BESOINS LEXICAUX A LA LUMIERE DE L'ANALYSE STATISTIQUE DU CORPUS DE TEXTES DU PROJET "BREF" - LE LEXIQUE "BDLEX" DU FRANCAIS ECRIT ET ORAL.

I. FERRANE, M. de CALMES, D. COTTO, J.M. PECATTE, G. PERENNOU.

IRIT - Université Paul Sabatier
118, route de Narbonne
31062 TOULOUSE Cedex - FRANCE

ABSTRACT

In this paper, we describe lexical needs for spoken and written French surface processing, like automatic text correction, speech recognition and synthesis.

We present statistical observations made on a vocabulary compiled from real texts like articles. These texts have been used for building a recorded speech database called BREF. Developed by the Limsi, within the research group GDR-PRC CHM (Groupe De Recherche - Programme de Recherches Concertées, Communication Homme-Machine — Research Group - Concerted Research Program, Man Machine Communication), this database is intended for dictation machine development and assessment.

In this study, the informations available in our lexical database BDLEX (Base de Données LEXicales - Lexical Database) are used as reference materials. Belonging to the same research group than BREF, BDLEX has been developed for spoken and written French. Its purpose is to create, organize and provide lexical materials intended for automatic speech and text processing.

Lexical covering takes an important part in such system assessment. Our first purpose is to value the rate of lexical covering that a 50,000 word lexicon can reach.

By comparison between the vocabulary provided (LexBref, composed of 84,900 items, mainly distinct inflected forms) and the forms generated from BDLEX, we obtain about 62% of known forms, taking in account some acronyms and abbreviations.

Then, we approach the unexpected word question looking into the 38% of left forms. Among them we can find numeration, neologisms, foreign words and proper names, as well as other acronyms and abbreviations. So, to obtain a large text covering, a lexical component must take in account all these kinds of words and must be fault tolerant, particularly with typographic faults.

Last, we give a general description of the BDLEX project, specially of its lexical content. We describe some lexical data recently inserted in BDLEX according to the observations made on real texts. It concerns more particularly the lexical item representation using phonograms (i.e. letters/sounds associations), informations about acronyms and abbreviations as well as morphological knowledge about derivative words. We also present a set of linguistic tools connected to BDLEX and working on the phonological, orthographical and morphosyntactical levels.

1. Introduction

Dans le domaine des *Industries de la Langue* les lexiques électroniques occupent une place importante. Dictionnaires et encyclopédies sont maintenant disponibles sous DOC ; pour le français, on peut citer entre autres le dictionnaire Zyzomis d'Hachette et le Robert électronique. Les systèmes de traitement de texte actuels disposent dans leur environnement, de lexiques pouvant être consultés pour vérifier l'orthographe ou la conjugaison d'un mot, pour la recherche de synonymes, etc. Les correcteurs automatiques font aussi appel à des lexiques.

Actuellement, tous ces matériaux lexicaux sont encore loin de satisfaire les besoins du traitement automatique de la parole et des textes. Ils sont insuffisants lorsqu'on aborde des traitements linguistiques mettant en jeu une analyse sémantique et syntaxique approfondie. Ils sont également inadaptés pour des traitements de surface tels que ceux qui interviennent dans la correction orthographique, la synthèse de la parole à partir de texte, et la dictée vocale. C'est pourquoi différentes équipes de recherche ont entrepris de développer leurs propres lexiques.

Dans cet article, nous décrivons les observations statistiques faites sur le vocabulaire extrait d'un corpus de textes réels constitués d'articles de journaux. Ceux-ci sont utilisés dans la base de données BREF destinée au développement et à l'évaluation des *machines à dicter*.

Cette étude met en évidence différents besoins en matériaux lexicaux. Elle montre aussi la nécessité de développer tout un ensemble de procédures pour traiter les inattendus qui, quelle que soit l'étendue des matériaux lexicaux utilisés, surviennent très fréquemment dans les textes usuels.

