue for hyphenating according to the (optimized) proposal. Of course, this will leave a small residue of errors, e.g. hope-äistykä, gemmaati-öllmiö, keitti-öikku, puolu-ellitamat.

The following ambiguous strings have not been included in the list (2b):

AUI	hau-issa / ha-uissa
AUU	lau-on / vara-uoma
EIA	rei-än / eläke-län
IOU	kavio-ura / hi-outua

It would be easy to generalize over some of the instances (2b). E.g., given that X stands for all vowels except ä, the HP would be after the middle ä in AI-X, EI-X, OI-X, UI-X.

5. Rules for two-vowel strings

For combinations of two vowels, two subcases may be distinguished. One concerns HPs due to vowel harmony (3a), the other certain nondiphthongs (3b).

(3a) If a two-vowel string runs counter to the requirements posed by vowel harmony, there is an HP between the vowels.

The following nonharmonic strings thus contain HPs (äy, frequent in English borrowings, has not been included). As above, plus-signs have been prefixed to the most frequent ones. Some of these strings are next to nonexistent.

STRING	EXAMPLE
A-Ä	sukkela-ällyinen
+ A-Ö	raaka-öljy
O-Ä	teko-äly
O-Y	kanto-yrittäjä
+ O-Ö	osto-öljy
U-Ä	
U-Ö	
U-Y	kaasu-yhtiö
+ Ä-A	määrä-ala
+ Ä-O	etelä-osa
Ä-U	epä-urheilullinen
+ Ö-A	henkilö-arvio
Ö-O	kiinteistö-osa
Ö-U	
+ Y-A	myrkkö-annos
Y-O	risteily-ohjus
Y-U	nyky-urheilu

Rule (3b) provides HPs for nondiphthongs. By definition, these strings contain a syllable boundary between the vowel segments.

(3b) The following nondiphthongs contain an HP.

STRING	EXAMPLE
+ A-E	anta-essa
A-O	Kala-ostos
+ E-O	te-os
+ O-A	auto-ani
U-A	auto-asta
+ Ä-E	vetä-essä
Ä-Ö	uikona-össä
Ö-E	näkö-eristys
Ö-Ä	näkö-äni
Y-E	nyky-ennuste
Y-Ä	kääntö-äkseen

The strings EA, EÄ, IA, IÄ, IO, IÖ, OE, UE have not been included since they would often overgenerate errors like korke⁺-akoulu, Re⁺-ading, lipe⁺-äkälä, ko⁺-ekäsittely. Cf. rule (3c) for some precisions.

(3c) The following frequent word-final strings are hyphenated as indicated.

STRINGS	EXAMPLES
+ E-AN, E-AT	kärke-an, kärke-at
E-ÄN, E-ÄT	hilpe-än, hilpe-ät
I-AN, I-AT	halti-an, halti-at
I-EN	sai-en (except after E)
I-ON, I-OT	valti-on, valti-ot
I-ÄN, I-ÄT	mini-än, mini-ät
I-ÖN, I-ÖT	nelli-ön, nelli-öt

6. A rule for consonant strings

The CV-rule properly hyphenates all domestic consonant combinations, e.g. kal-lijo, yk-si, mels-keessä. Problems are posed by simplex loans and by compounds where a word-medial (borrowed) part begins with a consonant cluster. Hyphenations such as Lundst*-röm, Whitb*-read, Voltpop*-rosentli, nakkisist*-roganoff are unacceptable. Rule (4) states some instances where the HP normally is located before certain word-medial two-consonant combinations.

- (4) There is an HP before the following word-medial strings (with a consonant or vowel to the left, and a vowel to the right).

