IDIAP

Rapport technique

Un interface d'indexation documentaire I d'i, version 2.0

Jean-Luc Cochard

Mai 1993

INSTITUT DALLE MOLLE D'INTELLIGENCE ARTIFICIELLE PERCEPTIVE
CASE POSTALE 609 - 1920 MARTIGNY - VALAIS - SUISSE
TELEPHONE : ++41 26 22.76.64 - FAX : ++41 26 22.78.18
E-MAIL : IDIAP@IDIAP.CH

Numéro : 93-03

Un interface d'indexation documentaire
I d'i, version 2.0

Mode d'emploi

Jean-Luc Cochard

Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP)
Case postale 609, CH-1920 Martigny, Suisse
Adresse électronique : `cochard@idiap.ch`

Résumé

Ce document présente un interface d'utilisation d'un système prototype d'indexation documentaire construit dans le cadre du projet *Specification and Prototyping of a System for the Intelligent Management of Information*[1]. Cet interface permet de commander l'indexation automatique de documents et de contrôler le résultat du traitement effectué. Dans un mode de fonctionnement particulier (mode de démonstration), il est possible d'obtenir une illustration des différentes étapes grâce à une visualisation graphique des résultats intermédiaires.

D'autre part, cet outil illustre les possibilités d'utilisation d'une boîte à outils graphique, `Tk`, et son intégration possible avec Prolog via un utilitaire appelé `expect`.

Cet interface d'indexation fait partie d'une triade d'outils qui comprend un interface d'administration [Coc] et un interface de recherche documentaire [Coc93] qui font l'objet de rapports distincts.

Mots-clé : interface homme-machine, interface graphique, indexation et recherche documentaire, traitement de la langue naturelle.

[1] Projet № 4023-26996, PNR 23, FNSRS

Table des Matières

1 Introduction **3**

2 Le tableau de bord **3**

3 La sélection de documents **4**
3.1 Le choix du répertoire 5
3.2 Le choix du document 6
3.3 La constitution de la file de traitement 6
3.4 L'édition de la file de traitement 7
3.5 La terminaison de la sélection 7

4 L'indexation et la consultation des résultats **8**
4.1 La commande d'indexation et les messages d'exécution 8
4.2 La consultation des résultats 9

5 Le mode de démonstration **11**
5.1 L'illustration de la décomposition en blocs 12
5.2 L'illustration de l'extraction de blocs significatifs 13
5.3 L'illustration de l'analyse syntaxique 13

6 La terminaison du programme **15**

7 L'organisation logicielle **16**
7.1 L'installation de **I d'i 2.0** et la commande Unix 16
7.2 Les opérations internes du lancement de **I d'i 2.0** 17
7.3 Les fichiers de langues 17

8 Commentaires et conclusion **19**
8.1 Problèmes connus de l'interface 19
8.2 Amélioration de l'analyseur linguistique 19

Préface

Ce document technique est un mode d'emploi de la version 2.0 de l'interface I d'i, d'indexation documentaire. Son contenu reprend une grande partie de celui qui décrivait **I d'i 1.4**. Si fonctionnellement, **I d'i 2.0** n'est pas très différent de sa version précédente, il l'est cependant sur deux points essentiels. Premièrement, l'apparence de l'interface a été complètement redéfinie avec comme objectif l'intégration d'effets 3D. Deuxièmement, **I d'i 2.0** a deux modes de fonctionnement : le mode normal qui était celui, unique, de sa version précédente et le mode de démonstration qui illustre graphiquement les étapes qui conduisent à une indexation documentaire.

La version 2.0 corrige quelques-uns des problèmes recensés dans la version 1.4, à savoir :

- Lorsqu'on fait un accès direct sur un répertoire qui n'existe pas, dans la fenêtre de sélection des documents, le système engendre un message d'erreur parasite dans la liste des documents.
- Lors d'une tentative d'indexation d'un document inexistant, le système se bloque.
- Absence de titres aux fenêtres car le bilinguisme devrait s'appliquer ici aussi.
- Concernant le traitement purement linguistique, l'heuristique devrait disparaître au profit d'un analyseur moins sensible à la taille des textes qu'on lui soumet.

1 Introduction

L'objectif de ce document est de fournir un mode d'emploi aussi complet que possible des commandes disponibles dans cet outil d'indexation automatique de documents : **I d'i 2.0**.

Cet outil permet de faire une indexation structurelle et linguistique de lettres administratives conformément à l'un des objectifs du projet de recherche *Specification and Prototyping of a System for the Intelligent Management of Information* [CHK93] qui a précédé la mise en place de cet outil.

L'indexation structurelle consiste à décomposer le texte d'une lettre en identifiant certaines composantes sensibles comme la *date de rédaction*, le *numéro de référence* et l'*adresse du destinataire*. Ces trois données[2] sont normalisées et constituent l'*index structurel* du document.

L'indexation linguistique est réalisée par un traitement linguistique particulier sur la zone du *sujet* de la lettre. Ce fut le point central du projet de recherche mentionné ci-dessus. Nous vous référons donc au rapport final du projet de recherche pour une description de ce traitement linguistique.

La suite de ce document décrit les différentes opérations implantées dans cet interface d'indexation. Dans la section 2, nous présentons la structure générale de l'interface ainsi qu'un scénario simple d'utilisation de ses possibilités. La section 3 présente l'environnement de sélection de documents. La section 4 explique le fonctionnement de la commande d'indexation et le rôle de la commande de consultation des résultats. La section 5 présente les possibilités offertes par le mode de démonstration qui fait apparaître des résultats intermédiaires lors du traitement de la commande d'indexation. La section 6 explique la manière de quitter le programme. La section 7 décrit le format de la commande Unix avec ses paramètres d'exécution, ainsi que les fichiers de description de la langue de l'interface. Et finalement, en conclusion, dans la section 8, nous présentons une liste non exhaustive d'améliorations possibles du produit.

2 Le tableau de bord

L'interface d'indexation documentaire, **I d'i 2.0**, se présente sous la forme d'un "tableau de bord" (cf. Figure 1) constitué de quatre zones avec de haut en bas :

[2] La limitation à ces trois éléments d'une lettre est purement arbitraire. Si l'expérience nous montrait qu'un autre élément du texte permette d'améliorer la qualité de la recherche documentaire, il serait tout à fait possible d'étendre la liste des zones sensibles afin d'y inclure ce nouvel élément d'information.

l'entête – cette zone identifie l'application par son nom et son icone, mais n'offre aucune possibilité d'interaction;

la file de traitement – cette file est symbolisée par une table qui a le titre "**Documents à indexer :**" et qui est assortie d'un *ascenseur* et du bouton de commande "**Éliminer**";

les messages d'exécution – tous les messages à l'intention de l'utilisateur sont centralisés dans cette zone anonyme;

les commandes – elles sont au nombre de cinq et seront décrites en détail dans les sections suivantes.

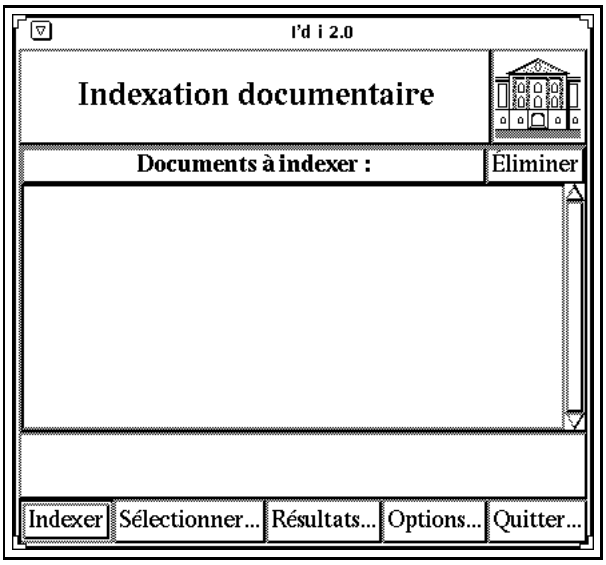


Figure 1 Le tableau de bord de l'interface **I d'i 2.0**.