Nous donnons ensuite une description de la base de données lexicales du français écrit et oral, BDLEX, dont l'objectif est de créer, d'organiser et de distribuer des matériaux lexicaux destinés au traitement automatique de la parole et des textes [Pérennou, 91].

Les projets BREF et BDLEX sont développés dans le cadre du GDR-PRC Communication Homme-Machine —groupe de recherches coordonnées du Ministère de la Recherche et de la Technologie, et du Centre National de la Recherche Scientifique.

2. Couverture de textes réels

L'accès au lexique joue un rôle crucial dans des applications comme la correction automatique, et le traitement automatique de la parole. Si un mot est inconnu du lexique, le système est mis en échec sans qu'il le sache toujours. En effet, l'accès étant tolérant aux fautes ou aux imprécisions de reconnaissance, il se trouvera toujours un mot plus ou moins proche pour remplacer celui qui est observé.

Le *taux de couverture* lexicale, ou proportion des mots d'un texte connus du lexique, est donc un des critères importants pour l'évaluation du niveau de performance des systèmes de ce type.

2.1. Résultats classiques

Différentes études statistiques ont déjà été effectuées sur ce point. On peut citer pour illustration les résultats obtenus par P. Guiraud [Guiraud, 59]. Ceux-ci établissent que 100 *mots* bien choisis assurent un taux de couverture d'environ 60%, tandis que 1 000 mots couvrent 85% et 4 000 mots 97,5%. La couverture des 2,5% restant peut être assurée par un corpus de 40 000 mots. Pour un complément sur ce type d'étude, on peut se reporter à [Catach, 84].

En terme de *formes de mots* le taux de couverture est dépendant de la langue considérée. En effet, des statistiques basées sur l'étude de corpus constitués à partir de lettres d'affaire et établies par Averbuch pour l'anglais [Averbuch, 87] et Mérialdo pour le français [Mérialdo, 88], il ressort que le taux de couverture assuré en anglais par un lexique de 20 000 formes, soit environ 97,5%, est équivalent à celui assuré, en français, par un lexique 10 fois plus important [Pérennou, 90].

Ces taux de couverture relativement élevés sont obtenus à partir de corpus où chaque forme est pondérée par sa fréquence d'apparition dans les textes considérés. Ainsi, le pourcentage des formes rejetées, généralement des formes rares ou très spécialisées, reste très faible.

2.2. Le corpus BREF

Les résultats ci-dessus sont faussés dès que le corpus étudié n'est pondéré d'aucune information fréquentielle et qu'il aborde des domaines aussi vastes que variés : *finance, politique, géographie, culture, spectacle, ...* Tel est le cas du corpus BREF, établi à partir d'articles de journaux.

2.2.1. Etat des matériaux étudiés

Nous donnons ici des statistiques portant sur le lexique des formes fléchies extraites d'un corpus de textes constitué en vue de la création d'une base de données de parole enregistrée : la base de données BREF. Celle-ci est destinée à l'évaluation de systèmes de reconnaissance de grands vocabulaires. Cette base est développée au LIMSI dans le cadre du GDR-PRC Communication Homme-Machine [Lamel, 91].

Nous nous intéresserons plus particulièrement, à la composition du vocabulaire apparaissant dans les textes du corpus BREF. Celui-ci nous a été communiqué sous la forme d'une liste de 84 900 mots, que nous appellerons par la suite LexBref. Chaque forme est représentée en lettres minuscules ; la distinction entre nom propre et nom commun est donc complètement perdue. Il en va de même pour les repérages typographiques conventionnels des sigles, des abréviations et de certains mots composés, les signes non alphanumériques ayant été effacés.

2.2.2. Taux de couverture lexicale

Nous avons procédé à la comparaison des formes de LexBref avec celles que nous avons générées à partir de notre base de données lexicales BDLEX.

La version BDLEX-1 comporte 23 000 entrées et permet d'accéder à un corpus comptant environ 270 000 formes fléchies. L'extension de ce lexique à 50 000 entrées constitue la version BDLEX-2.

