LA RECHERCHE DE TERMES DANS LES TEXTES SPÉCIALISÉS :

INTERNET ET LA CONSTITUTION DE CORPUS DE RÉFÉRENCE

ESIT

9 JANVIER 2002

Marc Van Campenhoudt

Centre de recherche Termisti

Institut supérieur de traducteurs et interprètes

Bruxelles

1 INTRODUCTION PROBLÉMATIQUE GÉNÉRALE

1.1 Rompre avec une tradition de compilation des dictionnaires

1.2 Quelques définitions de la notion de corpus

> « La réunion d'un grand nombre de textes indexés constitue une **corpus**, à l'intérieur duquel on se propose d'étudier et de quantifier certains faits lexicaux, syntaxiques, etc. Le corpus comprend en général des divisions ou **sous-corpus**, qui la plupart du temps ont une unité propre (chronologique, stylistique, etc. » (MULLER 1973 : 16.)

> « Ensemble d'énoncés d'une langue donnée (écrits ou oraux enregistrés) qui ont été recueillis pour constituer une base d'observation permettant d'entreprendre la description et l'analyse de la langue en question. » (ARRIVÉ, GADET et GALMICHE 1986.)

> « Si l'on veut étudier tels ou tels phénomènes dans une langue naturelle, il faut recueillir les *corpus* correspondants. Ce sont des ensembles d'énoncés (écrits ou oraux, selon les besoins) que le linguiste pose comme un échantillon représentatif de faits de parole (au sens saussurien) des locuteurs de cette langue. » (CHISS, FILLIOLET et MAINGUENEAU 1993 : 61.)

2 DES LOGICIELS À LA RESCOUSSE : CONCORDANCIERS, EXTRACTEURS ET EXTRACTEURS-ALIGNEURS

2.1 Le concordancier, toujours aussi performant

2.2 L'extracteur de candidats-termes

2.3 L'alignement de corpus

3 INTERNET ET LA DISPONIBILITÉ DES CORPUS

3.1 Recherche de corpus préconstitués

3.2 Recherche de textes spécialisés pour alimenter un corpus

3.3 Recherche de textes spécialisés multilingues à aligner

3.4 Réflexions sur les caractéristiques des textes disponibles

3.5 Internet comme outil de contact avec le spécialiste

4 LE TRAVAIL DE PRÉPARATION DU CORPUS

4.1 Le découpage en unités lexicales

[blanc]*Suivez-moi,*[blanc]*jeune-homme,*[blanc]*s'écria-t-elle.* [blanc] = 3 mots

4.1.1 PONCTUATION

4.1.2 DIACRITIQUES

En français : AaÀàÂâ, Bb, CcÇç, Dd, EeÈèÉéÊêËë, etc.

4.1.3 ALGORITHMES DE DÉCOUPAGE

Dit-il.
Penses-tu ?
Répète-le.
Dis-moi.
Va-t'en.
Laisse-le-moi.
Soyez-en sûr.
Cette personne-là.
Du papier-peint jauni.
Un papier peint en rose.
Suivez-moi, jeune homme !
Son chapeau portait un suivez-moi-jeune-homme.
Le poisson-chat.
Il s'abaisse.
La grand'place.
La grand-place.
Mais qu'en dira-t-on à l'avenir?
Je me moque du qu'en-dira-t-on.

4.2 Propreté de la mise en page

4.2.1 SAUVEGARDER CORRECTEMENT UN TEXTE DEPUIS INTERNET

- Du format HTML au format texte
- Du format PDF *(Acrobat Reader)* au traitement de texte via le format HTML
- Du format *PostScript* au format texte