Un scénario simple d'utilisation de **I d'i 2.0** se déroule de la manière suivante :

1. le lancement de l'interface à l'aide d'une commande Unix (cf. section 7.1);
2. la constitution d'une file de traitement en sélectionnant des documents du système de fichiers (commande "**Sélectionner**..." décrite dans la section 3);
3. le lancement d'une commande d'indexation qui traite séquentiellement tous les documents présents dans la file de traitement (commande "**Indexer**" décrite dans la section 4);
4. la consultation des résultats qui permet de connaître la qualité de traitement atteint par le système en affichant une trace d'exécution (commande "**Résultats**..." décrite dans la section 4);
5. la fin d'une session qui fait suite à une éventuelle itération des étapes précédentes (commande "**Quitter**..." décrite dans la section 6).

Il faut noter que la sélection du mode de démonstration via la commande "**Options**..." ouvre d'autres possibilités d'utilisation qui sont décrites dans la section 5.

3 La sélection de documents

La sélection de documents est une opération lancée par la commande "**Sélectionner**...". Elle fait apparaître une fenêtre temporaire (cf. Figure 2) qui permet de se déplacer dans le système de fichiers et

de choisir un document. Lorsqu'un document est sélectionné, son nom est transféré dans la file de traitement. Plusieurs possibilités d'interaction sont offertes pour la sélection d'un document. Il est premièrement possible de se déplacer dans le système de fichiers en sélectionnant un répertoire. Ensuite le choix d'un document permet de compléter la sélection.

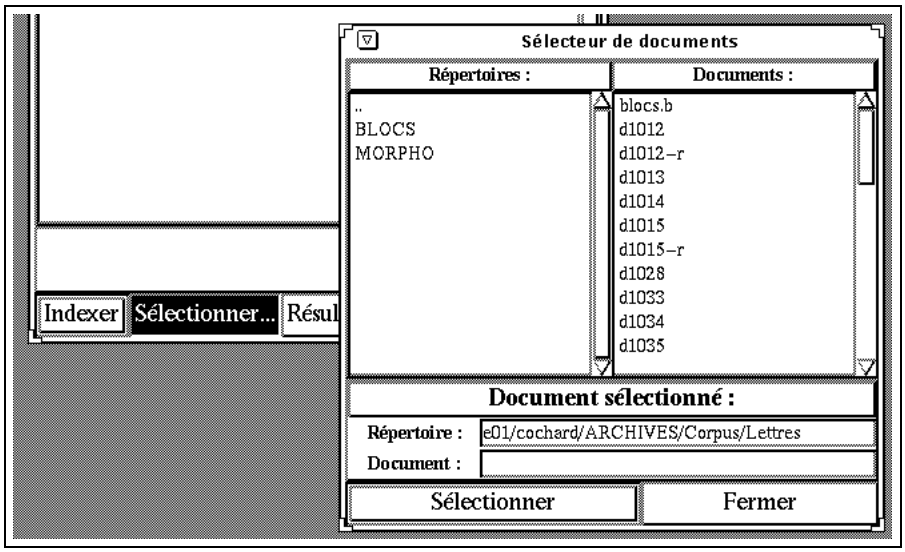


Figure 2 Lors de l'exécution de la commande "**Sélectionner...**", une nouvelle fenêtre temporaire "**Sélecteur de documents**" est affichée à l'écran.

3.1 Le choix du répertoire

Le déplacement dans le système de fichiers permet le positionnement dans la structure arborescente des répertoires. Afin de faciliter la sélection d'un répertoire, deux solutions sont proposées : l'accès contrôlé et l'accès libre.

L'accès contrôlé à un répertoire se fait en utilisant la liste des répertoires sous le titre "**Répertoires :**". Le premier élément de la liste est systématiquement "**..**", qui, selon la convention Unix, dénote le *père* du répertoire courant, tous les autres éléments étant les fils du répertoire courant. Le choix d'un répertoire se fait par sélection de l'élément avec la souris[3]. La sélection d'un nouveau répertoire courant donne automatiquement lieu à une mise à jour de tous les éléments de la fenêtre (cf. Figure 3). L'ascenseur, sur la droite de la liste, permet de se déplacer dans la liste en utilisant indistinctement un des trois boutons de la souris.

L'accès libre à un répertoire se fait en éditant le contenu de la zone appelée "**Répertoire :**". Les commandes suivantes d'édition de ligne sont disponibles :

Positionnement du curseur d'édition – le positionnement se fait par sélection à l'aide de la souris;

Insertion de caractères – l'insertion se fait à la position courante d'édition si le curseur de la souris est positionné dans la zone en question;

Effacement d'un caractère – les touches `Delete` et `Back Space` du clavier effacent le caractère à gauche du curseur d'édition;

Marquage d'une chaîne de caractères – le marquage qui consiste à repérer une chaîne de caractères (affichée en inverse vidéo), est effectué en gardant le bouton de sélection de la souris enfoncée durant un déplacement vers la gauche ou vers la droite;

[3] La "sélection avec la souris" est effectuée en appuyant sur le bouton gauche de la souris lorsque le curseur est sur l'élément en question.

Effacement d'une marque – la combinaison de touches `Control-D` permet d'effacer une marque;

Effacement du texte – la combinaison de touches `Control-N` permet d'effacer tout le texte de la zone d'édition;

Défilement du texte – lorsque le texte dépasse la taille de la zone d'édition, il est possible de déplacer la portion visible en gardant le bouton du milieu de la souris enfoncé durant un déplacement vers la gauche ou vers la droite.

Lorsque le contenu a été édité, la touche `Return` permet de sélectionner ce répertoire, ce qui met à jour l'affichage de la fenêtre.

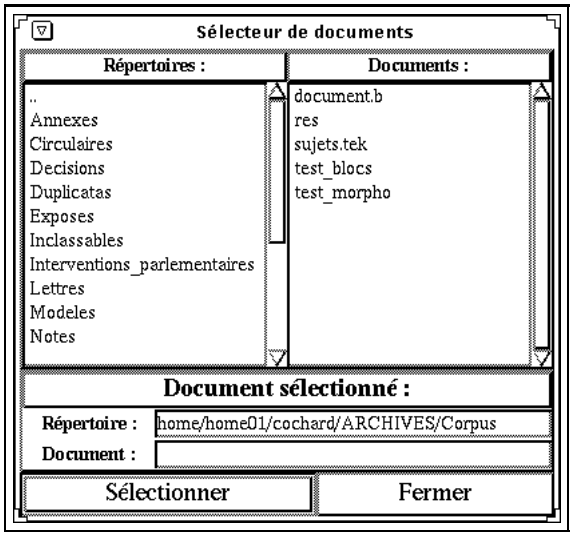


Figure 3 Après avoir sélectionné l'élément "**..**" dans la configuration présentée à la Figure 2, les différentes composantes de la fenêtre sont automatiquement mises à jour.

