NAWEB : un navigateur et analyseur morphologique des pages web pour l'espagnol

Santana Suárez, Octavio ; Hernández Figueroa, Zenón J. ; Rodríguez Rodríguez, Gustavo

NAWeb is a computer application developed in the frame of a larger project designed to tap the flow of linguistics information of Internet documents. It is a tool which is oriented to morphosyntactic analysis of web pages. Its simple inteface facilitates the acquisition of knowledge about the text analyzed in an interactive way.

1.- Introduction.

Le travail montré dans ce papier est la projection naturelle des efforts réalisés par le Groupe de Structures des Données et Linguistique Computationnelle de l'Université de Las Palmas de Grande Canarie pendant ces dernières années. Ces travaux ont étés centrés dans le cadre de la linguistique computationnelle et ont donné lieu, parmi d'autres résultats, au développement d'outils de reconnaissance et gestion morphologique [SANT93, SANT95, SANT97, SANT98, SANT99a, SANT99b], dont quelques uns se trouvent disponibles pour utilisation en ligne dans la page web du groupe (http://gedlc.ulpgc.es). On propose d'utiliser de tels outils comme partie des nouveaux logiciels dont le but est de profiter de l'abondance d'information linguistique qu'Internet représente.

Naweb est conçu pour l'exploration détaillée des documents individuels sous supervision directe de l'utilisateur, et rassemble les caractéristiques typiques d'un navigateur web tout en incluant des nombreuses options pour l'analyse des pages web récupérées.

Les modalités d'analyse qui peuvent être réalisées comprennent : (1) la détection des néologismes, c'est à dire, en principe, tout mot qui ne soit pas identifié par les outils de reconnaissance morphologique incorporés - plus tard, il faudra filtrer dans le cas des entités telles quelles des noms propres, des séquences spéciales, ou même des simples fautes d'orthographe -, (2) l'étude de l'usage des mots, par l'intermédiaire de plusieurs mesures quantitatives et qualitatives, (3) des aspects proches de la syntaxe tels quels l'étude des placements lexicaux ou des régimes prépositionnels.

2.- Architecture de NAWeb.

NAWeb est composé (figure 1) de 7 modules principaux : 1) celui de *navigation*, il s'agit d'un composant *TWebBrowser* conventionnel, fournissant au logiciel les fonctionnalités basiques du Microsoft Internet Explorer, 2) celui d'*extraction de texte*, 3) celui de *lemmatisation*, 4) celui de *désambiguation*, 5) celui de *classification*, 6) celui de *recherche* et 7) celui d'*exportation des résultats*.

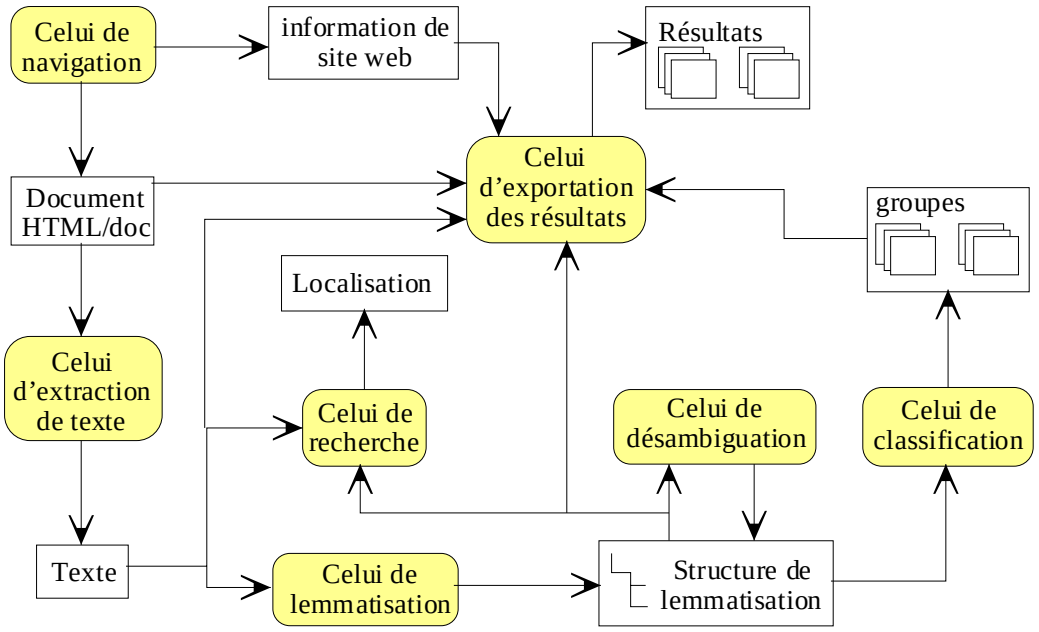

Figure 1 : Architecture de NAWeb

2.1.-Le module d'extraction de texte.