STRING	EXAMPLES
-BL	Lind-blad, Ek-blom, suurvaita-blokki
-BR	Kilpailu-bridge, kokonais-butto, Rem-brandt
-DR	perhe-draama
-FL	housu-flanelli, in-flaatio, Butter-fly
-FR	Wil-Fried, seinä-fresco, kohteliaisuus-fraasi
-GL	lithas-globullini, pre-glastaallinen (not after N)
-GR	Lind-gren, milliy-gramma (-"-)
-KL	hammas-klinikka, basso-klarinetti
-KR	Kulttuuri-kriisi, raamatun-kriittikki
-KV	jazz-kvartetto, säteily-kvantti
-PL	uus-platonismi, nenä-plastikka
-PR	apulais-professori, ilma-putkaaci
-CL	rupia-clearing
-QV	Lund-qvist, En-qvist
-SCH	Welt-schmerz

This rule is based also on strategic reasoning in the sense that there are a few domestic words containing some of these strings, in particular Ku, KR, PL, PR. Counter to (4), these words are hyphenated vik-lä, yuok-lata, kup-lia, etc. The proper solution is to list these words, amounting only to a few

tens, as exceptions to (4). This finite procedure makes it possible to use (4) as the general case.

But there are also borrowings diverging from the prescriptions of (4), e.g. ak-vaario, hyb-lidi, sop-laano, mig-legni. Furthermore there are many frequent nondomestic names such as Lind-ström, Sjö-stedt, whose HPS are not properly determined by the CV-rule and rule (4) as they stand. The only conceivable solution is an optimized frequency-based exception list. We have compiled such a list containing some 100 items. Rule (4) and the list together eliminate errors such as jazzk*-vartetto, lapsip*-sykiatri, Lindst*-röm, neekerig*-heltto.

7. The N/S-restriction

Trivial Finnish hyphenation programs contain only the CV-rule or some equivalent of it. In programs with higher ambition levels, the most persistent hyphenation errors concern certain word-medial strings consisting of:

vowel + \bar{n} or \bar{g} + vowel] + consonant + vowel

Errors: asia \bar{n} *-mistaja, asia*-nomistaja, asuna*-luvet, asul*-naluuet, kaavoitus*-sasto, kaavoitu*-sosasto. The problem is that there are three potential HPS in these strings but no fully safe non-lexical means of picking the right one(s). Cf.:

CORRECT HP(S)	EXAMPLE
V N/S- V C V	hevo*-s-o*-mistaja
V -N/S V -C V	pinta*-s*-i-vistys

Without a full root lexicon and a sophisticated operational derivational morphology there is no way of knowing on which side of the N/S there is an HP, or whether there is more than one HP in the string. These ambiguous strings are especially

likely to occur in derived words whose (all) components would not be found even in an extensive root lexicon. Some examples with us-derivatives have been provided above.

These errors may be safely avoided only by lowering recall in regard to the critical N/S strings.

(5) A word-medial pattern of Y N/S Y C Y is not hyphenated at all, provided the N/S is not at the first or last syllable boundary of the word.

(5) is a restriction on the CV-rule and leaves unhyphenated strings such as the following italicized ones: *kan-gānēdūs-taja*, *kaavoi-tūsosās-to*, *kehi-lyyagguā*, *be-toniseinämä*. The restriction concerning syllable boundaries is necessary if recall is not to be lowered so much as to render the algorithm practically useless. As it stands, (5) allows hyphenations like *ka-naia*, *kana-si* and has a recall lowering effect of some 2 %.

The practical problem with (5) is that it leaves a residue of fairly long (7-12 letters) unhyphenated strings (cf. section 8). In real systems, it would be advisable to implement (5) to be interactively checked by the human user.

8. Evaluation of implementations

Using Brodda's BETA-system several versions of the hyphenation rule set have been implemented. The versions differ in regard to recall and precision. Different performance levels might be appropriate for different purposes. Of course, an optimal implementation with full coverage and minimized error rate requires more memory space and processing time than one compromising in either or both respects.

First, we reiterate that plain CV-hyphenation, i.e. an algorithm containing only rule (1) while complying to the typographical requirements, has a recall of 95,7 % and a precision of 98,74 %.