3.2 Le choix du document

Lorsque le répertoire souhaité est sélectionné, il est possible de faire la sélection d'un document selon le même principe : par un accès contrôlé, via la liste des documents, ou par un accès libre, via la zone appelée "**Document :**". Les commandes d'édition décrites ci-dessus s'appliquent aussi à cette zone.

3.3 La constitution de la file de traitement

L'étape suivante est la transmission des coordonnées du document à la file de traitement. C'est le rôle du bouton de commande "**Sélectionner**" qui ajoute une ligne dans la file de traitement avec le chemin d'accès complet au document (cf. Figures 4.1, 4.2).

Pour des raisons ergonomiques, la commande "**Sélectionner**" est la *commande par défaut* de la fenêtre, c'est-à-dire qu'elle peut être activée depuis n'importe où dans la fenêtre en appuyant la touche `Return`. Pour visualiser cette particularité, le bouton de commande "**Sélectionner**" est placé en retrait par rapport à l'autre bouton "**Fermer**".

Remarque : Dans l'ensemble des interfaces de ce système de gestion des documents informatisés, la mise en évidence des commandes par défaut associées à des boutons est visualisée systématiquement par une mise en retrait du bouton et la touche `Return` est toujours la touche d'activation de cette commande.

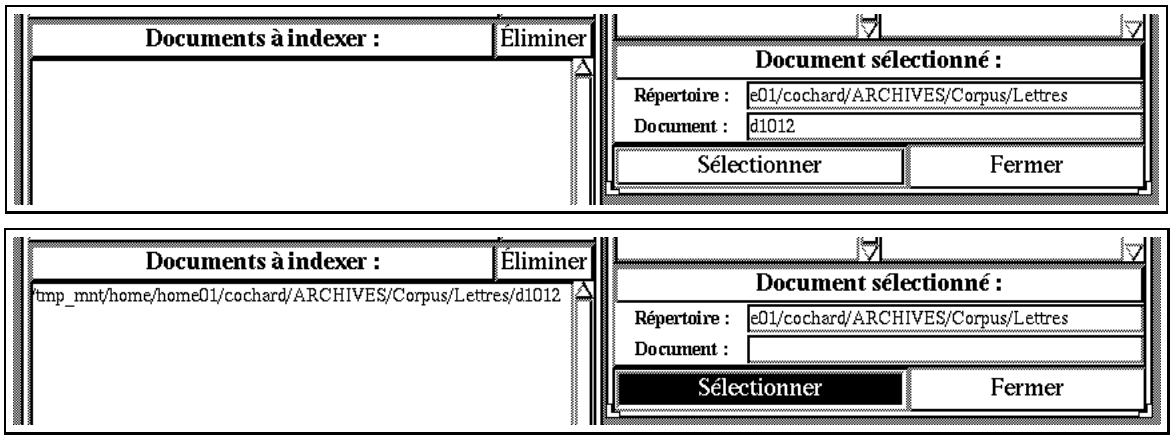


Figure 4 Lorsqu'un document est clairement identifié par son répertoire et son nom, le bouton de commande "**Sélectionner**" effectue le transfert d'information vers la file de traitement. De plus la zone "**Document :**" est effacée afin de permettre la sélection d'un nouveau document.

3.4 L'édition de la file de traitement

Comme il n'est pas exclu qu'un mauvais document soit inséré dans la file de traitement, il est possible de l'éliminer. Cette opération ne recquiert pas l'utilisation de la fenêtre "**Sélecteur de documents**" mais comme elle est hatituellement liée à cette étape, nous la décrivons dans cette section.

L'élimination d'une entrée dans la file se fait en la marquant avec la souris — la sélection la met en inverse-vidéo — et en utilisant le bouton de commande "**Éliminer**" (cf. Figures 5.1, 5.2).

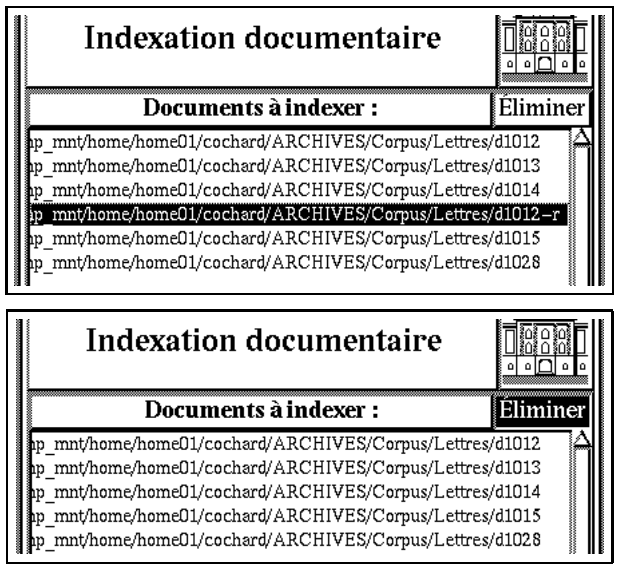


Figure 5 Lorsqu'un document est marqué dans la file de traitement, l'exécution de "**Éliminer**" fait disparaître le nom du document fautif de la file.

3.5 La terminaison de la sélection

Pour terminer une phase de sélection de documents, il est conseillé de faire disparaître la fenêtre de sélection en lançant la commande "**Fermer**". Le positionnement dans le système de fichiers est conservé entre la disparition et l'ouverture ultérieure de la fenêtre.

4 L'indexation et la consultation des résultats

Dans l'introduction nous avons présenté deux formes distinctes d'indexation : une indexation structurelle et une indexation linguistique. La première résulte d'une décomposition du document en blocs de texte et d'une identification de ces blocs. La deuxième utilise le résultat de la première pour isoler le bloc *sujet* et le traiter par des techniques d'analyse linguistique. Ces deux formes d'indexation sont groupées au sein de la commande "**Indexer**" qui fait subir au document une séquence de traitements et qui engendre les deux types d'index (cf. [Coc]).

4.1 La commande d'indexation et les messages d'exécution

Pour lancer le processus d'indexation, il est indispensable d'avoir constitué une file de traitement (cf. section 3). La commande "**Indexer**" (commande par défaut) traite un par un tous les documents associés à une entrée de la file de traitement, en éliminant ces entrées au fur et à mesure. Pour chaque document, une série de messages d'exécution est affichée; ce qui permet de suivre l'évolution du processus d'indexation et de patienter en connaissance de cause !

Il est possible que l'indexation ne se fasse pas complètement, c'est-à-dire jusqu'à la constitution d'un index linguistique, si bien que le traitement d'un document peut se terminer soit par le message "**a réussi !**", soit par le message "**a échoué !**". En cas d'échec, tous les traitements n'auront probablement pas été effectués. C'est le rôle de la commande "**Résultats**..." de consultation des résultats (cf. section 4.2) de présenter le résumé des opérations qui ont effectivement eu lieu sur chaque document.