Une page ou document web est basiquement un texte étiqueté en utilisant le langage HTML. Les étiquettes déterminent l'aspect visuel de la page à afficher dans la fenêtre d'un navigateur - on fixe les couleurs, la disposition et structure du texte et beaucoup d'autres détails. Pour le type d'analyse qu'on prétend réaliser - morphologique et morphostatistique -, les étiquettes HTML ne sont pas normalement pertinentes, et, quand elles donnent de l'information utile, le font du point de vue syntaxique ou de l'analyse des parties du texte. Elles ne sont jamais susceptibles d'être analysées, puisqu'elles ne font pas réellement partie du texte. Les types d'étiquettes HTML sont: 1) celles définissant les éléments inclus dans le texte sans interrompre son flux - l'étiquette pour le style de police de caractère gras appartient à cette catégorie - et 2) celles définissant les éléments de coupure du texte - avance de ligne ou changement de paragraphe, parmi beaucoup d'autres. Les premières doivent être éliminées. Celles définissant des éléments de coupure sont remplacées par une marque spéciale que le gestionnaire de l'analyse emploie comme séparateur - elles fournissent des renseignements sur la structure du texte qui peuvent devenir utiles. Il faut en plus tenir compte des marques de caractères spéciaux - permettant d'utiliser des tildes ou des alphabets nationaux - qui, au moment de disparaître, doivent être remplacées par le caractère correspondant.

NAWeb a été doté de la capacité additionnelle de travailler sur des documents dans le format MS-WORD, ce qui ouvre des possibilités énormes en mode local, un résultat très utile - même si l'objectif principal est de travailler sur Internet - ; pour les documents en format MS-WORD, le module d'extraction de texte se comporte d'une façon transparente, puisque le module de navigation a la capacité d'extraire ce type de textes.

2.2.-Le module de lemmatisation.

Le module de lemmatisation, figure 2, travaille sur le texte extrait du document une fois que le *module d'extraction de texte* l'a privé des marques de format. Il produit la *structure de lemmatisation* du document. Ses composants basiques sont : 1) un *Gestionnaire d'analyse*, 2) un *Sélecteur des mots*, 3) un *Reconnaisseur morphologique* et 4) un *Optimisateur des recherches morphologiques*.

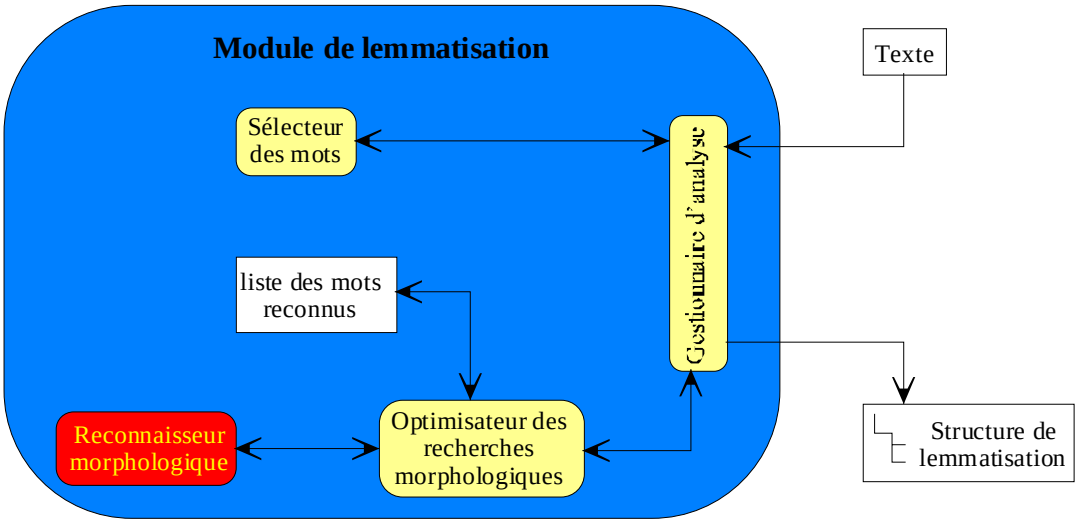

Figure 2 : Module de lemmatisation

Le produit du module de lemmatisation - *structure de lemmatisation* - est une structure hiérarchique dans laquelle on garde : 1) tous les mots retrouvés dans le texte, avec information sur leur localisation - ordre séquentiel dans le texte et position en termes de caractères -, et aussi les signes de ponctuation précédents et suivants - ce qui permet d'avoir des renseignements, même si très peu élaborés, sur la structure du texte - 2) pour chaque mot, les formes canoniques desquelles elle pourrait provenir, avec sa catégorie grammaticale, la flexion et d'autres caractéristiques de la reconnaissance.