4.3 La préparation structurelle du corpus

4.3.1 ASSURER LA PÉRENNITÉ DU TEXTE

4.3.2 CONSERVER LA MÉMOIRE DU TEXTE

4.3.2.1 LA NOTION DE DOCUMENT STRUCTURÉ

4.3.2.2 LA TEXT ENCODING *INITIATIVE* (T.E.I.)

- www.tei-c.org
- www.fas.umontreal.ca/EBSI/cursus/vol1no2/beaudry.html

Début typique de document T.E.I. : conserver la mémoire de la constitution du corpus

```
<teiHeader>
       <fileDesc>
         <titleStmt>
             <title>Thomas Paine: Common sense, a machine-readable transcript</title>
             <respStmt><resp>compiled by</resp>
                  <name>Jon K Adams</name>
             </respStmt>
         </titleStmt>
          <publicationStmt>
             <distributor>Oxford Text Archive</distributor>
          </publicationStmt>
          <sourceDesc>
             <bibl>The complete writings of Thomas Paine, collected and edited by Phillip
             S. Foner (New York, Citadel Press, 1945</bibl>
           </sourceDesc>
        </fileDesc>
    <teiHeader>
```

Échantillon de quelques balises T.E.I. :

<bibl>	contient une citation bibliographique structurée de façon informelle, et dont les sous-champs peuvent ou non être balisés explicitement ;
<div1>...<div1>	contient une subdivision de niveau un à sept des parties liminaires du corps ou des annexes d'un texte ;
<h>	marque un mot ou une expression comme étant graphiquement distincte du texte avoisinant, sans aucune interprétation quant à la raison de cette mise en valeur ;
<title>	contient le titre d'une œuvre, que ce soit un article, livre, journal, ou une série, y compris tout sous-titre ou titre alternatif ;
<p>	marque les paragraphes écrits en prose.

4.3.2.3 QUELLES INFORMATIONS PERTINENTES RETENIR POUR LA RECHERCHE TERMINOLOGIQUE ?

5 L'EXTRACTION DE CANDIDATS TERMES

5.1 Méthodes statistiques : Quelques notions de lexicométrie (D'après MULLER 1968)

5.1.1 ÉTENDUE DU CORPUS

Nombre de pages, nombre de lignes, nombres de mots, nombre de caractères ?

5.1.2 FORMES :

- *Token* = le nombre total de formes fléchies comptées dans un texte, y compris celles qui sont répétées maintes fois ;
- *Type =* le nombre total de formes fléchies différentes comptées dans un texte, les répétitions n'étant pas prises en compte.

5.1.3 LEMMES :

Le nombre de formes canoniques différentes (entrées de dictionnaire) observées dans un texte.

Problématique de l'homographie : *affluent, fils, mousse, voile*, etc.

5.1.4 MESURER LA RICHESSE LEXICALE

5.1.4.1 AVEC LEMMATISATION : LA RICHESSE LEXICALE

$$\text{Richesse lexicale} = \frac{\text{nombre de lemmes}}{\text{nombre de formes}}$$

5.1.4.2 SANS LEMMATISATION : LE TYPE-TOKEN RATIO

$$\text{Type-Token Ration} = \frac{}{\text{Nombre total de formes fléchies}} \times 100$$

5.1.5 MESURER LA CONNEXION LEXICALE

$M = 29/53 = 0,55$ $I_{manif\ 1} = 11/40 = 0,28$ $I_{manif\ 2} = 13/42 = 0,31$

	Présent dans Manif 1	**Absent dans Manif 1**	**Total**
Présent dans Manif 2	29	13	42
Absent dans Manif 2	11	∅	11
Total	40	13	53

$$C = \frac{\text{lexique commun}}{\sqrt{(\text{lexique 1} \times \text{lexique 2})}}$$

