Cahiers GUTenberg

LE TRAITEMENT DES TEXTES POLONAIS AVEC LE LOGICIEL TEX

Hanna KOLODZIEJSKA

Cahiers GUTenberg, nº 0 (1988), p. 3-10.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1988___0_3_0>

LE TRAITEMENT DES TEXTES POLONAIS AVEC LE LOGICIEL TEX

Hanna KOLODZIEJSKA
Instytut Informatyki
Uniwersytet Warszawski
PKiN p.850
00-901 Warszawa, Pologne

Résumé

Je présente dans ce document quelques difficultés d'utilisation du logiciel TEX pour les textes polonais, liées à l'insertion des lettres avec des signes diacritiques, à la division des mots et à certaines particularités de la typographie polonaise.

Introduction

Etant une langue slave, le polonais appartient au groupe des langues indo-européenes. Il est parlé par environ 50 millions de personnes en Pologne et à l'étranger. Le polonais n'est pas facile à apprendre à cause de sa grammaire assez compliquée, son ortographe et sa prononciation difficile, tellement différente de celle des langues romanes par exemple. Il est rarement connu par les étrangers, c'est pourquoi la plupart des documents scientifiques, ceux concernant les mathématiques en particulier, écrits en Pologne, sont publiés en anglais.

L'alphabet polonais

L'alphabet polonais est basé sur l'alphabet latin. Il est composé de 32 lettres suivantes:

a	A	f	F	m	M	ś	Ś
ą	Ą	g	G	n	N	t	T
b	B	h	H	ń	Ń	u	U
c	C	i	I	o	O	w	W
ć	Ć	j	J	ó	Ó	y	Y
d	D	k	K	p	P	z	Z
e	E	l	L	r	R	ź	Ź
ę	Ę	ł	Ł	s	S	ż	Ż

Il y a neuf lettres avec de différents signes diacritiques: ą, ć, ę, ł, ń, ó, ś, ź, ż. Chacune d'elles est prononcée de manière tout à fait différente que la correspondante lettre latine. Par exemple: le mot "róża" (*rose*) est prononcé en polonais de la même façon que le mot

rouja serait prononcé en français. Sans signes diacritiques nous prononcerions ce mot comme *rosa* en français.

Parfois l'omission des signes diacritiques change le sens du mot. Par exemple les mots "półka" et "polka" signifient respectivement *rayon* et *polka* (une danse d'origine tchèque). Ainsi le manque des signes diacritiques dans le texte polonais change essentiellement la prononciation des mots, parfois même leur sens et écorche le polonais.

Insertion des lettres polonaises

Rares sont en Pologne les claviers comportant tout l'alphabet polonais. Le plus souvent nous nous servons des claviers n'ayant que 26 lettres latines ce qui nécessite l'utilisation des commandes TEX pour accentuer certaines lettres. Suivant le signe diacritique nous distinguons quatre groupes des lettres accentuées:

(1.) ć ń ó ś ź (et les correspondantes lettres majuscules)—nous pouvons les obtenir par surimpression de l'accent aigu sur la lettre latine appropriée, par exemple:

\'c ⟶ ć

(2.) ż (Ż)—nous l'obtenons de la manière suivante:

\.z ⟶ ż
\.Z ⟶ Ż

(3.) ł (Ł)—cette lettre, qui apparaît dans l'alphabet polonais et lituanien, est appelée dans [K]: "Polish suppressed-L" et peut être obtenue à l'aide des commandes suivantes:

\l ⟶ ł
\L ⟶ Ł

(4.) ą ę (Ą Ę)—ce signe diacritique, appelé "ogonek" dans [ISO], manque dans TEX. Il diffère de la cédille par sa forme et avant tout par la courbure inversée. En utilisant la cédille nous obtenons les caractères suivants:

\c a ⟶ ą
\c e ⟶ ę

Leur qualité typographique n'est pas bonne. "Ogonek" qui apparaît dans ce document, symétrique à la cédille, a été obtenu par une simple modification du fichier **amr10.pxl** (le fichier **amr10.tfm** n'a suivi aucun changement).

Les lettres accentuées apparaissent très souvent dans les mots polonais si bien que l'utilisation des commandes TEX n'est pas commode. Par exemple pour obtenir la phrase suivante:

"Żółty liść spadł z dębu"
(*Une feuille jaune est tombée du chêne*)

il faut écrire:

```
\.Z\'o\l ty li\'s\'c spad\l\ z d\c ebu
```

ce qui est pratiquement illisible.
D'autres exemples: après l'insertion de la suite des caractères:

```
wzi\c a\'c si\c e w gar\'s\'c
```

nous obtiendrons l'expression:

"wziąć się w garść"
(reprendre son sang froid, se remettre)
et après l'insertion de `\L\'od\'z` c'est le mot "Łódź", le nom d'une grande ville polonaise, qui sera imprimé.

