УДК 821.133.1.05+519.688
ББК 80/84

Е. И. Бойчук
Кандидат филологических наук

ALGORITHME DE L'ANALYSE INFORMATIQUE DU TEXTE FRANÇAIS

Le but essentiel de la présente recherche est une description d'une des nombreuses méthodes de traitement du texte littéraire. Plus particulièrement, différents aspects de la linguistique informatique permettant d'effectuer des requêtes et d'utiliser les données seront abordés. La réalisation de ces opérations est effectuée par le chercheur dans une perspective de communication. Le système de traitement du texte proposé est permet la recherche des caractéristiques rythmiques d'un texte littéraire.

Mots clés: rythme, texte français, programme de l'analyse du rythme, assonance, allitération, prose, poésie.

E. I. Boichuk
Ph.D. in philology

COMPUTER ALGORITHM ANALYSIS OF FRENCH TEXT

This article describes an attempt to create a special computer program for analyzing the rhythm of the French text. This analysis is based on the treatment of such means of rhythm as assonance, alliteration and quantity of rhythmical groups' syllables. The program analyses different types of French text: poetry and prose.

Key words: rhythm, French text, program of rhythm analysis, assonance, alliteration, prose, poetry.

La linguistique informatique moderne propose une grande variété de programmes permettant de réaliser l'analyse linguistique complexe de textes en russe, en anglais, en allemand et en français (POLYGLOTE, Russian Context Optimizer, Яndex.Server, Link Grammar Parser, Hudlomer … etc.) [1,2,3]. Les modules d'analyse linguistique comprennent la morphologie, la syntaxe, la sémantique (y compris les dictionnaires raisonné-combinatoires). Il existe un module de classification sémantique des textes. Mais le plus souvent l'analyse du texte au niveau indiqué se limite à l'utilisation de synonymes thématiques, à la recherche par mots-clés et à l'indication de la fréquence des mots dans le texte.
Les deux intérêts principaux de ces programmes [1] sont les suivants :

- la possibilité de déterminer l'appartenance du texte à tel ou tel auteur. Le programme est en mesure de différencier le style individuel de chaque auteur et d'indiquer la proximité d'un texte en référence à un étalon choisi parmi les auteurs de la science fiction russe. Le texte d'entrée est analysé et le programme donne le nom de trois écrivains-auteurs potentiels. En outre, pour chaque auteur, le programme propose trois œuvres les plus proches du style du

[1] Programmes Linguoanalysator par D. Khmélev, Hudlomer par L. Délitsin [2,3]

texte d'entrée.

- l'identification du style fonctionnel du discours. Une classification automatique des styles fonctionnels du texte est effectuée (populaire, littéraire, périodique, commercial, scientifique …etc.) à partir de l'analyse de la quantité des mots.

Le programme Rhymes par N. Kétsaris est également d'un très grand intérêt. Ayant comme base le dictionnaire des rimes de A. Zalizniak, ce programme permet d'analyser les rimes russes en réalisant la comparaison phonétique des mots selon leur accent. De plus, ce programme trouve des synonymes et antonymes.

Mais aucun programme ne propose l'analyse complexe des caractéristiques rythmiques du texte littéraire.

Dans cet article, nous proposons la description d'une première étape de travail du programme informatique linguistique permettant d'analyser le rythme de la prose. Le programme effectue l'analyse syntaxique (parsing) d'un texte d'entrée, c'est-à-dire qu'il transforme le texte en système de données qui représente la structure syntaxique (notion informatique) d'une succession d'entrées et qui convient au traitement suivant.

Les conditions du traitement du texte sont déterminées par les règles de prononciation et de division en syllabes dans la langue française. À la création de ce programme, on utilisait les langues de programmation suivantes: QML (pour interface), JavaScript (pour la logique liée à l'interface), HTML (pour une présentation du texte), C++ и Qt (pour une création d'une partie fonctionnelle du programme).

L'analyse du texte prosaïque littéraire s'effectue par la recherche des moyens phonétiques, lexico-stylistiques et grammaticaux de rythmisation du texte.

Le screenshort nous montre que l'aspect phonétique est présenté par les moyens suivants: la longueur des groups rythmiques (qui révèle la quantité égale des syllabes dans les mots et dans les groups rythmiques), l'allitération, l'assonance, la rime. La case « rime » est subdivisée en différents types de rime.

L'aspect lexico-stylistique est présenté par les moyens suivants: anaphore, épiphore, répétitions de différents types, la fréquence des antonymes et des synonymes.