The complete version of the algorithm, BETA/FIN/HYP, mod-

els optimal hyphenation with maximal precision. It implements rules (1-5) in the form stated above. In particular, the N/S-restriction (5) is strictly interpreted. No chances are taken in hyphenating medial vowel-internal n's and s's unless full security obtains. BETA/FIN/HYP includes a list of some 250 exceptional words or parts of words running counter to the general rules. This list is in addition to the one mentioned in connection with rule (4).

The recall of BETA/FIN/HYP is 95,0 % and its precision, within these confines, 99,86 - 99,95 % depending upon what kind of text is being processed. It thus raises the precision of plain CV-hyphenation by more than 1,1 %. These figures derive from a test corpus comprising 8,739 words of running text with some 15,000 potential HPs. BETA/FIN/HYP thus eliminates almost all errors left by CV-hyphenation and lays only some 0,1 % from perfection. Its main limitation is practical and due to the fact that recall was compromised in order to maximize precision. Some 2 % of the word forms in running text come to contain sequences of 7-12 letters where no HP is found with certainty. Relevant examples are the italicized strings in e.g. *bud-jet-tiesitys*, *kor-keakouluis-sa*, *val-tionneuvos-ton*, *uv-delleenorien-toi-tu-mi-nen*, *mas-sayllipis-ton*, *kon-sert-tiyleisö-jen*.

The following example shows how BETA/FIN/HYP applies to running text. All HPs identified by the program are indicated by hyphens. This particular text contains some 450 words with 897 potential HPs and happens to be free from errors. All unhyphenated strings longer than 7 letters (N=12) have been underlined. Not a single one occurs at the actual border of two lines, i.e. deficient recall causes actual harm only in a fraction of the instances (and then hyphenation should be interactively checked). All potential HPs not spotted by BETA/FIN/HYP (N=41; 4,6 %) are indicated by a manually added slash /.