2.2.1.- Le Sélecteur des mots.

Le sélecteur des mots réalise un processus d'exploration progressive du texte fournit par le sous-module de nettoyage : dans le premier appel on avance depuis le début en sélectionnant les caractères jusqu'à former le premier mot ; dans celles subséquents, on reprend l'exploration depuis le caractère dans lequel on s'est arrêté la dernière fois et on avance jusqu'à compléter un autre mot. Le processus se répète moyennant des pétitions du gestionnaire d'analyse jusqu'à ce que tout le texte ait été parcouru. Pour l'extraction des mots on distingue parmi cinq classes de caractères: alphabétiques, numériques, signes de ponctuation, terminateurs et d'autres. Quelques symboles peuvent avoir des rôles différents selon le contexte dans lequel ils se trouvent : ainsi, un point peut servir comme signe de ponctuation - il est en même temps un terminateur de mot -, ou bien comme connecteur - classe *d'autres* - dans une adresse URL, par exemple. Ce que le sélecteur des mots extrait appartient à une de trois catégories possibles : séquence alphabétique - les mots eux mêmes -, séquence alphanumérique - formée par lettres et chiffres, comme les identificateurs typiques dans l'informatique - et d'autres séquences - y inclus des caractères spéciaux, comme le point dans les adresses URL. En plus, ces séquences sont accompagnées des renseignements concernant les signes de ponctuation dans son entourage.

2.2.2.- Le reconnaisseur morphologique.

Les séquences des caractères produites par le sélecteur de mots sont étiquetées selon leur catégorie ; parmi elles seulement les séquences alphabétiques sont considérées comme des mots par l'outil de reconnaissance

morphologique et, par conséquent, seulement elles seront soumises à son activité (les autres seront classés comme non reconnues).

L'outil de reconnaissance morphologique est un module externe qui travaille prenant un mot et donnant comme résultat la liste des formes canoniques desquelles il pourrait provenir, et les catégories grammaticales qui seraient applicables. Pour obtenir ce résultat on commence par décomposer le mot dans ses couples possibles racine_terminaison, préfixes, et, dans le cas des verbes, les pronoms enclitiques. La racine passe à un module d'indices donnant sa localisation pour qu'un module d'accès externes vérifie si la racine admet la terminaison, détermine à quelle flexion ou dérivation celle-ci correspond, et en déduise sa forme canonique et fournisse sa catégorie grammaticale.

2.2.3.- L'optimisateur des recherches morphologiques.

L'analyse morphologique des textes obtenus constitue une partie fondamentale des applications développées ; par conséquent, l'utilisation efficace du module produisant cet analyse a une grande influence dans la performance globale. L'analyseur morphologique est en fait composé de deux sous-modules qui réalisent séparément la lemmatisation des formes verbales et des formes non-verbales. Chacun des modules est capable de reconnaître à peu près 450 formes par seconde, d'après des essais réalisés sur un processeur Pentium II à 300MHz avec 128Mo de mémoire volatile. Comme un mot peut appartenir à une des deux catégories, il est toujours nécessaire de mettre en oeuvre les deux processus de reconnaissance, et alors on peut espérer que la vitesse moyenne obtenue soit environ une moitié - entre 220 et 250 mots par seconde.

Dans un texte, les mots ne sont pas distribués de façon uniforme, par contre, typiquement, un nombre très réduit s'y répètent beaucoup et un groupe pas très large n'apparaît qu'une seule fois. Quand un mot apparaît par deuxième fois ou suivantes dans le texte, on retombe dans l'effort d'analyse lors de sa première apparition. Puisqu'il y a des mots qui se répètent beaucoup, il paraît une alternative intéressante la possibilité d'éviter des successifs appels au reconnaisseur, dont on obtiendrait à nouveau les mêmes données de la première fois. La solution consiste donc en la mise en oeuvre d'un type de structure à accès rapide dans laquelle on garderait les données résultantes de la reconnaissance de chaque mot rendu par l'analyseur morphologique - une telle structure serait d'autant plus justifiée que grand le niveau de répétition des mots. L'architecture de la reconnaissance serait donc modifiée de façon à ce que lors de l'obtention d'un mot du texte, on n'interrogerait pas directement l'analyseur morphologique, mais on consulterait préalablement la *liste des mots reconnus* pour vérifier s'il a été déjà lemmatisé ; la surcharge représentant la consulte des mots apparaissant par première fois, et qu'en tout cas il faut lemmatiser, sera amplement compensée par la supérieure vitesse d'accès à la structure par rapport au processus de reconnaissance morphologique.