Exemple de l'interface du programme:

L'analyse de l'aspect grammatical est basée sur l'analyse morphématique des mots, sur une identification des termes homogènes (marqué le plus souvent par la ponctuation), sur la fréquence des propositions ayant différents buts communicatifs (question, exclamation, réticence) et sur les particularités de l'ordre des mots dans la proposition (inversion, chiasme …etc.). Le programme révèle la fréquence des moyens indiqués dans le texte, ce qui permet de suivre la périodicité de leur apparition (celle-ci étant la marque essentielle du rythme dans toutes ses manifestations). La fréquence de tel ou tel moyen rythmique nous permet de déterminer les particularités du style individuel de chaque auteur.

L'analyse du texte est dynamique puisque le programme permet de changer ses paramètres. Il est à noter que le programme proposé est auxiliaire dans la procédure d'analyse rythmique du texte. Cela s'explique par le fait que l'analyse complexe de tous les paramètres de rythmisation du texte aux différents niveaux langagiers comprend plusieurs étapes (y compris la perception individuelle du texte par le lecteur et l'analyse purement phonétique réalisée par d'autres programmes).

De cette façon, on ignore dans le programme présent certains moyens de rythme (phonétiques surtout) qui dépendent des conditions subjectives telles que la manière individuelle de lire, la vitesse de lecture, le timbre et l'intensité de la voix du lecteur, les pauses, les césures …etc. Tous ces moyens sont à analyser à l'aide des programmes PRAAT et SpeechAnalyzer.

Dans le cadre de l'aspect phonétique, la détermination de la plupart des moyens du rythme mentionnés est possible essentiellement grâce à la division du

texte en syllabes. L'élaboration des règles du programme permettant la division du texte en syllabes représente une très grande difficulté (ce qui s'explique en premier lieu par le fait que l'ordinateur ne voit que les signes). La division du texte en syllabes par le programme est basée sur les règles de lecture : la combinaison des sons et des diphtongues (classifiés par le programme comme des syllabes : iè, ieu, eai …etc.) et les différentes positions des consonnes (dont la prononciation dépend des sons qui l'entourent).

Les règles de prononciation du « e »muet (qui dépend, entre autres, du style et de la vitesse du discours, ainsi que de la manière individuelle de lire) sont présentées dans le programme sous la forme d'une règle commune de la position de « e » muet après deux consonnes devant la troisième. Cependant, il est fait exception des cas de la position jointe de deux mots et des combinaisons suivantes: n+consonne+e+blanc, m+ consonne +e + blanc, par exemple, nbe, nce, nde et d'autres.

Les cas de la prononciation du « e » muet dans les mots à une syllabe, tels que le, me, te, ce, se, ne, de sont présentés dans le programme par le schéma suivant: blanc+consonne+e: be, ce, de, fe, ge, je, ke, le, me, ne, pe, re, se, te, ve, ze, ce qui signifie que leur position est initiale.

Pour différencier les cas de la prononciation ou de la non-prononciation de « *–ent* » à la fin des mots, les règles et les exceptions suivantes ont été précisées comme suit :

- « *-ent* » à la fin des mots ne se prononce pas, à l'exception des mots en « *-ment »* et en « *–scent* » où la terminaison se prononce et représente une syllabe, ainsi que les mots en « *–ent »* qui ont deux syllabes ou plus (par exemple : *argent, abat-vent, abstinent, accent, accident, adhérent, adjacent, afférent …*etc) ;
- les verbes de la 3-ième personne du pluriel en « *–ment »* (qui ne se prononce pas) sont relevés afin de ne pas les confondre avec les cas de l'emploi des adverbes en « *–ment »* prononcé (p.ex. affament, affirment, aiment, allument, animent, arment, …etc).