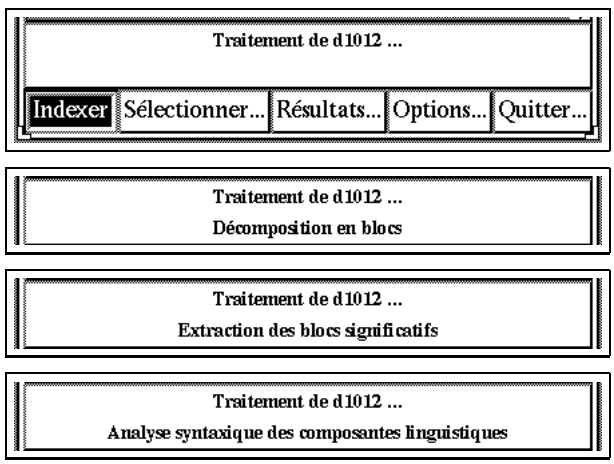
Les différentes copies d'écran de la Figure 6 donnent la liste complète et chronologique des messages d'exécution produits durant le processus d'indexation. Chaque message est associé à un traitement précis dont le rôle est décrit ci-dessous :

Décomposition en blocs – cette étape rendue obligatoire par le format uniquement ASCII des documents à indexer, effectue un découpage géométrique des zones du texte en blocs;

Extraction des blocs significatifs – durant cette étape, l'ensemble des blocs est confronté à une grammaire de mise en page qui décrit différents modèles de rédaction de documents. Si le document candidat satisfait les règles de la grammaire, ses blocs reçoivent une étiquette et seuls certains d'entre eux sont retenus pour les besoins de l'indexation;

Analyse syntaxique des composantes linguistiques – c'est une des deux étapes centrales, avec la suivante, du processus d'indexation. Actuellement, seuls les sujets contenant des mots connus sont analysés sans pour autant que le niveau de détail de l'analyse ne soit connu;

Création d'une clé d'indexation – durant cette étape, quelques résultats d'analyse sont pris en compte — les plus vraisemblables — et les informations capitales pour l'indexation sont retenues et organisées afin de refléter certaines dépendances linguistiques profondes.



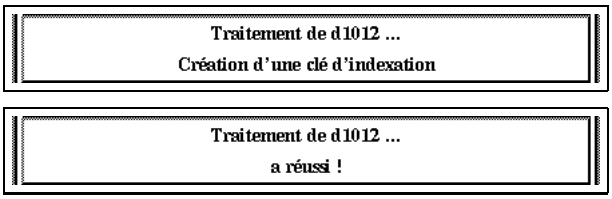


Figure 6 Lors du démarrage de l'indexation, le premier document de la file est pris en compte et un premier message est affiché. Les autres messages viennent compléter le premier en indiquant la phase d'indexation en cours. Le processus s'achève normalement après la création d'une clé d'indexation.

4.2 La consultation des résultats

La consultation des résultats d'indexation peut se faire après, ou durant, le processus d'indexation. La commande "**Résultats**..." ouvre une fenêtre "**Trace d'exécution**" dans laquelle s'inscrit, pour chaque document, une trace de son indexation.

La Figure 7 donne un exemple d'une telle trace avec un échantillon de tout ce qui peut arriver, ou presque ! durant une session de travail. Chaque ligne de la trace est produite par le traitement d'un document et contient les informations suivantes : le nom du document à indexer suivi du nom sous lequel il est enregistré dans la base de données des documents indexés. Suit un résumé des traitements subis. Une marque dans une colonne signifie que le traitement correspondant a réussi. Chaque colonne correspond à une étape précise :

blocs – décomposition en blocs;

docum. – extraction des blocs significatifs (analyse du document);

syntaxe – analyse syntaxique des composantes linguistiques;

index – création d'une clé d'indexation.

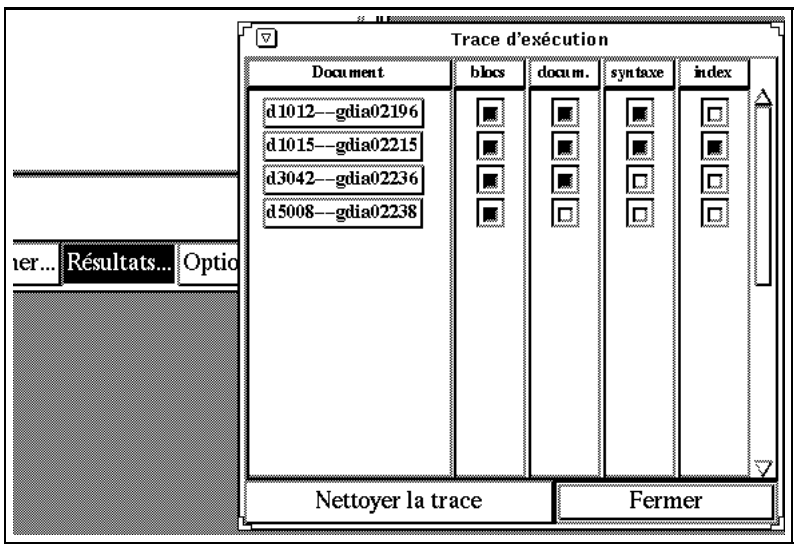


Figure 7 Un exemple d'une trace d'exécution offrant un échantillon des cas de figures possibles.

Afin de compléter l'information de l'utilisateur sur les résultats obtenus ou sur les raisons d'un échec, le nom du document est configuré comme un bouton de commande qui ouvre la fenêtre "**Commentaire d'exécution**". Cette fenêtre contient un complément d'information associé à l'entrée sélectionnée par la souris. Les Figures 8 à 11 présentent les commentaires associés à chacun des résultats présentés dans

l'exemple de la Figure 7. Le bouton de commande "**Fermer**" (commande par défaut) de cette fenêtre de commentaire (cf. Figure 8) permet de la faire disparaître.

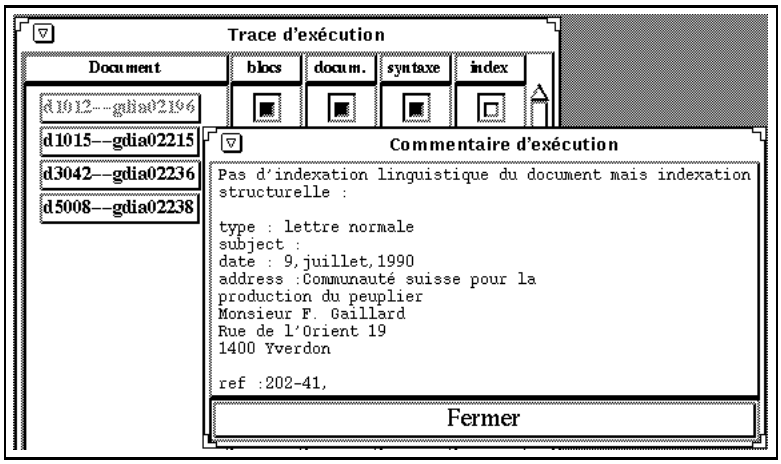


Figure 8 Cet exemple illustre un cas d'échec partiel d'indexation dû à l'absence d'un sujet dans la lettre. La fenêtre présente notamment les blocs qui ont été identifiés avec leur indentification, en particulier la date "**date :**", l'adresse du destinataire "**address :**", et le numéro de référence "**ref :**", le sujet "**subject :**" étant absent. Par ailleurs, chaque document est typé "**type :**". Seuls deux types sont reconnus actuellement : les *lettres normales* avec un destinataire précis et les *lettres circulaires*.