KAS-KU-JA, VIT-SE-JÄ, PI-LA-PIIR-HOK-SIA, KO-ME-DI-OI-TA, PARS-
SE-JA, HU-PAI-LU-JA, HUO-LIA JNE. TAR-JOIL-LAAN IH-MIS-TEN KU-
LU-TER-TA-VIK-SI PÄI-VIT-TÄIN YM-PÄ-RI MAA-IL-MAN SUUN-MAT-TO-
MI-EN TUO-TAN-TO JA LE-VI-TYS-KO/NEIS-TO-JEN VOI-MAL-LA. JOUK-
KO-TIE-DO-TUK-SEN ULOT-TU-MAT-TO-MIS-SA ELÄÄ VIE-LÄ EDELL-IS-TÄ
HUI-KE-AM-PI KAS-KU-JEN JA VIT-SI-EN FLOO-RA, JO-KA EI OLE SO-
VE-LI/AI-SUUS-SYIS-TÄ PÄKS-SYT MO-NEN-KAAN PAI-NO-KO-NEEN PU-
RIS-TUK-SEEN, MUT-TA JO-KA OH-KUU SUUS-TA SUO-HUN KUL-KE-VA-NA
PE-RIN-TEE-NÄ ELIN-VOI-MAA. OSA TÄS-TÄ KAS-KUS-TOS-TA PÄI-VIT-
TÄIN UNOH-TUO. MUT-TA AJÄN-KOH-TAI-SIS-TA TA-PAH-TU-MIS-TA
SYN-TYV VÄ-LIT-TÖ-MÄS-TI UU-DET KAS-KUT JA UU-DET VIT-SIT. KAN-
SAN-PE-RIN-TEEN LA-JI-NA KAS-KUT OVAT TAA-JAAN SI-KI/Ä/VIÄ,
RÖN-SYI-LE-VIÄ JA SA-MAL-LÄ EH-KÄ HA-TA-RIM-MIN PUT-KIT-TU-JA.
ENG-LAN-TI-LAI-NEN ES-TES-TIK-KO SPAR/SHOT ON TO-DEN-NUT, ET-
TÄ HUO-NO-RIN-TUT-KI-MUK-SEN ALAL-LA MO-NET JÄL-JET VIE-VAT
LUO-LAAN, MUT-TA YH-DET-KÄÄN EI-VÄT TU-LE ULOS. LI-SÄK-SI VÄI-
TE-TÄÄN, ET-TÄ HUO-MO-RIN-TUT-KI-MUS-TA PUI-SE-VAM-PAÄ PUU-HAA
SAA HA-KE-MAL-LÄ HA-KEA. KYL-LÄ TÄ-MÄN-TA-PAI-SIT-LÄ EVÄIL-LÄ
JO KE-SAN-NOI-TUI-SI MI-KÄ TA-HAN-SA PEL-TO. KUN RU-VE-TAAN
JÄL-JIT-TÄ-MÄÄN KAS-KU-JEN JA NY-RY-PÄI-VÄN VIT-SI-EN MEN-NEI-
SYIT-TÄ, JOU-DU-TAAN PI-AN TE-KE-MI-SIIN SA-TU-JEN JA ERI-TYI-
SES-TI PI-LASATU-JEN KANS-SA, JOI-DEN JUO-RET JOH-TA-VAT IN-TI-
AAN JA KII-NAAN, MIS-TÄ ON TA-VAT-TU PI-LA/SA/TU-JEN VAR-HAI-
SIM-MAT LÄP-TEET. PI-LÄ/SA/DUIL-LA ON OL-LUT VA-KIIN-TU-NUT
ASE-MAN-SA MYÖS AN-TI-KIN KREI-KAS-SA, MIS-SÄ SA-DUN-KER-TO-
JI-EN OH-JEL-MIS-TOON OVAT MUON MUAS-SA KUU-LU-NEET KE-VY-ET JA
USEIN EROOT-TIS-SÄ-VYI-SET PI-LÄ/SA/DUT. NI-I-DEN AI-HEL-MIIN
POH-JAA-VAT MO-NET OVI-DI/UK-SEN ME-TA-MOR-FOO-SIT, JOI-TA ON
LÖY-DET-TÄ-VIS-SÄ SUO-MA-LAIS-TEN-KIN KIR-JA-KAUP-PO-JEN JA
KIR-JAS-TO-JEN HYI-LYIL-TÄ MAAN KIE-LEL-LE KÄÄN-NEP-TYI-NÄ. IL-
MEI-SES-TI KAN-SAN-OMAI-NEN PI-LA/SA/TU-PE-RIN-NE ON EU-ROO-
PAS-SA LE-VIN-NYT LIUK-KAIM-MIN KES-KI/AJAL-LA KIER-TE-LE-VI-EN
TEI-NI-EN HUO-LIL-LA. HEI-DÄN SUO-SI-MAN-SA, ALUN-PE-RIN RANS-
KAS-TA LÄH-TE-NYT JA I200-1300-LU-VU-LÄ YM-PÄ-RI EU-ROOP-PAÄ
PIRS-TOU-TU-NUT FAB-LI-KUX-RU-NOUS, JO-KA AM-MEN-SI AI-HEEN-SA
KAN-SAN KER-TO-MIS-TA PI-LÄ/SA/DUIS-TA JA MUO-TON-SA SO-FIS-