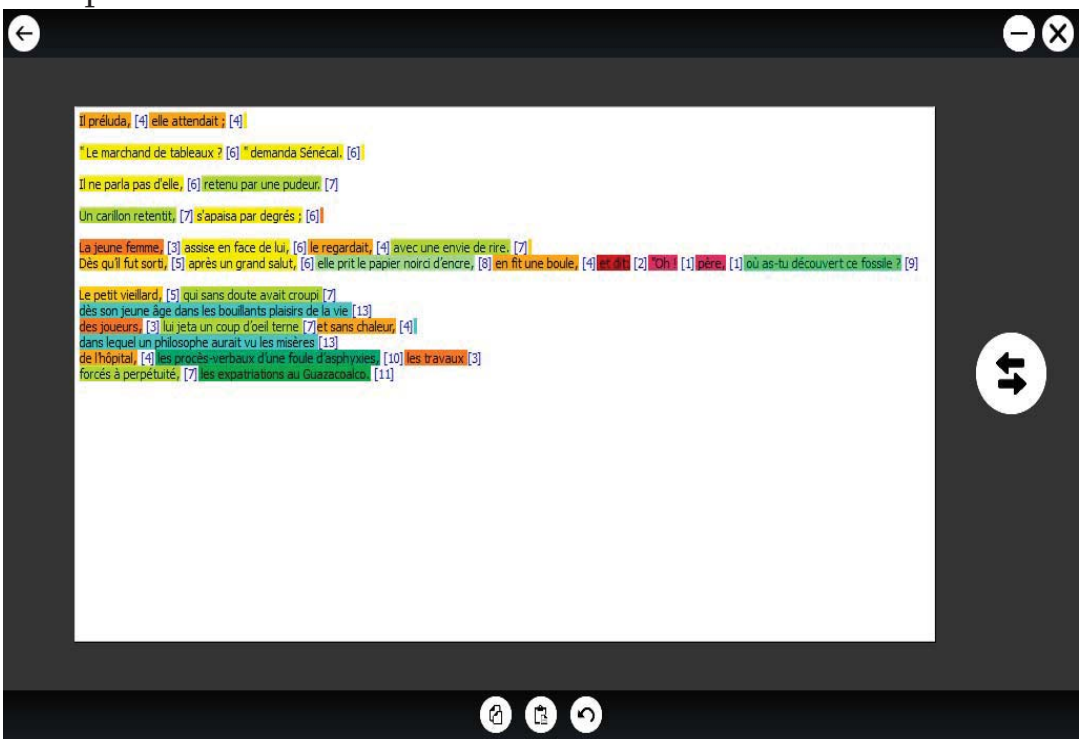
Le programme inclut les cas des abréviations telles que M. (Monsieur), Mme (Madame) …etc., qui, le plus souvent, représentent plus de deux syllabes à la différence de leur équivalent écrit.

Sur cette base des règles de la division du texte en syllabes (et quelques autres), est créé l'algorithme de la détermination des unités rythmiques du texte. Il est difficile de soumettre aux règles phonétiques communes la détermination des unités rythmiques dans le cadre du programme proposé. Cela s'explique par l'impossibilité pour le programme de déterminer les parties du discours et les termes de proposition. Pour cette raison, la division en unités rythmiques du texte

s'effectue sur la base de la ponctuation et sur l'emploi des conjonctions de coordination et de subordination.

Bien sûr cette division n'est pas exacte, mais elle permet cependant la division en unités rythmiques ce qui facilite beaucoup la procédure de traitement du texte. Cette division et la possibilité de déterminer la quantité de syllabes permet de révéler les unités comprenant une quantité de syllabes identique (ce qui est étroitement lié au rythme du texte : les unités équivalentes du point de vue de leur composition syllabique ou les unités représentants une succession de syllabes avec une différence d'une ou deux syllabes sont classées comme les plus rythmiques).

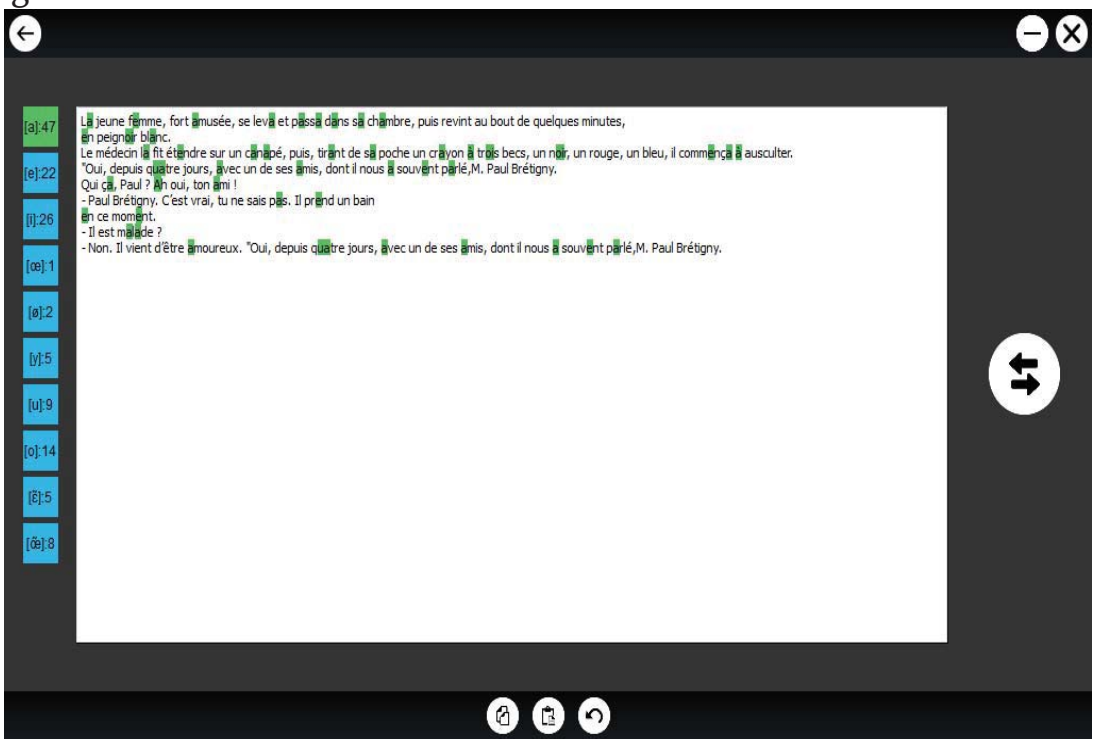
Citons l'exemple de la division du texte en unités rythmiques avec une indication de la quantité des syllabes à l'intérieur de chaque unité. À chaque quantité correspond sa couleur :