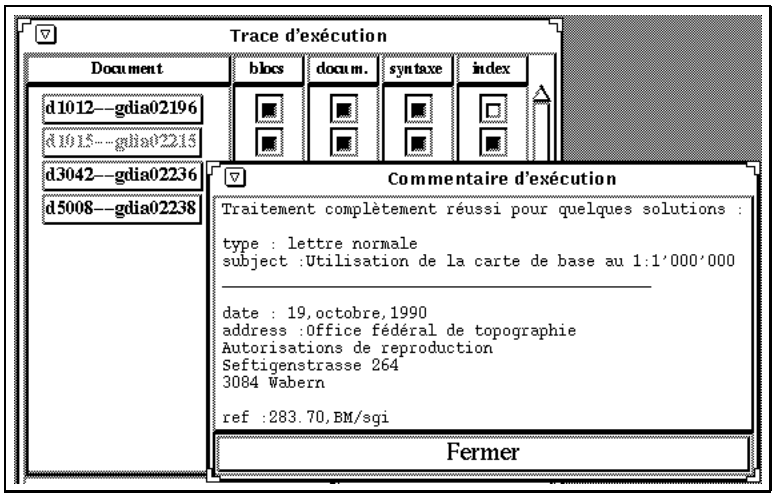


Figure 9 Cet exemple présente un cas d'indexation linguistique et structurelle complet. Les blocs identifiés sont affichés comme dans la Figure 8. Cette fois, un sujet était présent dans la lettre et il a ainsi pu donner lieu à un index linguistique. Cet index a été construit à partir de "**quelques solutions**", à savoir les plus vraisemblables. Deux autres stratégies sont possibles pour la construction d'un index : il peut s'agir d'un index construit à partir de "**la meilleure solution**" ou bien de "**toutes les solutions**". Pour l'instant, le choix d'une stratégie est déterminé uniquement par l'algorithme.

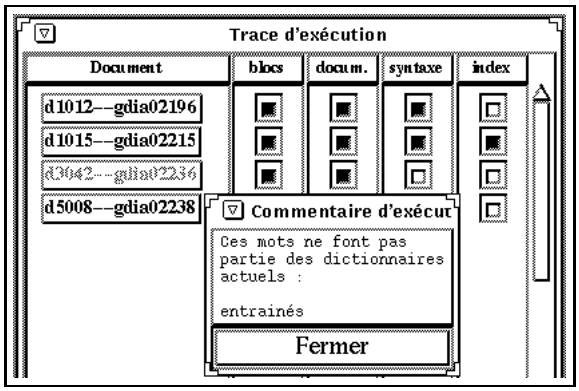


Figure 10 Dans cet exemple, le sujet contient le mot "**entrainés**" qui est mal orthographié et qui, par conséquent, ne fait pas partie de notre dictionnaire linguistique.

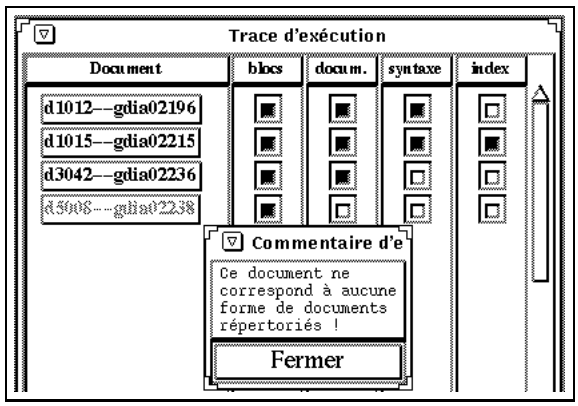


Figure 11 Ce dernier exemple illustre un cas d'analyse structurelle qui échoue. Ce document ne respecte pas les règles de mise en page qui ont été établies sur la base du corpus à disposition. Il s'agit soit d'une forme rare de mise en page, soit d'un bug possible bien que très improbable (!) de l'analyseur de structure.

Le bouton de commande "**Fermer**" (cf. Figure 7) permet de faire disparaître la fenêtre de trace d'exécution et le bouton de commande "**Nettoyer la trace**" (cf. Figure 7) permet, comme son nom l'indique, d'effacer le contenu de la trace d'exécution.

5 Le mode de démonstration

Le mode de fonctionnement de démonstration est une alternative au mode de fonctionnement de la commande "**Indexer**" décrit dans la section 4, qui est le mode par défaut. Pour activer le mode de démonstration, il faut lancer la commande "**Options**..." et sélectionner le mode "**démo**" (cf. Figure 12).

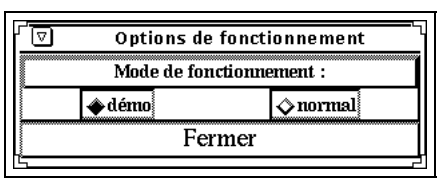


Figure 12 Présentation de la fenêtre d'"**Options de fonctionnement**" avec le mode "**démo**" enclenché.

Dans l'état actuel de cet environnement de démonstration, trois étapes de l'indexation sont illustrées

graphiquement : la décomposition en blocs du document brut, l'extraction des blocs significatifs et l'analyse syntaxique du sujet du document. Seule la création de l'index linguistique n'est pas illustrée.

5.1 L'illustration de la décomposition en blocs

Dans son mode de démonstration, lorsque le processus d'indexation est lancé via la commande "**Indexer**", le message suivant apparaît dans la zone des messages d'exécution : "**Traitement illustré de** ...". La première étape qui est illustrée graphiquement est celle de la décomposition en blocs (cf. Figure 13).

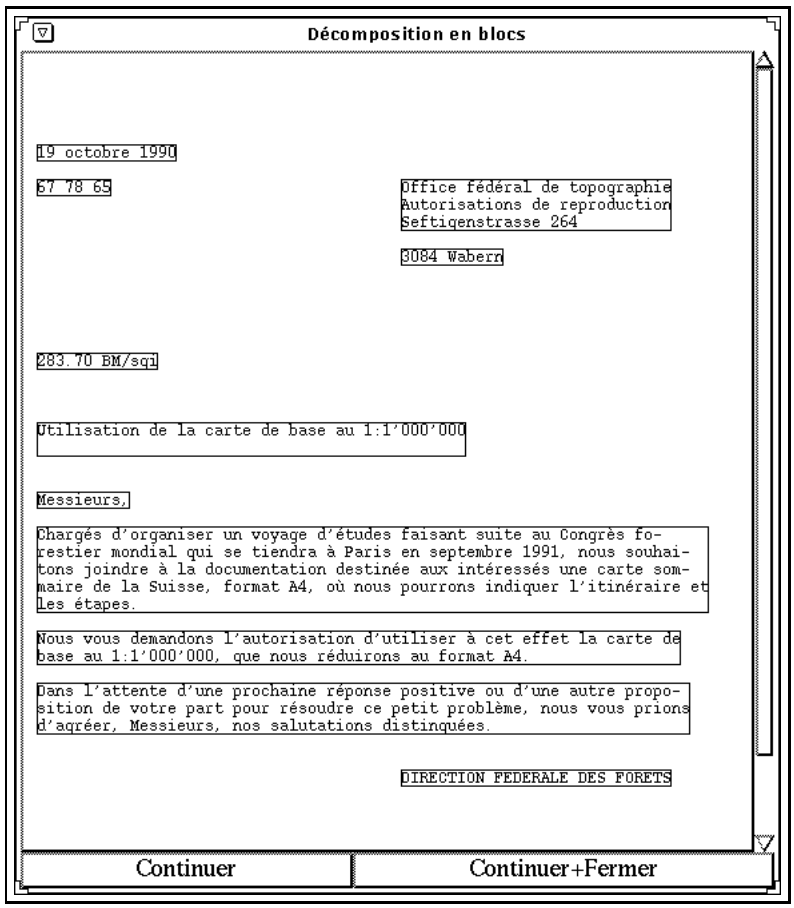


Figure 13 Le texte du document apparaît dans la fenêtre de "**Décomposition en blocs**" et le résultat de la décomposition est synthétisé sous forme de cadres entourant des portions de texte.