Dans la figure 1 nous avons représenté le pourcentage de formes de LexBref qui ont été trouvées dans BDLEX-1 et BDLEX-2.

Nbre de formes trouvées / Corpus de référence	Recherche directe (1)	Fautes d'accent (2)	Pourcentage
BDLEX-1 (23 000 entrées)	40 931	1 542	50%
BDLEX-2 * (50 000 entrées)	9 415	183	11, 3%
LexBref par rapport à BDLEX-2	50 346	1725	**61,3%**

* : Complémentaire de BDLEX-1 par rapport à BDLEX-2

Fig.1- Résultats obtenus par comparaison de LexBref aux formes générées à partir de BDLEX-1 et BDLEX-2.

La colonne (1) donne les résultats obtenus à partir d'une recherche directe qui a permis de reconnaître le plus grand nombre de formes. Nous avons ensuite supposé que d'autres formes pouvaient être trouvées, moyennant la correction d'une faute d'accent.

Les résultats portés en colonne (2) de ce tableau ont été obtenus en utilisant le correcteur orthographique et typographique VORTEX [Pérennou, 86, 91], [Pécatte, 90].

Pour affiner l'analyse, nous nous sommes intéressés aux sigles et aux abréviations qui pouvaient figurer dans ce corpus. Pour cela nous l'avons comparé à une liste de sigles, fournie par M. Plénat dans le cadre du GDR-PRC CHM, et à une liste d'abréviations. Les résultats de cette recherche sont portés dans la figure 2 ci-dessous.

Corpus de référence	Formes trouvées	Pourcentage
Sigles de Plénat (1 000 sigles)	380	0,45%
Abréviations (280 abrév.)	70	0,08%
Sigles et abréviations de LexBref reconnus	450	**0,53%**

Fig.2 - Résultats obtenus par comparaison de LexBref à une liste de sigles et une liste d'abréviations.

En observant les figures 1 et 2, on constate qu'un ensemble important de formes de LexBref, environ 38%, n'ont pas été identifiées.

L'étude de la structure de ce lexique résiduel, LexR, peut être un moyen de préciser les besoins en matériaux et outils lexicaux en vue d'augmenter la proportion de formes reconnues.

2.2.3. Analyse du corpus résiduel LexR

Pour déterminer les différents types de mots inattendus et leur proportion dans le corpus initial, LexBref, nous avons appliqué diverses procédures que l'on peut qualifier de non lexicales puisqu'elles ne font intervenir aucune consultation de lexique.

◆ *Formes numériques*
Dans le corpus de BREF figurent des nombres cardinaux et ordinaux, exprimés en chiffres arabes (*1991, 200e*,...), ou en chiffres romains (*XVIIe, XV*, ...). On trouve également des nombres exprimant un pourcentage (*5%, 75%*, ...). Ces unités représentent environ 1,5% du corpus LexBref.

♦ *Mots étrangers et Noms propres*
La grande diversité des sujets abordés dans un quotidien et la portée internationale des faits relatés font que de nombreux mots étrangers apparaissent dans les textes (*amnesty, congress, perestroïka, glasnost...*).

Une analyse basée sur des critères particuliers, comme l'étude des finales de mots n'appartenant pas à la langue française, mais fréquentes dans d'autres langues ou encore caractéristiques de noms propres (*-y, -ss, -ski, -nn, -ff, -v, -oux* ...), nous a permis de distinguer un premier groupe de mots d'origine étrangère (*academy, congress, ...*) représentant environ 15,5% du corpus initial. Ce pourcentage inclut également les mots pouvant correspondre à des noms propres français ou étrangers (*Châteauroux, Einstein, Gorbatchev, Stravinski, Bonn, ...*).