Pour faciliter l'insertion des lettres polonaises nous pouvons adopter la solution qui ressemble à celle trouvée dans notre Institut pour d'autres logiciels. Nous distinguons un signe particulier (`@` — *à commercial*) informant que la lettre qui le suit doit être accentuée.

Nous insérons:	pour obtenir:
`@a`	ą
`@c`	ć
`@e`	ę
`@l`	ł
`@n`	ń
`@o`	ó
`@s`	ś
`@z`	ź
`@r`	ż

Nous faisons de même pour les lettres majuscules. Nous écrivons `@r` pour la lettre 'ż' afin de la distinguer du 'ź' décrit par `@z` (il y a deux signes diacritiques différents au-dessus de la même lettre). La désignation `@r` n'est pas fortuite. Nous l'avons choisie, parce que la lettre 'ż' et la consonne composée 'rz' sont prononcées en polonais de la même façon (comme *j* dans le mot français *jeudi*). Les expressions présentées ci-dessus en tant qu'exemples seraient insérées de la manière suivante:

```
@R@o@lty li@s@c spad@l z d@ebu
wzi@a@c si@e w gar@s@c
@L@od@z
```

Cette solution facilite l'insertion du texte polonais et le rend un peu plus lisible en entrée. Voilà le texte complet des définitions TEX ajoutées:

```
\catcode'@\active
\def@{\leavevmode\futurelet\next\lettertest}
\def\lettertest
    {\ifx\next a\def\diacr{6}\else
     \ifx\next A\def\diacr{7}\else
     \ifx\next c\def\diacr{1}\else
     \ifx\next C\def\diacr{1}\else
     \ifx\next e\def\diacr{0}\else
     \ifx\next E\def\diacr{0}\else
     \ifx\next l\def\diacr{4}\else
     \ifx\next L\def\diacr{5}\else
     \ifx\next n\def\diacr{1}\else
     \ifx\next N\def\diacr{1}\else
     \ifx\next o\def\diacr{1}\else
     \ifx\next O\def\diacr{1}\else
     \ifx\next r\def\diacr{2}\else
```

```
\ifx\next R\def\diacr{3}\else
\ifx\next s\def\diacr{1}\else
\ifx\next S\def\diacr{1}\else
\ifx\next z\def\diacr{1}\else
\ifx\next Z\def\diacr{1}\else
\fi\fi\fi\fi\fi
\fi\fi\fi\fi\fi
\fi\fi\fi\fi\fi
\fi\fi\fi
\ifcase\diacr
   {\c \next}\or
   {\accent"13\next}\or
   {\accent"5Fz}\or
   {\accent"5FZ}\or
   {\char'401}\or
   {\setbox0=\hbox{L}\hbox to\wd0{\hss\char'40L}}\or
   {\setbox0=\hbox{a}a\kern-\wd0\char24\kern-.23em}\or
   {\setbox0=\hbox{A}A\kern-\wd0\kern.16em\char24\kern-.12em}\fi
\let\next}
```

Le signe @ défini comme *actif* et la lettre qui le suit seront remplacés par des commandes respectives. Il faut remarquer que dans ce cas-là le contrôle de macro-instructions à l'aide de `\tracingmacros` devient plus compliqué à cause des très grandes dimensions du fichier `.log`.

Outre une certaine commodité d'insertion des lettres polonaises cette solution a encore un autre avantage. Même après avoir créé avec METAFONT des polices des caractères polonais, la façon d'insérer les textes resterait invariable. Après avoir effectué de simples modifications des fichiers `.tfm` chaque lettre avec un signe diacritique serait traitée comme ligature. Par exemple `@c` constituerait la ligature 'ć'. Le traitement des lettres accentuées comme ligatures permettrait à TEX de diviser les mots polonais (voir [A], [K]). Il suffit seulement, conformément à ce que W. Appelt a proposé dans [A], de changer la catégorie du signe @ pour 11:

```
\catcode'\@=11        \lccode'@='@
```