La seule interaction possible dans cette fenêtre est le déplacement dans le texte à l'aide de l'ascenseur. Pour passer à l'étape suivante ici, comme pour les autres fenêtres à suivre, deux solutions sont offertes. La commande "**Continuer**" lance de traitement de l'étape suivante en conservant à l'écran la fenêtre de décomposition en blocs. La commande "**Continuer+Fermer**" lance aussi le traitement de l'étape suivante en faisant disparaître la fenêtre.

Remarque : Il faut noter qu'une limitation actuelle de la taille du texte affichable dans ce genre de fenêtre graphique ne permet pas d'afficher l'entier de certains documents relativement longs. Ce qui fait que les cadres des blocs calculés apparaissent où il n'y a pas de texte ! Cette limitation devrait disparaître dans une version ultérieure de l'interface.

5.2 L'illustration de l'extraction de blocs significatifs

L'étape suivante est celle d'extraction des blocs significatifs dont certains serviront d'index structurel du document. Le sujet du document sera la donnée du processus d'analyse linguistique. L'illustration de ce traitement structurel présente les différents blocs de texte en regard de l'étiquette qui les qualifie (cf. Figure 14).

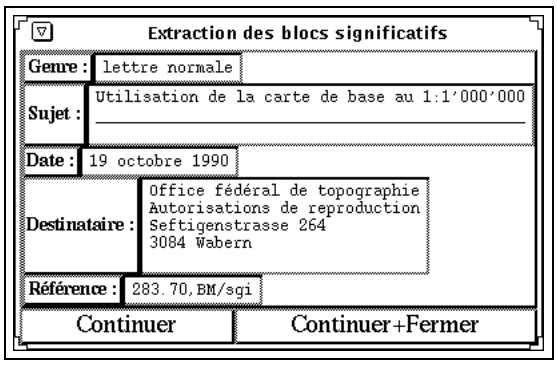


Figure 14 Les composantes significatives du document présenté à la Figure 13 sont affichées, assorties de l'étiquette qui catégorise leur rôle dans le document.

5.3 L'illustration de l'analyse syntaxique

Cette troisième étape dont les résultats sont présentés graphiquement est la plus complexe au niveau de l'interaction. C'est une conséquence de l'éventail des résultats syntaxiques obtenus. Idéalement, un analyseur syntaxique ne devrait fournir qu'un résultat, celui qui décrit la structure syntaxique sous-jacente de la phrase qui est analysée. Pratiquement, et particulièrement dans notre approche de l'analyse syntaxique, nous nous éloignons de cet idéal selon deux axes distincts : premièrement, la stratégie d'analyse indéterministe fournit plusieurs résultats distincts acceptables compte tenu des connaissances internes du système et, deuxièmement, le processus aboutit souvent à des analyses partielles. Dans ce deuxième cas, le résultat final de l'analyse est une séquence d'arbres syntaxiques.

La fenêtre d'"**Analyse syntaxique des composantes linguistiques**" prend en considération tous ces paramètres. La structure générale de son contenu est la suivante (cf. Figure 15) avec de haut en bas :

le texte à analyser – cette zone affiche pour mémoire le sujet du document qui a été analysé;

les arbres syntaxiques – cette zone est découpée verticalement en autant de fenêtres de résultats qu'il y a de constituants partiels dans le résultat de l'analyse;

les ascenseurs – ils permettent de commander un déplacement horizontal et vertical dans la fenêtre de résultats *courante*;

les commandes d'agrandissement – le bouton de commande "**Taille d'origine**" et le potentiomètre à sa droite, permettent de faire varier la largeur de la fenêtre de résultats *courante*;

les commandes principales – ces commandes permettent d'afficher des résultats d'alternatives d'analyse syntaxique — commandes "**Précédent**" et "**Suivant**" — et de poursuivre le traitement à l'aide des commandes usuelles — "**Continuer**" et "**Continuer+Fermer**".

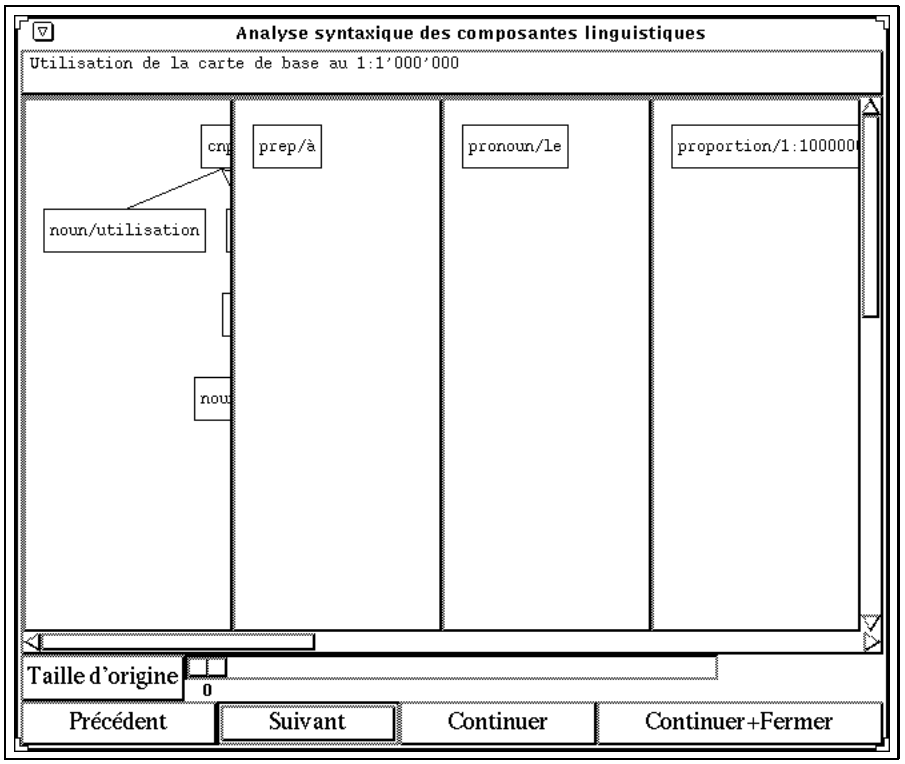


Figure 15 Présentation du résultat de l'analyse de "`Utilisation de la carte de base au 1:1'000'000`". Dans ce cas, l'analyseur n'a pas réussi à produire une analyse syntaxique complète de la phrase. Le résultat est composé de quatre analyses partielles, la première recouvre le texte "`Utilisation de la carte de base`", la deuxième, le texte "`à`", résultat de la décomposition de "`au`" en "`à le`", la troisième, le texte "`le`" et la quatrième, le texte "`1:1'000'000`".

Le contenu des nœuds de l'arbre syntaxique est extrêmement réduit par rapport aux informations utiles à leur constitution. Mais ce choix délibéré reflète très précisément les données que l'analyseur syntaxique transmet à l'analyseur sémantique responsable de la création de l'index linguistique. Il s'agit donc essentiellement de la catégorie morpho-syntaxique et pour les feuilles de l'arbre, cette information est complétée par la forme de citation du mot.

Les opérations élémentaires qui peuvent être effectuées dans cette fenêtre sont les suivantes : détermination de la fenêtre de résultats courante, déplacement dans cette fenêtre, modification de la taille de cette fenêtre et affichage de solutions alternatives dans cette fenêtre.

La détermination de la fenêtre courante

La détermination se fait simplement en sélectionnant une des fenêtres avec la souris. Une conséquence visible de cette opération est de placer la fenêtre sélectionnée en retrait par rapport aux autres. Une autre conséquence est de faire apparaître, sous le curseur du potentiomètre, la largeur de cette fenêtre.