5.1.6 L'INFORMATION MUTUELLE (CHURCH & HANKS 1990, D'APRÈS KRAIF 1997)

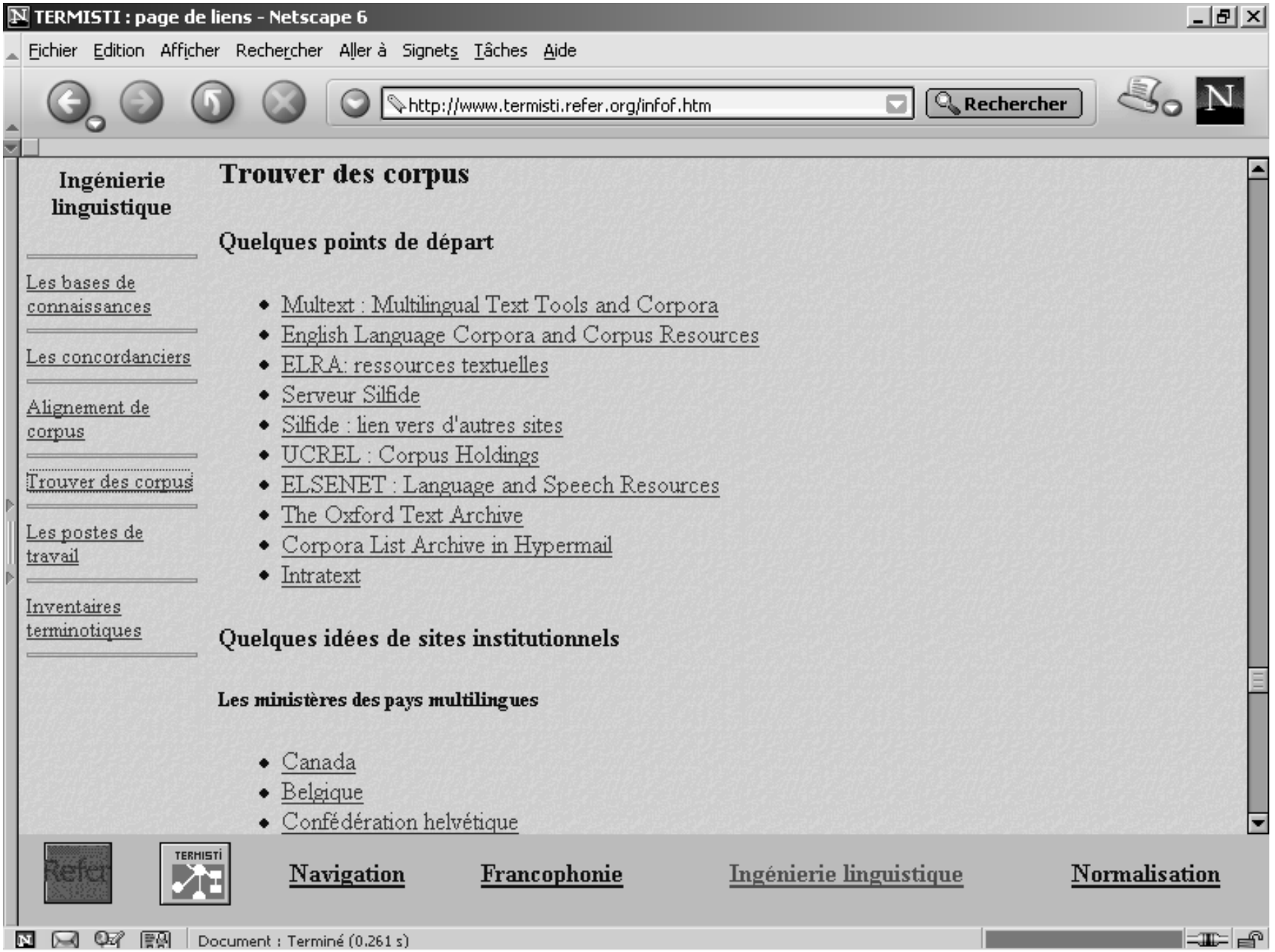
Pour deux formes x et y : Où P(x,y) est la probabilité d'observer x et y ensemble, P(x) et P(y) sont les probabilités respectives d'observer x et y indépendamment.