Il est évidemment possible de passer aux ligatures de ce type même dans le cas où nous insérons le texte en utilisant les commandes du PLAIN TEX pour créer les lettres accentuées. Il suffit uniquement de modifier leur définition:

```
\def\.#1{\if#1z{@r}\else
         \if#1Z{@R}\else
         {\accent 95 #1}\fi\fi}
\def\'#1{\if#1c{@c}\else
         \if#1C{@C}\else
         \if#1n{@n}\else
         \if#1N{@N}\else
         \if#1o{@o}\else
         \if#1O{@O}\else
```

```
10          \if#1s{@s}\else
11          \if#1S{@S}\else
12          \if#1z{@z}\else
13          \if#1Z{@Z}\else
14          {\accent 19 #1}\fi\fi\fi\fi\fi\fi\fi\fi\fi\fi}
15 \let\ccc=\c
16 \def\c#1{\if#1a{@a}\else
17          \if#1A{@A}\else
18          \if#1e{@e}\else
19          \if#1E{@E}\else{\ccc#1}\fi\fi\fi\fi}
20 \def\l{@l}
21 \def\L{@L}
```

Remarquons qu'après cette dernière modification les lettres non polonaises ayant des signes diacritiques seront traitées par TEX comme jusqu'à présent.

Polices des caractères polonais

Les problèmes concernant la qualité typographique des lettres polonaises avec des signes diacritiques et la division des mots nécessitent la création de nouvelles polices de caractères avec le logiciel METAFONT.

Toutes les lettres polonaises devraient se trouver dans ces polices, les minuscules ainsi que les majuscules. Même en adoptant, en tant que solution temporaire, la modification des polices déjà existantes il faut que toutes les 18 lettres accentuées polonaises (9 minuscules et 9 majuscules) s'y trouvent.

Remarquons que dans la police **Télétype** les signes diacritiques des lettres 'ł' et 'ż' avaient été remplacés par d'autres caractères. Par conséquent cette police n'est pas utile pour le polonais.

La division des mots polonais

Les problèmes concernant la division des mots avec des lettres accentuées ont été présentés dans les documents: [A], [D1], [D2], [RS] consacrés à l'adaptation du système TEX à des langues autres que l'anglais. La plupart de ces problèmes sont aussi valables pour la langue polonaise.

Aucun dictionnaire polonais n'indique les divisions des mots ce qui empêche l'utilisation du programme PATGEN ([L]). Il reste à écrire "à la main" les motifs requis par TEX, de même que J. Désarménien l'a fait pour le français.

Nous n'avons pas encore fini notre travail aux motifs pour le polonais, nous pouvons toutefois indiquer les difficultés essentielles.

Voilà les règles de division du polonais (d'après [S]):

(1.) **Critères.** La division des mots s'accomplit conformément à deux critères: phonétique (entre les syllabes) et morphologique (entre le préfixe et la racine ou, dans le cas des mots composés, entre les éléments de composition). Le critère morphologique est supérieur au phonétique.

(2.) **Il est interdit de diviser:**

- les mots composés d'une seule syllabe;
- les groupes de deux consonnes désignant un son: **ch, cz, dz, dź, dż, rz, sz,** par exemple: **mu-cha, o-cho-czy;**
- les groupes composés d'une consonne suivie de la lettre 'i' et d'une voyelle, par exemple: **cio-tka, ciot-ka;**
- les groupes composés des lettres 'dzi' suivies d'une voyelle, par exemple: **sie-dzieć.**

(3.) **Groupes de consonnes.** Ils se divisent de façon arbitraire, par exemple: **wa-rstwa, war-stwa, wars-twa, warst-wa.** La coupure juste après tout le groupe est la seule qui est interdite. Il faut toutefois diviser les groupes composés de deux consonnes identiques, par exemple: **pan-na, las-so.**

(4.) **Division entre le préfixe et la racine.** Conformément au supérieur critère morphologique la coupure comme **wez-brać,** par exemple, est la seule autorisée (préfixe "wez-", racine "brać"). D'autres coupures: **we-zbrać** ou **wezb-rać** sont interdites.

(5.) **Mots composés.** Il faut les diviser entre les éléments de composition, par exemple: **pół-noc** (*mi-nuit*). La division: **pó-łnoc** n'est pas correcte.

Après l'analyse des règles présentées ci-dessus nous remarquons deux problèmes essentiels: la division des groupes de consonnes et la division du mot entre le préfixe et la racine.

Le nombre de différents groupes de consonnes qui apparaissent dans les mots polonais est très grand. Il est pratiquement impossible de construire les motifs qui permettraient de trouver tous les points de division d'autant plus que personne, jusqu'à présent, n'a établi la liste complète de suites de consonnes possibles. La solution la plus simple est de diviser les mots juste après la voyelle, avant tout le groupe de consonnes. Selon la troisième règle présentée ci-dessus ce point de division est correcte. Il faut toutefois remarquer que les règles mentionnées, observées en Pologne depuis 1936, sont plus libérales que les précédentes formulées en 1918. L'essentielle différence concerne la division des groupes de consonnes (ici: la 3ème règle). La règle précédente avait désigné exactement les points de division et exclu, en particulier, la coupure entre la voyelle et le groupe de consonnes. N'étant plus valide, elle est profondément enracinée dans les esprits des Polonais pour qui la coupure comme **wa-rstwa** reste toujours "bizarre" et "à éviter".