♦ *Néologismes*
La création lexicale est un phénomène linguistique fréquent dans les médias : *groupuscularisation, zapping, ...* Beaucoup de mots sont créés à partir de noms propres issus des milieux politique, artistique ou littéraire : *antigaulliste, mitterrandien, maccarthysme, hitchcockien, nabokovien, ...*

La plupart sont produits par dérivation mais il existe de nombreux exemples obtenus par composition, comme par exemple *vrai-faux* (*vraie-fausse facture, vrai-faux passeport, ...*). Quelques néologismes sont obtenus selon des procédés plus marginaux comme le verlan (*ripoux, chébran ...*) et les mots-valises (*motel, confipote ...*).

Nous avons examiné les néologismes dérivationnels construits de manière régulière, par application de règles dérivationnelles sur un mot de la langue ou un nom propre —[Ferrané, 91] pour le traitement morphologique dans BDLEX.

A partir d'une liste d'affixes productifs comme les préfixes *anti-, dé-, inter-, néo-, sur-,...* et les suffixes *-ation, -ien, -isme, -iste, -is(er), -ité, -ment, ...*, nous avons procédé à une recherche dans LexR qui nous a permis d'estimer respectivement à 0,5% et 5,5% les mots de LexBref initialement rejetés et susceptibles d'être analysés dans un deuxième temps comme préfixés ou bien suffixés —lors du traitement des suffixes nous avons pris en compte les variations flexionnelles (par exemple les mots comme *hitchcockiennes* sont détectés).

La figure 3 ci-dessous reprend les différentes estimations faites dans cette seconde phase d'étude du corpus LexBref.

Critères de recherche	Exemples de formes sélectionnées	Pourcentage par rapport à LexBref
Nombres	*1991, XXVIIe,...*	1,5%
Mots étrangers et noms propres	*congress, amnesty, roscoff, gorbatchev*	15,5%
Mots supposés préfixés	*interafricain, néobaroque,...*	0,5%
Mots supposés suffixés	*hitchcockiennes, groupuscularisation, zapping, ...*	5,5%
Mots extraits de LexBref par procédure non lexicale		23%

***Fig.3** - Analyse du corpus résiduel LexR.*

Parmi les 15% restant, on trouve notamment des sigles qui n'ont pas été répertoriés dans la liste de référence que nous avions à notre disposition (TF1, ADN,...).

On trouve encore des néologismes, des noms propres et des mots étrangers d'emprunt pour lesquels aucune procédure non lexicale n'a pu être appliquée.

Enfin, on rencontre des mots incorrectement écrits (le plus souvent à la suite d'une faute typographique) et d'autres qui seraient reconnus par un lexique plus étendu que BDLEX-2.

3. Le projet BDLEX

Le projet BDLEX regroupe un ensemble de matériaux lexicaux et d'outils linguistiques.

3.1. Matériaux lexicaux

Selon l'application visée, différents lexiques peuvent être dérivés de BDLEX. La version BDLEX-1 est organisée en base de données relationnelle gérée par le SGBD ORACLE sur station SUN. Les informations disponibles permettent d'aborder nombre d'applications en traitement automatique de la parole et des textes.

Ce sont :
- la graphie accentuée,
- la transcription phonologique incluant les frontières de syllabe et de pied,
- la représentation en phonogrammes mettant en évidence les associations lettres/sons,
- la morphologie flexionnelle : conjugaison des verbes, flexion des noms et adjectifs,

- la morphosyntaxe,
- des indices de fréquence d'apparition dans les textes,
- la dérivation et la composition.

Le corpus de BDLEX-1 de 23 000 entrées a été étendu à 50 000 entrées, en particulier en ce qui concerne les informations graphiques et morphosyntaxiques. BDLEX a déjà été décrit dans [Pérennou, 90] et [Ferrané, 91].

Nous ne détaillons ici que les informations introduites plus récemment en fonction des observations effectuées sur des textes réels, tels que le corpus de BREF décrit dans le paragraphe 2. Il s'agit notamment de la représentation en phonogrammes des entrées lexicales, des informations relatives aux sigles et aux abréviations ainsi qu'à la morphologie dérivationnelle

- *Phonogrammes*

Ce sont des associations élémentaires de lettres et de sons —voir par exemple [Catach, 78]. Ils jouent un rôle important en correction automatique et en synthèse de la parole à partir de texte.