La structure de la *liste des mots reconnus* est celle d'une table à dispersion des clés - les mots - mise en oeuvre dans la mémoire principale.

2.3.-Le module de désambiguation.

La plupart des mots qui sont reconnus peuvent avoir plus d'une fonction grammaticale. Le reconnaisseur indique toutes les possibles catégories grammaticales et flexions avec lesquelles une forme donnée accorde, mais il ne détermine pas - ce n'est pas sa fonction - quel est le rôle joué dans un texte déterminé. Quand on utilise le reconnaisseur morphologique sur un document, ce qu'on obtient est une liste trié des mots, chacun avec une liste des possibles formes canoniques desquelles il peut provenir, et, pour chaque forme canonique, une liste des catégories et flexions avec lesquelles accorde - quelques mots seront marqués comme « non reconnus » car ils ne figurent pas dans la base de données du reconnaisseur. Dans une première approche, une telle reconnaissance suffit pour un grand nombre d'occasions, mais des études plus fines demandent qu'on puisse identifier avec plus de précision la fonction que chaque mot accomplit.

Les deux problèmes qui empêchent de connaître exactement la fonction de chaque mot sont : 1) celui des mots non reconnues - pour ne pas figurer dans la base - n'a pas de grandes possibilités d'automatisation : la solution nécessiterait que l'utilisateur leur affecte à la main une catégorie - il faudrait les tenir en compte dans le futur pour son addition à la base des données dans des ultérieures révisions - ; pour alléger le travail de l'utilisateur, on pourrait éviter l'examen de toute la liste des mots et que ce soit au programme de chercher et montrer - dans son contexte - les mots qu'il a préalablement marqué comme « non reconnues » en donnant l'option d'en choisir une catégorie - en fonction du contexte du mot, ou encore il pourrait en recommander une - et 2) le problème de la reconnaissance multiple peut aussi être attaqué de façon manuelle mais, étant donné son volume plus large - le nombre de mots non reconnus est beaucoup plus petit - et l'existence

d'options pour choisir, il paraît faisable l'utilisation d'un certain mécanisme de désambiguation automatique - au moins partiel.

Le problème de la désambiguation morphologique a été traditionnellement traité moyennant deux techniques différentes : 1) les méthodes probabilistiques basés sur la statistique - prédominantes depuis le début des années 1980 - résolvent presque toutes les ambiguïtés, mais au coût d'un haut taux d'erreur, et 2) les modèles basés sur des règles font peu d'erreurs, mais ils laissent des ambiguïtés sans résoudre. La plupart des systèmes stochastiques obtiennent leurs probabilités à partir des corpus étiquetés à la main ; on utilise aussi des lemmatiseurs basés sur des modèles de Markov et dérivés des corpus non étiquetés, qui obtiennent des hauts taux de succès ; bien que quelques développements basés sur règles ne sont pas en reste.

La désambiguation n'a pas été abordée par le Groupe de Structures des Données et Linguistique Computationnelle de l'Université de Las Palmas de Grande Canarie comme ligne de recherche que dans une période relativement récente et, même si on a déjà obtenu des résultats intéressants dans l'identification et classification des règles de désambiguation [SANT02], il n'y a pas de doute qu'ils sont encore à atteindre des réussites plus grandes.

On a suivi le même critère de modularisation qu'on a appliqué au reste des éléments, de sorte qu'on a inclus un module de désambiguation basé sur l'état actuel des travaux en développement qui peut être remplacé sans problème à mesure qu'on obtient des résultats plus raffinés. Basiquement, ce qu'on a fait est de profiter de l'ensemble des règles existantes pour les appliquer au résultat de la lemmatisation du texte d'un document. En principe, et puisque, même si ce n'est qu'un peu, la désambiguation a besoin d'une consommation extra de ressources et temps, et même si ce n'est pas toujours, on a choisi de ne pas l'appliquer de façon automatique - ceci figure comme une option que l'utilisateur doit activer quand cela lui convient sur un texte préalablement lemmatisé.