La couleur permet d'évaluer visuellement le texte et d'y trouver facilement les unités ayant le même nombre de syllabes ou un nombre approchant (c'est à dire une différence de deux ou trois syllabes). Le plus souvent la comparaison des unités rythmiques identiques est complétée par d'autres moyens : rythmiques (rime, allitération, assonance), phonétiques, lexicaux et grammaticaux (par ex. la répétition et l'énumération).

Dans l'objectif de déterminer les consonnes et les voyelles du fragment de texte proposé, ont été formulées des règles permettant au programme de différencier les consonnes et les voyelles prononcées et non prononcées en fonction de leur position et de leurs combinaisons. Afin de faciliter le travail sur

de longs extraits, le programme permet de visualiser la fréquence d'une consonne en la surlignant dans le texte étudié.



Dans le cadre d'une recherche sur les assonances, on précise non seulement la répétition des voyelles accentuées mais aussi celle des voyelles non accentuées. Cette recherche est intéressante au regard de l'étude du rythme de la prose au travers des assonances en tant que phénomène littéraire général et non seulement poétique.

Dans ce programme informatique, on révèle des cas d'assonance comprenant parfois un groupe de sons qui ont des ressemblances du point de vue phonétique, par exemple, le signe [a] représente le groupe de sons suivants [a, a□, wa, ua, ya, aj, ja], le signe [e] représente les sons [ej, je, □, j□, □j, y□, u□] etc.

La création de ce programme pour l'analyse du rythme du texte est avant tout nécessaire pour faciliter en quelque sorte le travail d'un chercheur dans le décompte des syllabes en vue de révéler leur égalité quantitative. Il permet une détermination automatique des répétitions à tous les niveaux de langue: phonétique, lexical, stylistique et grammatical. Tout cela est nécessaire pour améliorer la productivité du travail sur le texte en économisant le temps du chercheur et en l'aidant à éviter les erreurs de toutes sortes liées au facteur humain.

Le matériel de cet article représente une étape initiale du travail sur la création du programme informatique. Tous les changements et toutes les démarches ultérieures seront décrits dans un cycle d'articles consacrés à cette question.

ЛИТЕРАТУРА

1. Официальная страница приложения Link Grammar Parser [Электронный ресурс]. URL: http://www.link.cs.cmu.edu/link/ (дата обращения 10.05.2013).
2. Хмелев Д. Официальная страница приложения ЛингвоАнализатор [Электронный ресурс]. URL: http://www.rusf.ru.books/analysis/ (дата обращения 10.05.2013).
3. Официальная страница приложения Худломер // Тенета-Ринет 2000 Худломер [Электронный ресурс]. URL: http://www.teneta-rinet.ru/2001/hudlomer/ (дата обращения 10.05.2013).

REFERENCES

1. *Official page of the application's Link Grammar Parser*. Available at: http://www.link.cs.cmu.edu/link/ (accessed 10 March 2013).
2. *Khmelev D. Official page of the application LingvoAnalizator*. Available at: http://www.rusf.ru.books/analysis/ (accessed 10 March 2013).
3. *Official page of the application Hudlomer // Teneta-Rinet 2000 Hudlomer*. Available at: http://www.teneta-rinet.ru/2001/hudlomer/ (accessed 10 March 2013).

Информация об авторе

Бойчук Елена Игоревна (Российская Федерация, Ярославль) – Кандидат филологических наук. Старший преподаватель. Ярославский государственный педагогический университет. E-mail: Elena-boychouk@rambler.ru

Information about the author

Boichuk Elena Igorevna (Russian Federation, Yaroslavl) – Ph.D. in philology. Senior lecturer. Yaroslavl state pedagogical university. E-mail: Elena-boychouk@rambler.ru