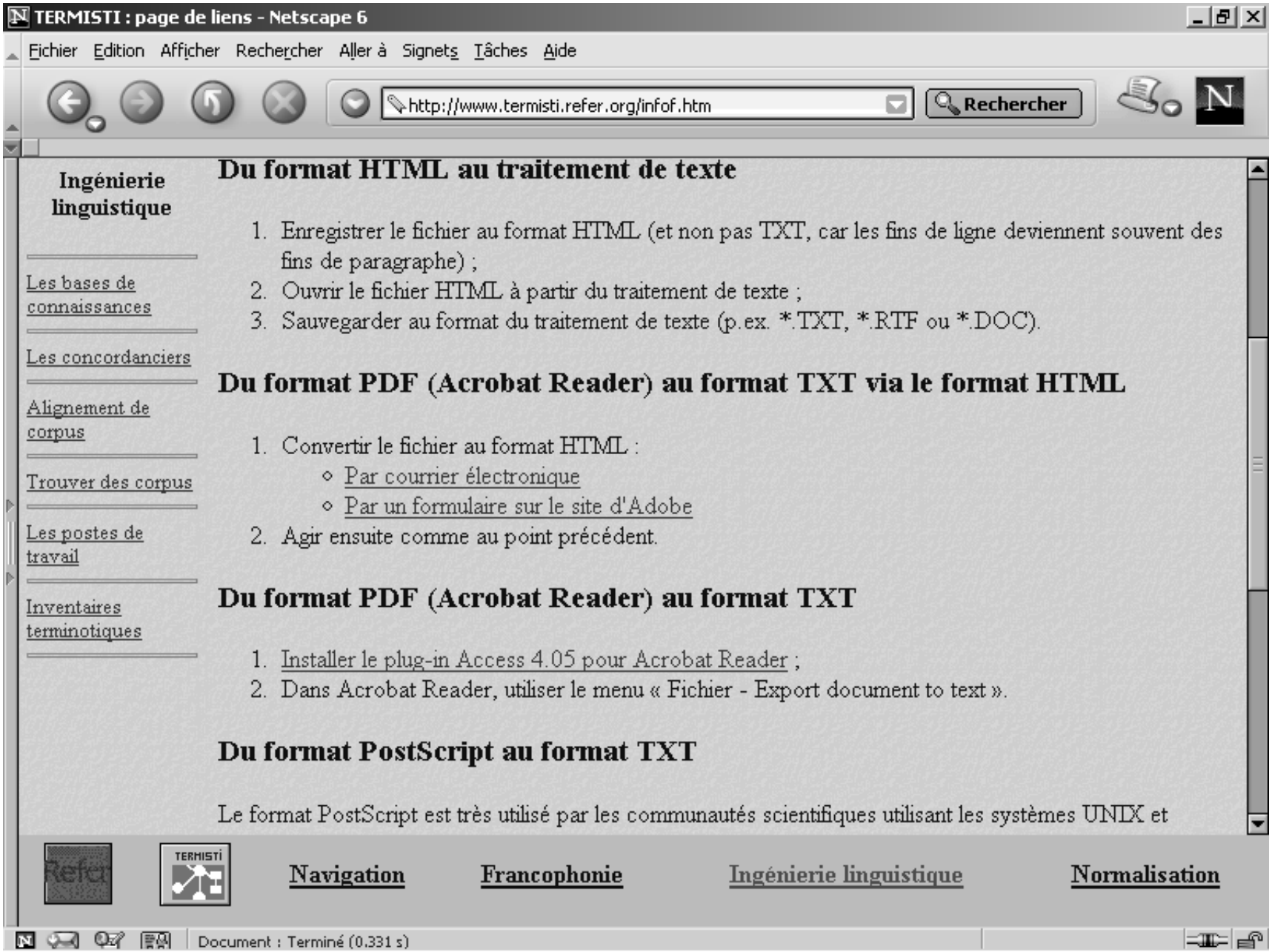
$$I(x,y) = \log 2 \ \frac{P(x,y)}{P(x) \times P(y)}$$