Le processus de désambiguation automatique opère en parcourant la structure qui résulte de la lemmatisation d'un document. On trouve les mots qui ont un dégrée de reconnaissance supérieur à un - ceux auxquels on peut affecter plus d'un lemme. Lorsqu'on trouve un mot dans ces conditions, on prend en compte le mot antérieur et suivant, et on essaie quelles combinaisons de catégories en résultent valides ou pas.

Le programme admet aussi la possibilité que l'utilisateur affecte des catégories aux mots, de façon à réaliser une « désambiguation manuelle » qui peut être utile en petites doses, surtout pour fournir des points de repère pour la désambiguation automatique dans des textes particulièrement complexes.

2.4.- Module de Classification.

Le module de classification a comme but d'engendrer des listes avec les mots du texte groupés selon des critères divers ; il est composé de deux parties : 1) le module de *classification métrique* et 2) le module de *classification morphologique*.

Le module de *classification métrique* ne dépend pas réellement du résultat de la lemmatisation, puisque ce qu'il offre est des tries de mots en fonction des critères tels que sa fréquence d'apparition, son rapport alphabétique directe ou inverse et sa longueur - des caractéristiques qui peuvent se calculer directement à partir du texte sans lemmatiser. À l'exception de ceux en relation avec le calcul des distancies entre mots, toutes les classifications métriques se font lors de l'analyse du texte ; la classification par distances - activée séparément - fait le trie par proximité à celle que l'utilisateur sélectionne de l'ensemble des mots du texte - celle-là doit être réalisée chaque fois que l'utilisateur choisit un mot différent - ; deux classifications possibles ont été envisagées - l'utilisateur choisit celle qu'il préfère - en fonction de la distance de Levenshtein [WEB01,WEB02] ou de la sous-séquence commune la plus longue [CORM90, DÍAZ93, GUSF97] - pas nécessairement contiguë - entre le mot choisi et les autres.

Le module de *classification morphologique* distribue les mots du texte en les groupant par ses catégories grammaticales ; il obtient des listes séparées des verbes, substantifs, adjectifs et d'autres formes, ainsi que des mots non reconnues et des séquences alphanumériques non classifiables comme des mots. La classification se réalise initialement avec le résultat de la lemmatisation, et elle est reconsidérée si l'on effectue un processus de désambiguation ou d'affectation des catégories aux mots classifiés comme non reconnus. Les mots avec ambiguïté sont classifiés selon toutes ses possibilités.

3.- Interface de NAWeb.

zone des menus et des barres d'outils

zone des vues et d'édition

zone d'analyse et des données

Figure 4 : Aspect général de NAWeb

NAWeb montre comme caractéristique la plus remarquable la division de la fenêtre de l'application en trois franges horizontales de haut en bas : 1) *zone des menus et des barres d'outils*, 2) *zone des vues et d'édition* et 3) *zone d'analyse et des données*. Les différents éléments de l'interface correspondent presque directement avec l'architecture interne de l'application, tel parallélisme est logique puisque, étant donné son caractère interactif, on demande une intervention attentive par l'utilisateur, à laquelle l'interface doit contribuer de façon prioritaire.

3.1.-Zone des menus et des barres d'outils.

La zone des menus et des barres d'outils montre : 1) le menu principal de l'application et ses sous-menus et 2) deux barres avec des boutons typiques de navigation et une aire pour l'introduction de la direction à laquelle on veut naviguer. L'ensemble formé des menus, barres d'outils et le contenu du premier onglet de la zone des vues et d'édition se conforment à l'aspect typique d'un navigateur Internet. La possibilité d'activer ou désactiver l'analyse - *Analizar* - représente une nouveauté du navigateur NAWeb.

Le menu principal de NAWeb offre cinq options: 1) *Archives*, 2) *Chercher*, 3) *Annotations*, 4) *Options* et 5) *À propos de*. Le sous-menu *Archivos* [*Archives*] offre cinq possibilités: 1) *Abrir* [*Ouvrir*] permet de naviguer dans un fichier local sélectionné grâce à un dialogue dans lequel on a accès a la structure du système d'information de la machine personnelle ou lancer l'exécution d'une autre instance de NAWeb pour réaliser des navigations parallèles sur des différents documents, 2) *Guardar* [*Sauvegarder*] enregistre localement la page affichée - permettant de différer son étude pour une occasion ultérieure sans nécessité de re-connexion au réseau -, 3) *Guardar como* [*Sauvegarder comme*] est utilisé pour enregistrer une copie avec un nom différent à celui utilisé préalablement, 4) *Imprimir* [*Imprimer*] fournit une version en papier du document et 5) *Salir* [*Sortir*] termine l'exécution de NAWeb - de l'instance sur laquelle est appliquée, s'il y en avait plusieurs ouvertes.