Chaque entrée lexicale de BDLEX dispose d'une représentation en phonogrammes, comme cela est illustré dans le figure 4. Les associations lettres/sons ont été obtenues par un alignement entre la graphie accentuée et la représentation phonologique de l'entrée.

GR_AC	PH_S	F	CS	PHONOGRAMMES
axe	Aks	ə	N	(ə,A)(x,ks)(ə,ə)
balut	/bA/y		N	(b,b)(a,A)(h,€)(u,y) (t,€)
hache	/*Aʃ	ə	N	(h,*)(a,A)(ch,ʃ) (ə,ə)
skate	/skEjt	ə	N	(s,s)(k,k)(a,E)(t,t)(ə,ə)

*Fig.4 - Extrait de BDLEX. : représentation en phonogrammes —^ : lettre ne correspondant à aucun son; * : h aspiré ; / : frontière syllabique.*

On compte, en français, une centaine de phonogrammes de base. Cependant, lorsqu'on prend en compte des mots d'emprunt étrangers, ce nombre augmente considérablement : 450 phonogrammes recensés pour les 23 000 entrées de BDLEX-1.

- *Sigles et abréviations*

Des travaux, à l'IRIT, portant sur le développement d'outils linguistiques ont déjà donné lieu à la conception d'un noyau lexical de sigles et d'abréviations.

Comme cela est représenté dans la figure 5, un sigle dispose d'informations concernant la graphie, la phonologie et la morphosyntaxe.

GR_AC	GR Ext	PH_S	F	CS	CF	GN
c.-à-d	c'est-à-dire	/sE/tA/dir	ə	A		00
F	franc	/frɑ̃		N	Mn	01
M.	monsieur	/mə/sjŒ		N	MS	00
MM.	messieurs	/mE/sjŒ	z"	N	MP	00
kg	kilogramme	/kI/10		N	Mj	00
kilo	kilogramme	/kI/10		N	Mn	01
S.V.P.	s'il vous plaît	/sil/vu/plE		A		00

Fig.5 - Extrait de BDLEX : Sigles et abréviations.

Les travaux sur les sigles sont développés en liaison avec M. Plénat [Plénat,91].

- *Morphologie dérivationnelle*

L'introduction dans BDLEX d'un ensemble de connaissances morphologiques dérivationnelles doit permettre non-seulement de lier entre elles certaines entrées de BDLEX, mais également de procéder à l'analyse morphologique de néologismes dérivationnels.

En effet, bon nombre de ceux qui apparaissent dans les textes réels sont inconnus du lexique. Cependant, ils peuvent généralement être rattachés à une entrée lexicale : l'entrée dont ils dérivent (ou *base*). Ainsi, en appliquant la règle associée au suffixe *-ment*, formateur de noms masculin à partir d'une base verbale, on peut lier la forme *aboutissement*, trouvée dans LexBref et non répertoriée dans BDLEX-2, à l'entrée *aboutir*, verbe connu du lexique.

A l'heure actuelle 68 préfixes et 107 suffixes, essentiellement des suffixes à base verbale ou bien formateurs de verbes, ont été répertoriés dans BDLEX [Ferrané, 91].

3.2 Outils linguistiques

Dans le cadre de BDLEX, nous avons développé différents outils linguistiques utiles pour la création et l'utilisation des matériaux lexicaux. Ceux-ci opèrent aux différents niveaux de la structure textuelle ou du message vocal.

Sont disponibles actuellement :

- *Géner*, le générateur de formes fléchies,
- *Amflex*, l'analyseur morphologique flexionnel,
- *VortexPlus*, le correcteur orthographique qui peut également être employé comme lemmatiseur tolérant aux fautes (utilisable avec BDLEX-1 ou BDLEX-2),

- différentes fonctions d'accès particulières utilisées par les psychoneurolinguistes,
- *GEPH*, un système expert en phonologie [Tihoni, 91],
- *TEXOR* pour le prétraitement linguistique des textes en vue de la synthèse à partir de texte,
- *ASYSE*, un générateur d'analyseur linguistique à base d'ATN et d'opérations sur les schémas, en particulier l'unification.