5.2 Approches lexico-syntaxiques : l'exemple d'*Adepte-Nomino*

6 EN GUISE DE RÉFLEXION FINALE


TERMISTI : page de liens - Netscape 6
Fichier Edition Afficher Rechercher Aller à Signets Tâches Aide
http://www.termisti.refer.org/infof.htm
Rechercher
Ingénierie linguistique
Les bases de connaissances
Les concordanciers
Alignement de corpus
Trouver des corpus
Les postes de travail
Inventaires terminotiques
Les concordanciers
S'initier (en anglais) à l'usage des concordances et des corpus, par Catherine N. Ball, Georgetown University
A Summary of Text Analysis Tools
TACT : Text Analysis Computing Tools, Université de Toronto
SATO - Système d'Analyse de Texte par Ordinateur, à l'Université du Québec à Montréal
SATO Internet : version en ligne de SATO
WordSmith Tools : Université de Liverpool, Mike Scott's Web
Multiconcord : Université de Birmingham
Monoconc : Athelstan
Alignement de corpus
Navigation
Francophonie
Ingénierie linguistique
Normalisation
http://www.linq.uqam.ca/sato/outils/sato.htm


TERMISTI : page de liens - Netscape 6
Fichier Edition Afficher Rechercher Aller à Signets Tâches Aide
http://www.termisti.refer.org/infof.htm
Rechercher
Ingénierie linguistique
Les bases de connaissances
Les concordanciers
Alignement de corpus
Trouver des corpus
Les postes de travail
Inventaires terminotiques
Trouver des corpus
Quelques points de départ
Multext : Multilingual Text Tools and Corpora
English Language Corpora and Corpus Resources
ELRA: ressources textuelles
Serveur Silfide
Silfide : lien vers d'autres sites
UCREL : Corpus Holdings
ELSENET : Language and Speech Resources
The Oxford Text Archive
Corpora List Archive in Hypermail
Intratext
Quelques idées de sites institutionnels
Les ministères des pays multilingues
Canada
Belgique
Confédération helvétique
Refer
TERMISTI
Navigation
Francophonie
Ingénierie linguistique
Normalisation
Document : Terminé (0.261 s)


TERMISTI : page de liens - Netscape 6
Fichier Edition Afficher Rechercher Aller à Signets Tâches Aide
http://www.termisti.refer.org/infof.htm
Rechercher
Ingénierie linguistique
Les bases de connaissances
Les concordanciers
Alignement de corpus
Trouver des corpus
Les postes de travail
Inventaires terminotiques
Du format HTML au traitement de texte
1. Enregistrer le fichier au format HTML (et non pas TXT, car les fins de ligne deviennent souvent des fins de paragraphe) ;
2. Ouvrir le fichier HTML à partir du traitement de texte ;
3. Sauvegarder au format du traitement de texte (p.ex. *.TXT, *.RTF ou *.DOC).
Du format PDF (Acrobat Reader) au format TXT via le format HTML
1. Convertir le fichier au format HTML :
Par courrier électronique
Par un formulaire sur le site d'Adobe
2. Agir ensuite comme au point précédent.
Du format PDF (Acrobat Reader) au format TXT
1. Installer le plug-in Access 4.05 pour Acrobat Reader ;
2. Dans Acrobat Reader, utiliser le menu « Fichier - Export document to text ».
Du format PostScript au format TXT
Le format PostScript est très utilisé par les communautés scientifiques utilisant les systèmes UNIX et
Refer
TERMISTI
Navigation
Francophonie
Ingénierie linguistique
Normalisation
Document : Terminé (0,331 s)

MI(x,y)	Freq(x,y)	Freq(x)	Freq(y)	X	y
10	161	1419	4764	requested	Anonymity
8.2	14	1419	1529	requested	permission
7.8	5	1419	698	requested	asylum
7.3	5	1419	968	requested	copies
7.1	4	1419	935	requested	detailed
6.8	4	1419	1090	requested	background
6.2	9	1419	3744	requested	documents
6.0	5	1419	2519	requested	protection
5.7	6	1419	3498	requested	additional
5.4	4	1419	2928	requested	meetings
5.0	199	1419	190545	requested	by
5.0	9	1419	8983	requested	information