Le sous-menu *Búsquedas* [*Recherches*] est divisé dans quatre sections. La première offre deux options pour la localisation des occurrences exactes des mots ou des formes canoniques - on cherche les mots correspondants à une forme canonique déterminée. La deuxième présente trois options de recherche d'éléments complexes tels que colocations simples, périphrases ou régimes prépositionnels - dans tous les cas on peut configurer par un dialogue la fréquence minimale avec laquelle doivent apparaître les occurrences qu'on veut considérer. La troisième section offre deux options pour localiser les mots les plus pareils à un donné, selon le type de distance à appliquer - sous-séquence commune la plus longue non contiguë (SCML) ou distance de Levenshtein (DL). La quatrième section comprend une seule option commutable *Marcar/Desmarcar* [*Marquer/Démarquer*] qui active l'effet des antérieures sur la vue du texte : chacune des options de recherche des trois premières sections fonctionne en ressortant avec une couleur les occurrences trouvés, dans un processus accumulatif - ce qui permet à l'utilisateur d'étudier ses distribution et corrélation - ; comme la vue du texte est aussi modifiée par les mécanismes de synchronisation quand l'utilisateur se déplace par les autres zones d'information, l'option *Marquer* bloque la synchronisation de façon à ne pas interférer avec le résultat des recherches - ce qui s'active automatiquement chaque fois qu'une recherche est démarrée et l'utilisateur peut l'activer ou désactiver dans n'importe quel moment.

Le sous-menu *Anotaciones* [*Annotations*] est conçu pour extraire des renseignements de contexte sur des éléments sélectionnés dans la vue du texte et les déplacer sur la vue des annotations où ils peuvent être modifiés avec des notes de l'utilisateur et enregistrés sur un fichier. Il offre trois options: 1) *Anotar contexto* [*Annoter contexte*], rajoute le contexte - région de 80 caractères autour du mot - dans la vue des annotations, 2) *Limpiar* [*Nettoyer*], vide la vue des annotations et 3) *Guardar* [*Sauvegarder*], offre un cadre de dialogue avec lequel on peut sélectionner un nom de fichier et une position pour enregistrer l'information montrée dans la vue des annotations.

Le sous-menu *Opciones* [*Options*] permet de configurer le transfert des documents, spécifie la charge des images, vidéos, animations et sons ; ce sont des éléments qui ont un coût important de transmission et qui normalement n'ont pas de l'intérêt lorsqu'il s'agit de faire un étude linguistique du texte contenu dans un document - et non son dessin graphique - ; écarter sa décharge accélère le processus de transfert, et donc on gagne un temps précieux qu'on peut utiliser pour l'étude lui même.

La première barre d'outils est composée de quatre boutons qui sont activés et désactivés selon les circonstances de la navigation et qui représentent les cinq mouvements basiques d'une navigation par Internet: 1) *Anterior* [*Antérieure*], fait revenir sur la page antérieure dans l'ordre de décharge, 2) *Siguiente* [*Suivante*], avance sur la page suivante dans l'ordre de décharge, 3) *Parar* [*Arrêter*], arrête la décharge d'une page et 4) *Recargar* [*Récharger*], charge à nouveau une page.

La deuxième barre contient: 1) un champ d'entrée pour taper une URL qui en se déployant permet de renaviguer sur une page accédée préalablement et 2) un cadre de sélection, *Analizar* [*Analyser*], permettant d'activer ou désactiver l'analyse des documents accédés. Le processus déclenché par *Analizar* est contrôlé par un sous-menu d'options commutables permettant de choisir les analyses qui seront réalisées selon la complexité et la dépendance parmi eux : les études métriques sont toujours effectuées, pouvant sélectionner les morphologiques - activées par défaut -, celles de distribution et les tries des mots par ressemblance à un donné.

3.2.-Zone des vues et d'édition.

La zone des vues et d'édition se trouve organisée par onglets, figure 5, chacun présentant une forme différente de voir le document accédé. L'onglet initialement actif montre la forme « web » du document tel qu'il s'affiche dans tout navigateur - avec des couleurs, graphiques, navigation par les hyperliens y contenus, etc.