4. Conclusion

Comme nous l'avons illustré à partir du lexique extrait du corpus de BREF, le traitement automatique de la parole et des textes requiert un ensemble de matériaux lexicaux importants et variés, incluant les sigles et les abréviations, ainsi que des éléments de morphologie. Ils doivent être complétés d'outils linguistiques améliorant le traitement (correction, analyse morphologique, ...).

Ceux-ci doivent, non seulement, prendre en compte les besoins classiques aux plans morphologique et syntaxique, mais encore ceux plus particuliers relatifs aux inattendus variés qui apparaissent dans les textes et les messages vocaux.

Le projet BDLEX s'est développé dans ce contexte, avec pour objectif de rendre disponibles différents matériaux et outils linguistiques. C'est ce qui a été partiellement réalisé dans le cadre du GDR-PRC Communication Homme-Machine.

Les extensions en cours visent à l'enrichissement du vocabulaire et au développement des traitements phonologiques et morphologiques répondant aux besoins mis en évidence dans cet article.

5. Bibliographie

[Averbuch, 87] A. Averbuch et 21 co-auteurs, *Experiment with the TANGORA 20,000 Word Speech Recognizer*, CH2396-0/37/0000-0701,1987.

[Catach, 78] N. Catach, *L'orthographe*, Que sais-je ?, Presses universitaires de France, 1978.

[Catach, 84] N. Catach, *Les listes orthographiques de base du français (LOB)*, Nathan Recherche, 1984.

[Ferrané, 91] I. Ferrané, *Base de données et de connaissances lexicales morphosyntaxiques*, Thèse de doctorat de l'Université Paul Sabatier, Toulouse III, 1991.

[Guiraud, 59] P. Guiraud, *Problèmes et méthodes de la statistique linguistique*, D. Reidel Pub. Company, 1959.

[Lamel, 91] L.F. Lamel, J.L. Gauvain, M. Eskénazi, *BREF, a Large Vocabulary Spoken Corpus for French*, Proceedings of EUROSPEECH 91, Genova, 24-26 September 1991, Vol.2, pp. 505-508.

[Mérialdo, 88] B. Mérialdo, *Multi-Level Decoding for Very Large Size Dictionary Speech Recognition*, IBM Journal of R&D, 1988.

[Pécatte, 92] J.M. Pécatte, *Tolérance aux fautes dans les interfaces homme-machine*, Thèse de doctorat de l'Université Paul Sabatier, Toulouse III, 1992.

[Pérennou, 86] G. Pérennou, *La vérification et la correction automatique des textes : le système VORTEX*, Technique et Science Informatique, n°4, 1986, pp. 285-305.

[Pérennou, 90] G. Pérennou, *Le projet BDLEX de base de données et de connaissances lexicales et phonologiques*, Premières journées du GDR-PRC Communication Homme-Machine, EC2 Editeur, Paris, 24-25 Novembre 1988, pp. 81-111.

[Pérennou, 91] G. Pérennou, D. Cotto, M. de Calmès, I. Ferrané, J.M. Pécatte, J. Tihoni, *Composantes phonologique et orthographique de BDLEX*, Deuxièmes journées du GDR-PRC Communication Homme-Machine, EC2 Editeur, Toulouse, 29-30 Janvier 1991, pp. 351-362.

[Plénat, 91] M. Plénat, *Vers d'une phonémisation des sigles*, Deuxièmes journées du GDR-PRC Communication Homme-Machine, EC2 Editeur, Toulouse, 29-30 Janvier 1991, pp. 363-371.

[Tihoni, 91] J. Tihoni, G. Pérennou, *Phonotypical Transcription Through the GEPH Expert System*, Proceedings of EUROSPEECH 91, 2nd European Conf. on Speech Com. and Tech., Genova, Italy, pp.767-770, 1991.