Le déplacement dans la fenêtre courante

Le déplacement dans cette fenêtre est possible à l'aide des ascenseurs, lorsque la détermination de la fenêtre courante a eu lieu.

La modification de la taille de la fenêtre courante

Les deux éléments d'interaction qui permettent de modifier la taille de la fenêtre courante sont le bouton de commande "**Taille d'origine**" et le potentiomètre sur sa droite. En fait seule la largeur de la fenêtre

peut être modifiée (cf Figure 16). Cette largeur peut varier entre 0 et 600 points. La sélection d'une largeur à l'aide du potentiomètre peut se faire, soit en faisant glisser le curseur — le résultat en est désagréable car le contenu en est réaffiché par à-coups — soit en sélectionnant un point de la zone de déplacement de ce curseur — ce qui ne provoque qu'un seul réaffichage de la fenêtre.

Le bouton de commande "**Taille d'origine**" ne sert qu'à redonner la taille originale à la fenêtre courante qui a été calculée par le programme.

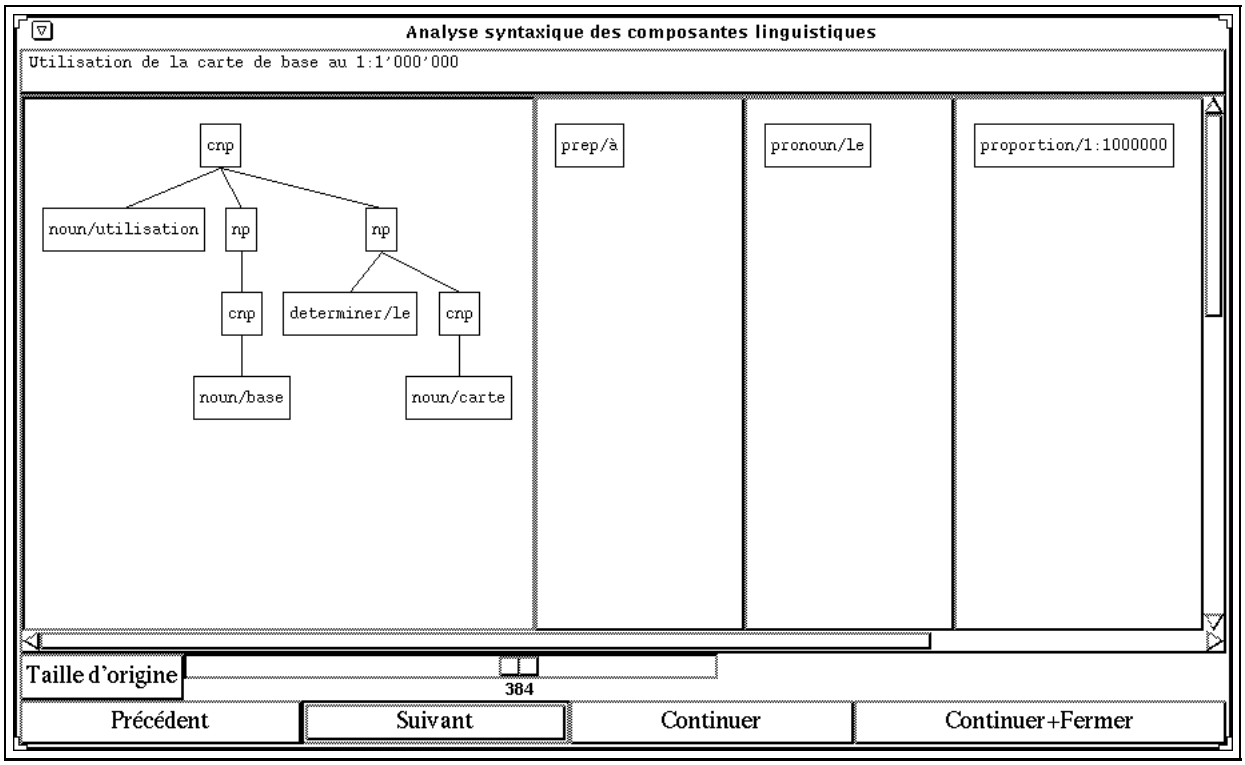


Figure 16 Présentation des résultats partiels d'analyse après édition de la largeur ce certaines fenêtres de résultats.

L'affichage des solutions alternatives

Plusieurs résultats d'analyse peuvent être associés à une fenêtre de résultats donnée. Les boutons de commandes "**Précédent**" et "**Suivant**" (commande par défaut) servent précisément à parcourir cette liste dans une direction et dans l'autre.

6 La terminaison du programme

La commande "**Quitter**..." du tableau de bord permet de mettre fin à une session de travail avec **I d'i 2.0**. Une confirmation est demandée avant de tuer tous les processus impliqués dans ce traitement (cf. Figure 17). La combinaison de touches `Control-C` est une autre solution pour mettre fin à une session. Elle donne lieu à la même demande de confirmation.

Figure 17 Fenêtre de confirmation de la commande de fin de session. Le dialogue est modal, ce qui oblige l'utilisateur à répondre à la question posée avant de faire autre chose, et la réponse par défaut est "**Non**".

7 L'organisation logicielle

Pour pouvoir utiliser **I d'i 2.0**, il y a un certain nombre de prérequis plus ou moins simples à satisfaire :

- disposer d'une SPARCstation de SUN Microsystems;
- disposer du logiciel SICStus-Prolog 0.7 #9[4];
- disposer de l'interprète `wish` de la boîte à outils `Tk3.2`[5] [Ous91];
- disposer de l'interprète `expect` (extension du language `Tcl`)[6] [Lib90];
- disposer des programmes `indexation.tcl`, `in_pro` et `ir`[7].

7.1 L'installation de I d'i 2.0 et la commande Unix

Pour installer l'interface d'indexation documentaire, il faut que les logiciels `prolog, expect et wish` soient accessibles directement, en complétant éventuellement le contenu de la variable d'environnement `PATH`.

Il faut aussi créer un répertoire dans lequel seront centralisés les index et les documents archivés. Ce répertoire doit être repéré par la variable d'environnement `ARCHIVEHOME` et doit avoir un sous-répertoire `Docs`. La série de commandes Unix suivante satisfait ces demandes :

```
% cd ~
% mkdir Archives
% mkdir Archives/Docs
% cat <<END >>.cshrc
setenv ARCHIVEHOME ${HOME}/Archives
END
```

La commande Unix qui démarre l'environnement démontré dans ce rapport est la suivante :

```
% in_pro indexation.tcl french ir >& /dev/null
```

Comme cet interface est bilingue, français-allemand, il est possible de démarrer l'environnement en précisant que les titres et messages doivent être fournis en allemand :

```
% in_pro indexation.tcl german ir >& /dev/null
```

Pour des raisons élémentaires de confort d'utilisation, il est conseillé de créer un "alias" plus concis.

[4] Adresse électronique à contacter : `sicstus_request@sics.se`.
[5] Site pour faire un "ftp" : `barkley.berbeley.edu`.
[6] Site pour faire un "ftp" : `ftp.cme.nist.gov`.
[7] Adresse électronique à contacter : `cochard@idiap.ch`.

7.2 Les opérations internes du lancement de I d'i 2.0

Le lancement de l'exécution de **I d'i 2.0** est une opération assez longue. Quelques explications sur le chargement et l'interaction des processus de ce système seront utiles pour donner un sens à la commande Unix décrite ci-dessus. En outre, cela permettra à l'utilisateur de patienter en connaissance de cause.