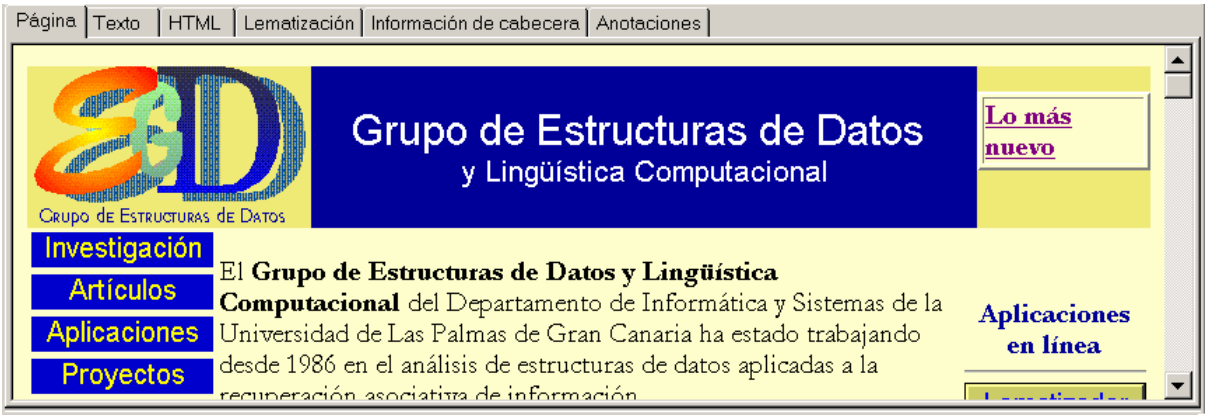

Figure 5 : Zone des vues et d'édition

Le deuxième onglet montre le contenu du texte du document - la vue du texte -, dépourvu de ses aspects graphiques. Le texte est obtenu par extraction du contenu du composant TWebBrowser. Le texte apparaissant dans cet onglet constitue le point de départ pour tous les analyses et recherches : l'information qu'on montre dans les autres parties de l'interface est toujours synchronisée avec ce texte-là.

Dans la partie supérieure on inclut une barre d'outils locale avec cinq boutons ; les quatre premiers servent à se déplacer séquentiellement en avant - suivante et dernière - et en arrière - antérieure et première - par la liste des occurrences écartées dans le texte ; le cinquième bouton a le même effet que l'option *Marcar/Desmarcar* [*Marquer/Démarquer*] du sous-menu *Búsquedas* [*Recherches*].

Un menu flottant permet de copier au « clipboard » la partie sélectionnée dans la vue du texte et sélectionner tout le texte contenu dans cette vue-ci. Quand l'option *Marcar* [*Marquer*] est activée, on peut sélectionner un morceau du texte avec la souris ou avec les touches du pointeur ; quand *Desmarcar* [*Démarquer*] est active, toute action de la souris ou du clavier démarre les mécanismes de synchronisation qu'empêchent de maintenir la sélection.

Le troisième onglet montre le contenu du document dans le langage HTML - vue HTML - ; ce qui réalise la fonction basique d'offrir des renseignements complémentaires par rapport à la structure du document. Ceci répond au même menu flottant que celui de la vue du texte, mais sans restrictions.

Le quatrième onglet montre la vue lemmatisée du texte - vue de la lemmatisation -, figure 6. Comme résultat de l'analyse du contenu de la vue du texte, le premier cadre à gauche présente une liste de tous les mots dans l'ordre d'apparition dans le texte - un code de couleurs identifie le dégrée d'ambiguïté de chaque mot - ; le deuxième cadre montre le résultat de la lemmatisation du mot sélectionné dans le premier : des formes canoniques et des catégories grammaticales. Si l'option d'analyse morphologique est désactivée, le deuxième cadre montrera un message indiquant que les mots n'ont pas été lemmatisés - ceci disparaîtra dès qu'on active la lemmatisation dans le sous-menu d'analyse.