L'autre problème concerne la division du mot entre le préfixe et la racine. Il y a environs trente préfixes d'origine slave: **bez-, na-, nad-, naj-, o-, ob-, od-, po-, pod-, prze-, przed-, we-, wes-, wez-,** etc. Il n'y a aucune règle pour distinguer automatiquement le préfixe d'un mot, par exemple il y a deux façons de décomposer le mot "nadrobić": **na-drobić** (*émietter*) et **nad-robić** (*faire qqch en plus*). Puisque le critère morphologique est supérieur au phonétique, la solution proposée ci-dessus de diviser le mot juste après une voyelle n'est pas satisfaisante. Il faut absolument exclure la division à l'intérieur d'un préfixe éventuel.

En adoptant la solution proposée nous éviterons les coupures fautives, mais beaucoup

de points de division correctes ne seront jamais trouvés. En outre il faut ajouter des motifs pour empêcher la coupure avant le groupe de consonnes qui termine le mot, par exemple: **-szcz**, **-rstw**, **-ństw**, **-ńcz**, **-rsz**.

Remarquons qu'une liste d'exception n'est pas utile pour la langue polonaise dans laquelle non seulement les verbes se conjuguent mais aussi les noms, les adjectifs et les pronoms se déclinent.

La typographie polonaise

Il n'existe pas de difficultés essentielles pour adapter TEX à la typographie polonaise. Une des différences concerne les guillemets qui ont la forme suivante: „ ". Les guillemets « » ne sont utilisés qu'à l'intérieur d'une autre citation, comme dans la phrase suivante:

Autor pisze: „przeczytałem «Hamleta» Szekspira w wieku 10 lat"
(*L'auteur dit: «j'ai lu «Hamlet» de Shakespeare à l'âge de 10 ans»*)

Les guillemets qui se trouvent au début d'une citation peuvent être définis à l'aide de la virgule:

```
\def\"{\leavevmode\hbox to.5em{,\hss,}}
```

Une autre différence concerne la façon de distinguer certains mots du texte. Assez souvent nous imprimons séparément les lettres d'un tel mot, par exemple:

"Wyraz t e k s t jest wyróżniony"
(*Le mot "tekst" est différencié*).

Mentionnons que dans la typographie polonaise, basée sur la tradition française, on mesure en points didot.

Conclusions

L'adaptation de TEX pour le polonais est possible bien qu'elle exige un certain effort. Parmi les problèmes présentés dans ce document celui concernant la division des mots est sûrement le plus difficile à résoudre.

Remarquons pour terminer que pour accomplir ce travail une profonde connaissance linguistique est nécessaire.

Remerciements

Je tiens à remercier Anna Borkowska de m'avoir aidée à rédiger en français ce document ainsi que mes collègues, Janusz S. Bień et Krzysztof Szafran, pour leurs conseils et leur amical intérêt.

Bibliographie

[A] Appelt (W.), *The Hyphenation of Non-English Words with TEX*, in: TEX for Scientific Documentation, Proc. of the First European Conference, Addison-Wesley, 1984, pp.61–65.

[D1] Désarménien (J.), *How to run TEX in French*, Stanford University, Report No. STAN-CS-1013, 1984.

[D2] Désarménien (J.), *The use of TEX in French: hyphenation and typographie*, in: TEX for Scientific Documentation, Proc. of the First European Conference, Addison-Wesley, 1984, pp.41–59.

[ISO] ISO, International Organisation for Standardization, *Information processing—Coded character sets for text communication*, Standard No. ISO 6937, 1983.

[K] Knuth (D.E.), *The TEXbook*, Addison-Wesley, Reading, Mass., 1984.

[L] Liang (F.M.), *Word Hy-phen-a-tion by Com-put-er*, Stanford University, Report No. STAN-CS-83-977, 1983.

[RS] Romberger (S.), Sundblad (Y.), *Adapting TEX to Languages that use Latin Alphabetic Characters*, in: TEX for Scientific Documentation, Proc. of the First European Conference, Addison-Wesley, 1984, pp.27–40.

[S] Szymczak (M.), *Słownik ortograficzny języka polskiego wraz z zasadami pisowni i interpunkcji*, PWN, Warszawa, 1981.