TIS-TEN KIR-JOIT-TA-JI-EN KY-NIS-TÄ, OLI KAI-KES-SA REA-LIS-TI-
SUO-DES-SAAN PI-LAI-LE-VAA, AR-VO-JA HÄ-PÄI/SE/VÄÄ JA ADAS-SAAN
SUO-RAS-TAAN KA-PINAL-LIS-TA. PIL-KAN-TE-ON KOH-TEI-NA OLI-VAT
KLEEN-SÄ AVIO-LIIT-TO JA AVIO-ELÄ-MÄ SE-KÄ EN-NEN KAIK-KEA HEN-
GEL-LI-NEN SÄÄ-TY PAP-PEI-NEEN, MUNK-KEI-NEEN JA NOR-NI-NEEN.
HEI-DÄN ELÄ-MÄN-TA-PO-JEN-SA NUR-JA PUOL-TA ESI-TEL-TIIN KAIH-
TE-LE-MAT-TO-MIN JA IVALI-LE-VIN SA-NA-KÄÄN-TEIN, JO-T-KA ANI
HAR-VOIN KIER-TYI-VÄT SÄÄ-TYYN KUU-LU-VI-EN EDUK-SI. DOC-CAC-
CIO ON KÄYT-TÄ-NYT DE-CA-ME-RONEN-SA, SA-MOIN KUIN CHAU-CER
CAN-FER-BU-RYN TA-HI/NOI/DE-N-SA AI-NEIS-TO-NA LÄ-HES YK-SIN-
OMAAAN KUU-LE-MI-AAN JA LU-KE-MI-AAN PI-LA/SA/TU-JA. NÄI-DEN
TEOS-TEN KER-TO-MUK-SET KIE-LI-VÄT NY-RY-AJAN IH-MI-SEL-LE EH-
KÄ PAR-HAI-TEN MEN-NEIT-TEN AI-KO-JEN PI-LA/SA/TU-JEN SUO-SI-
TUM-MIS-TA AI-HEIS-TA. PI-LÄ/SA/DUT KU-TEN KAS-KUT-KIN OVAT
AI-HE-PII-RIL-TÄÄN REA-LIS-TI-SIA JA POI-MI-VAT AI-NEK-SEN-SA
AR-KI-PÄI-VÄN ELÄ-MÄN TI-LAN-TIIS-TA. SA-TU-JEN KÄR-KI ON-KIN
MO-NES-TI SUUN-NAT-TU IH-MEI-TÄ HA-VIT-TE-LE-VIA JA TOI-SEN-
LAI-SES-TA ELÄ-MÄS-TÄ HAA-VEI-LE-VIA VAS-TAAN, JA USEIM-MI-TEN
NE JO-T-KA AN-TAU-TU-VAT UNEL-MIEN-SA VAL-TAAN, JOU-TU-VAT KAT-
KE-RAS-TI PE-TE-TYIK-SI JA NO-LA-TUIK-SI. KAS-KUIS-TA PI-
LA/SA/DUT EN-SI-SI-JAI-SES-TI ERO-A-VAT SII-NÄ, ET-TÄ JÄL-KIM-
MÄI-SET OVAT TA-VAL-LI-SES-TI MO-NI/EPI/SO/DI-SIA JA NI-I-DEN
VAIH-TE-LE-VÄT NÄYT-TÄ-MÖT JA KER-SE-LI-AÄS-TI ETENE-VÄT TA-
PAH-TU-MAT LUO-VAT USEIN ESI-TYK-SEL-LE TAI-DO-KAS-TA DRA-MAAT-
TI-SUUT-TA JA JÄN-TE-VYIT-TÄ. SUO-MA-LAIS-TEN PI-LA/SA/TU-JEN
TUN-NE-TUIK-MAT HEN-KI-LÖT OVAT EPÄI-LE-MÄT-TÄ MAT-TI JA PI-RU.
MAT-TI, KÖY-HÄ JA HEI-VE-RÖI-NEN POI-KA, ON PES-TAU-TU-NUT PI-
RUL-LE REN-CIK-SI, JA NÄI-DEN KAH-DEN VOI-MAIN-MIT-TÄ-LYIS-TÄ
SA-DUT RA-KEN-TU-VAT. MAT-TI SEL-VIT-TÄÄ KIL-VAT PI-RUA ÄLXK-
KÄÄM-PÄ-NÄ LÄ-HES POIK-KEUK-SET-TÄ EDUK-SEEN. MAT-TI VOIT-TAAN
PI-RUN KIL-PA/SYON-NIS-SÄ MÄT-TÄ-MÄ-LÄ PUU-ROA RIN-NAL-LEEN
SI-TO-MAN-SA SÄK-KIIN. MOU-KA-RIN-HEIT-TO-KIL-VAS-SA MAT-TI
PE-RIL VOI-TON UH-KÄÄ-MAL-LÄ HEIT-TÄÄ PI-RUN HO-PEÄ/MOU/KÄ-RIN
PIL-VEN-LON-GAN PÄÄ-LE, JO-L-LOIN PI-RU HÄ-TÄÄN-TYNEE-NÄ LUD-
VUT-TAA KO-KO KIL-VAN. LAS-TEN KU-VA-LEH-TI JUE-KAI-SI AI-KOI-
MAAN VUO-SI-KAU-SIA MA-TIS-TA JA PI-RUS-TA KER-TO-VAA SAR-JA-
KU-VAA OT-SI-KOL-LA MAT-TI JA PEIK-KO. MA-TIS-TA JA PI-RUS-TA
KER-TO-VAT PI-LÄ/SA/DUT TOIS-TA-VAT PE-TE-TYS-TÄ PA-HO-LAI-SES-