Le programme qui gère le démarrage de tout ce système est `in_pro` qui a deux rôles : le premier est de lancer l'exécution des programmes qui figurent en paramètres, à savoir `indexation.tcl french` et `ir`; le deuxième est d'assurer la communication entre ces deux programmes "fils".

L'interface défini dans `indexation.tcl` est un programme de taille modeste qui est assez rapidement chargé. Par contre `ir` est un très gros programme — sa taille actuelle est supérieure à 10 MB — et son chargement prend du temps.

7.3 Les fichiers de langues

L'interface **I d'i 2.0** est bilingue, français-allemand. Les noms, titres et messages dans chacune des deux langues sont centralisés dans deux fichiers paramètres : `.IndexationFrench` et `.IndexationGerman`. Il s'agit de fichiers texte qui peuvent être édités et modifiés à la convenance de l'utilisateur. Ces fichiers sont des fichiers de ressources dans la terminologie X [QO90] et toute adaptation du contenu doit respecter les règles des fichiers de ressources de X. Pour que **I d'i 2.0** trouve ces fichiers paramètres, il est indispensable que chaque utilisateur en ait une copie dans son répertoire `$HOME`.

Le fichier paramètre pour le français a le contenu suivant :

```
*version.text: I'd i 2.0
*Button.font: *Times-medium-r-*-18-*
*RadioButton.font: *Times-medium-r-*-14-*

*CheckButton.font: *Times-medium-r-*-14-*

*select_b.text: Sélectionner...
*options_b.text: Options...
*index_b.text: Indexer
*quit_b.text: Quitter...
*show_log_b.text: Résultats...
*select_action_b.text: Sélectionner
*cut_action_b.text: Éliminer
*yes_b.text: Oui
*no_b.text: Non
*close_b.text: Fermer
*continue_b.text: Continuer
*close_continue_b.text: Continuer+Fermer
*next_b.text:       Suivant
*previous_b.text: Précédent
*default_size_b.text: Taille d'origine

*main_title.text: Indexation documentaire
*list_title.text: Documents à indexer :

*file_selector.text:Sélecteur de documents
*dirs_title.text: Répertoires :
*files_title.text: Documents :

*selection_title.text:  Document sélectionné :
*dir_label.text: Répertoire :
```

```
*file_label.text: Document :

*quit_question.text:Sortie du programme
*quit_mess.text: Voulez-vous vraiment quitter le programme ?

*mess.treatment.text:  Traitement de
*mess.illustrated_treatment.text:   Traitement illustré de
*mess.blocs.text: Décomposition en blocs
*mess.document.text:Extraction des blocs significatifs
*mess.type.text: Genre :
*mess.date.text: Date :
*mess.sujet.text: Sujet :
*mess.adresse.text: Destinataire :
*mess.ref.text: Référence :
*mess.parsing.text: Analyse syntaxique des composantes linguistiques
*mess.indexing.text:Création d'une clé d'indexation
*mess.best.text: pour la meilleure solution :
*mess.some.text: pour quelques solutions :
*mess.all.text: pour toutes les solutions :
*mess.succeeded.text: a réussi !
*mess.failed.text: a échoué !
*mess.initialisation.text: Chargement du système
*mess.done.text: terminé
*mess.fichier_absent.text: Pas de fichier à sélectionner

*mess.b_err1.text:  Décomposition impossible du document en blocs. Comportement \
très étrange !
*mess.b_err2.text: Document inexistant !
*mess.d_err1.text: Ce document ne correspond à aucune forme de documents répertoriés !
*mess.h_err1.text:  Pour des raisons de taille du sujet, le document n'est \
pas soumis à l'indexation linguistique
*mess.p_err1.text: Ces mots ne font pas partie des dictionnaires actuels :
*mess.i_err1.text: Pas d'indexation linguistique du document mais indexation \
structurelle :
*mess.i_res.text: Traitement complètement réussi

*options.text: Options de fonctionnement
*fonctionnement.text: Mode de fonctionnement :
*demo_flag.text: démo
*normal_flag.text: normal

*resultats.text: Trace d'exécution
*exec_doc_name.text:Document
*exec_blocs.text: blocs
*exec_struct.text: docum.
*exec_parsing.text: syntaxe
*exec_index.text: index
*clear_b.text: Nettoyer la trace

*comment_win.text: Commentaire d'exécution
```

8 Commentaires et conclusion

I d'i 2.0 a été développé sur des SPARCstation SUN et fonctionne dans l'environnement OpenWindows de SUN. L'interface graphique a été réalisé à l'aide de la boîte à outils graphique Tk; l'indexation est prise en charge par un programme écrit en SICStus-Prolog 0.7 #9 et la communication entre ces deux applications est réalisée à l'aide d'une extension de Tcl, appelée Expect.

Cette application en collaboration avec **I d'a 1.0** et **I de r 2.0** préfigure ce que pourrait être un environnement d'indexation automatique et de recherche documentaire dans un cadre de travail comme les Archives fédérales suisses ou tout autre institution qui manipule de grosses bases de données textuelles.

8.1 Problèmes connus de l'interface

Dans sa version actuelle, **I d'i 2.0** souffre d'un certain nombre de problèmes de jeunesse qui seront corrigés prochainement. Voici, pour information, une liste de problèmes déjà recensés :

- Le transfert des résultats syntaxiques entre Prolog et l'interface est très lent.
- Absence d'options de fonctionnement qui permettrait, par exemple, de déterminer le niveau de création des index.
- Création d'une erreur interne non fatale lorsqu'on tente de consulter un commentaire sur une trace d'exécution qui n'est pas complète.
- Un même document peut être indexé plusieurs fois sans que le système ne réagisse.

8.2 Amélioration de l'analyseur linguistique

L'utilisation de l'analyseur linguistique (le programme `ir`) laisse entrevoir quelques faiblesses qu'il est indispensable de corriger avant d'envisager d'étendre ses dictionnaires ou sa couverture grammaticale.

- La stratégie d'analyse doit être modifiée pour permettre à la fois d'effectuer une analyse partielle même en présence de mots inconnus.

Toute suggestion d'amélioration ou compte rendu de problèmes à l'adresse électronique figurant dans l'entête de ce document sont les bienvenus.

Références

[CHK93] Jean-Luc Cochard, Michael Hess, and Andreas Kellerhals. A linguistically based information retrieval system for administrative letters. Technical report, IDIAP, Institut Dalle Molle d'Intelligence Artificielle Perceptive, 1993. To be published.

[Coc] Jean-Luc Cochard. Un interface d'administration de documents indexés, i d'a 1.0. Rapport technique, IDIAP, Institut Dalle Molle d'Intelligence Artificielle Perceptive. À paraître.

[Coc93] Jean-Luc Cochard. Un interface de recherche documentaire, i de r 2.0. Rapport technique RT-93-04, IDIAP, Institut Dalle Molle d'Intelligence Artificielle Perceptive, mai 1993.

[Lib90] Don Libes. The expect user manual – programmatic dialogue with interactive programs. Nist ir, National Institute of Standards and Technology, November 1990.

[Ous91] John K. Ousterhout. An X11 toolkit based on the Tcl language. In *Proceedings of the 1991 Winter USENIX Conference*, 1991.

[QO90] Valerie Quercia and Tim O'Reilly. *X Window System User's Guide for X11 R5.* O'Reilly and Associates, Inc., 1990.