TA KER-TO-VANSA-DUN KAN-VAM, JON-KA UNZ HELLERES
 VAF-TU SA-TA-KUN-TA JA JOT-KA LAI-NAU-TU-VAF MEIL-LE SOU-REK-SI
 OSAK-SI IDÄS-TÄ. TÄ-MÄN AI-HEEN KAN-SAIN-VÄ-LI-SES-TI TUN-NE-
 TUN PO-TEU-TUS LIE-NEE ODYS-SEI-AS-TA TUT-TU PO-LY-FE-MOS-MO-
 TI-VI, JOS-SA JÄT-TI-LÄI-SEN AI-NO-AN SIL-MÄN SO-KAI-SEE
 ÄLYYN-SÄ TUR-VAU-TU-NUT ODYS-SEUS, >>EI-KU-KAAN>>.

Recall is lacking and unhyphenated strings arise especially when word-medial vowel combinations and n's or s's in short syllables occur intermittently. The word hä-päiseyää is a pertinent example. Potentially, i.e. depending upon morphological structure, there could be HPS as in hä-pä-i-s-e-vää.

The long unhyphenated strings due to intentionally lowered recall may be largely eliminated, at least to 70 %, by implementing the N/S-restriction (5) as an interactive rule introducing secondary hyphens. Their correctness should then be checked by the human user. Such a secondary hyphenation rule not dependent upon morphological information runs an error risk of 4,6 % when hyphenating in the contexts V_nV, V_sV, and a somewhat higher error risk of 13,3 % when hyphenating in the contexts VnV_CV, Vsv_CV (at non-first syllable boundaries this risk is 22 %, for n no less than 35 %!).

Word-medial ViCV-strings beyond the first syllable boundary must be hyphenated Vi-CV even at the risk of producing a few errors like isoi+sä. Otherwise, recall will be lowered to an impracticable level.

The following list exemplifies the performance of BETA/FIN/HYP upon words that are especially difficult to hyphenate properly while trying to optimize recall. The list was compiled for the very purpose of testing the system against as tough instances as possible and is thus not representative of running text. Errors belonging to the 0,14 - 0,05 % precision deficit are prefixed by a star and the correct HP is provided in parentheses. Unidentified potential HPS are again indicated by a manually added slash, and unhyphenated strings longer than 7 letters underlined.

Application of BETA/FIN/HYP to difficult test words

KIE-LI-ALU-EI-LA	VLI-OPIS-TU
PAH-AST/AS-SA	NA-KO-ALA-PAIK-KA
VAC-TA-OSAI-LA	SÄH-KO-ERIS-PYS
UR-HET-LU/ELA-MÄN	MO-NI/ALA/YRIT-TÄ-JI-EN
NAVY-TE-LY-OSAS-TU	TI-MAN-TI/ALAN
UH-KA/YRI-TYS	VA-RA-OSI-EN
UL-KO-ASI-AIN-VA-LIO/KUN-TA	ERIN-OMAI-SEN
YDIN-ASEI-TA	KE-HI-TYS/YH-TSIS-TYO-OSAS-TU
PUO-LIU+NEEN (PUO-LI-)	SO-TI-LAS-VI-SAN/OMAI-NEN
YU-LEN/ARAK-SI	KAN-SA-LAIS/OPIS-TU
YK-KUS/AVUS-TA-JA	NI-MEN-OMAA
KES-KUS/AM-RAF-TI-KOU-LU	KE-HI-TYS/YH-TEIS-TYO
AVAIN/ALOIT-LA	YH-TEY/DEN/OF-TO-AM-NE
LAA-JEN-SUS/OSAN	KAR-SI-NI+SESI-TYK-SEN (KAR-SI-MIS-)
ASI-AN-OSAI/SE-NA	NAIS-ASIA/NAI-SIA
TOI-SIN/ADAT-TE-LI-JOI-TA	ASUIN-ALU-EI-TA
VIK-STROM	ASTROM
ER-BLOM	LIND-BLAD
WIL-FRIED	MAN-CHESS-TER
REB-BRANDTIL-LA	2DE-LAC-ROIX (DE-LA-)
REN-NO-SKAN-DI-AS-SA	ULL-STEN
DRAGS-FJÄRD	BUT-TER-FLY
KI-LING-TUN	IN-FLAA-TIO
SI-NE-BRYCHOP-FIN (?-CHOP-)	SE-GER-STAM
HUS-QVAR-NA	WHIT-BREAD
PORTS-MOUT-HIN	ZA-GORT-SCHIK
PIHL-GREN	HOCH-SCHUL-KOR-SE
PIE+NE*LÄIN-KLINIK-KA (PIEN-)	JAZ+ZIL-TA-MAT (JAZZ-)
NEE-KE-RI-GHEP-TU	VAN+HUUS/IAN
KOH-AJO	TYO-ELA-KE
TYO-IAN	DEOS-OT-TO-MIES
IA+NI+KUI-NEN (IAN-)	ELA-KE/IAN
FIN-LAY/SON	MUO-TI-SHOW
HUU-ME-PRO-TES-TI	UUS-ENG-LAN-TI-LAI-NEN
OUUT-YH-TY-MÄN	HUU-ME-PRO-TES-TI

GOLDS*-MITH (GOLD-)	TYOP*-ROSES-SI (TYO-)
PIK-KU-PLA/NEET-VA	SYYS-ALE
OU*-SISO/LA-TIONIS-MI (OUS-)	YLEIS-AVAIN
YU-KLU-BI	LI-SA-KRAA-SA
SA-TET-LY-SPAN-DAR-DI	MUR-RAYN
NIGHT-CLUB	TO-YO-TA (TOY-)
<u>RAY/KOVD</u>	AI-OIN
LEU-OIS-SA	LA-EIS-SA
HAL-KE-AL-SI	KU-TI-AI-SI
LI-PE-OI-DA	RE-PE-AI-SI
KES-KIR*-UROOP-PA-LAI-NEN (KES-KI-)	
PUO-LU-EIS-SA	SAIP-PU-OI-TA
PE-SY-SI-TA	KA-RY-AI-SI
TAI-AN	AI-ON
HAD-EN	TIE-YH-DIS-TYS
VA-LIO-AK-KU	VAL-TIO-ELIN
YH-TIO-ESI-TYS	KIU-AS
LIO/B-TA	SAI/ET-TA
TAV-ON	TE-KO-ADY
MAA-RA-ALA	KA-SU-YH-TIO
RIS-TET-LY/OH-JUS	AN-TA-ES-SA
TE/OS	VE-TA-ES-SA
KA-LA-OS-TOS	KAAN-TY-AK-SEEN
NY-KY-EN-NUS-TEET	KAR-KE-AN
SA-LI-EN	HIL-PE-YI
VAL-TI-OT	VAL-TI-OIS-SA
BAS-SO-KLA-RI/NET-TI	KUUT-TUO-RI-KRII-SI
SA-TET-LY-KVANT-TI	OUSP*-LA-TONIS-MI (OUS-)
RUP-LA-CLEA-RING	HAM-MAS-KLI/NIR-KA
SEI-NA-FRES-KO	HOU-SU-PLA/NEL-LI
PER-HE-DRAA-MA	KIL-PAI-LU-BRID-GE
LI-HAS-GLO-BU-LII-NIA	VA-LU-GAA-FIT-TI
MEZ-CO-SOP-RAA-NO	KAL-LIO-GNEIS-SI
TAP-LA	VIX-LA
LAP-SI-PSY-KI/AT-RI	ETR-LA-SLAA-VI-LAI-NEN
KOS-KEN-KOR-VA-SNAP-SI	AM-MAT-TI-STA-TUS
POH-JOIS-SLO-VEE-NI	AU-TO-STE-REOT
MAT-KA-STI-PEN-DI	EM-MEN-TRAL

It is particularly relevant to examine how *SESTA/ELN/HYP* fares in regard to compounds with borrowed constituent parts, to strings of three or more vowels, and to native words with such special shortened prefixes as *uus-*, *pien-*. The vast majority even of these strings is correctly hyphenated. Some of the errors, e.g. *puoliu*-neen* and *karsimil*-sestykseen*, might seem surprising. However, in real texts, they belong to the almost vanishingly small 0,14 - 0,05 % error share and are due to recall optimization of certain word-medial or word-final suffixal elements.

The errors occur partly in borrowings or other foreign words. These are impossible to incorporate exhaustively in lexical lists that cannot contain thousands of items if the system is to be practicable. Partly the errors appear in a few individual stems or words running counter to otherwise pervasive regularities (e.g. the errors occurring at the first syllable boundary in compounds such as *pie*-neläinklinkka*, *uu*-sisolationismi*, where restrictions are not possible to state since *uus*, *pieni* have inflectional stems with *i* and/or *e* after *pie-* or *uus-*).

Finally, the following list demonstrates how *BEVA/FIN/HYP* treats word tokens containing nonletter characters. These are quite frequent in running text and should be properly treated by a high-quality hyphenation algorithm. As above, all the HPS identified by the program are marked by a hyphen (the lower form of each pair). Of course, not all of these hyphens would actually occur in a production version since several nonletter characters qualify as such as HPS. The upper form in each pair is the unhyphenated word.

Application of BETA/PIN/HYP to test words containing non-letter characters

ETELÄ-AMERIKA
ETE-LÄ-AMS-RIK-KA

LEIKKI-IKÄINEN
LEIK-KI-IKÄI-NEN

KILPA-AJO-ORI
KIL-PA-AJO-ORI

SEINÄJOKI-ALAHÄRMÄ-YLIHÄRMÄ
SEI-NÄ-JO-KI-ALA-HÄR-MÄ-YLI-HÄR-MÄ

TEOL.YLIOPP.
TEOL.-YLI-OPP.

TEOL.YO.
TEOL.-YO.

DIPL.INS.
DIPL.-INS.

ECC:STÄ
ECC:-STÄ

UNIC:IIN
UNIC:-IIN

NAIMATON/NAIMISSA/LESKI/ERONNUT
NAI-MA-TON/-NAI-MI-SIS-SA/LES-KI/ERON-NUUT

RAA'ANLAINEN
RAA-AN-LAI-NEN

D*API
D*API

O'HARA
O'HA-RA

HENRY'S
HEN-RY'S

"TELSET"-YRITYS
*"TEL-SEP"-Y-RI-TYS ("TEL-SEP"-YRI-TYS)

1-VUOTIAS
1-VUO-TIAS

22-VUOTIAS
22-VUO-TIAS

12.00
12.00

13,76
13,76

100.000.000
100.000.000

"YLI-MÄÄRÄ"
"YLI-MÄÄ-RÄ"