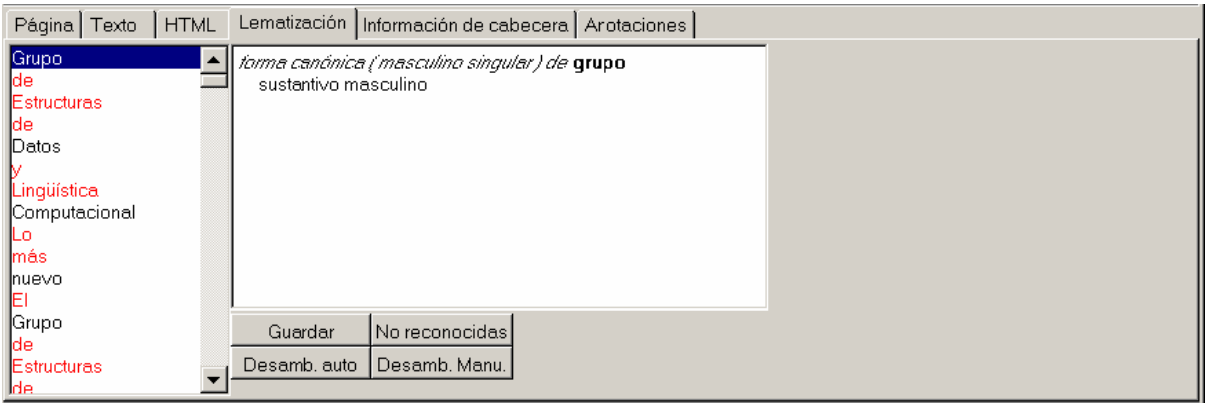

Figure 6 : vue de la lemmatisation

Au-dessous du deuxième cadre s'affiche un groupe de boutons - inactifs si l'analyse morphologique n'a pas été réalisé - servant pour : 1) sauvegarder le résultat de la lemmatisation dans un fichier à format HTML qu'on peut consulter quand on veuille, 2) affecter à la main des catégories aux mots non reconnus, 3) désambiguër automatiquement les mots à reconnaissance multiple, et 4) désambiguër à la main. Les deux options manuelles - 2 et 4 - opèrent par le moyen d'un cadre d'interaction avec l'utilisateur se dépliant dans l'aire vide à droite et offrant une liste des possibles interprétations pour le mot non-reconnu ou ambiguë; l'utilisateur peut choisir une des options pour l'appliquer sur l'apparition actuelle du mot ; il y a aussi la possibilité de laisser l'ambiguïté - et ses occurrences suivantes - sans résoudre pour l'instant.

La section *Information de cabecera* [*Information d'entête*] offre des renseignements techniques sur le serveur auquel on a accédé - sans relevance linguistique. La section *Anotaciones* [*Annotations*] est une aire d'édition dans laquelle l'utilisateur peut travailler librement; aussi les annotations de contexte demandées grâce au sous-menu *Anotaciones* [*Annotations*].

3.3.-Zone d'analyse et données.

La zone d'analyse et données est utilisée pour organiser, selon des différents critères, les résultats obtenus dans l'analyse du texte, figure 7 ; elle est composée de six onglets: 1) *Categorías*, 2) *Canónicas*, 3) *Ordenaciones*, 4) *Resultados*, 5) *Distribución* y 6) *Segmentos*.

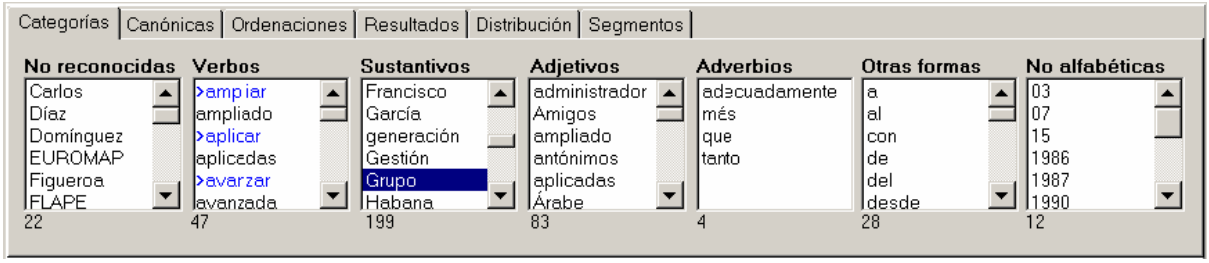

Figure 7 : Zone d'analyse et données. Vue par catégories

L'onglet *Categorías*[*Catégories*] montre les mots groupés selon sa catégorie grammaticale. Six boîtes sont affichées, avec des listes de *verbes*, *substantifs*, *adjectifs*, *adverbes*, *d'autres formes*, *mots non reconnus* et *séquences non alphabétiques*. La liste des séquences non alphabétiques inclut tant des noms et dates que toute autre séquence formée de lettres et caractères « étranges » - dans la web du journal « El País » apparaît « ciberp@is ». Un même mot peut apparaître dans plus d'une boîte - sauf les non reconnus ou séquences non alphabétiques - soit parce qu'il apparaisse dans le texte faisant des fonctions différentes ou parce que ses apparitions n'aient pas été désambiguëées.

L'onglet *Canónicas*[*Canoniques*] montre les formes canoniques correspondantes aux mots du texte dans quatre listes différentes triées selon: 1) relation *alphabétique*, 2) *fréquence*, 3) relation *alphabétique inverse* et 4) *taille*.

L'onglet *Ordenaciones*[*Tries*] montre quatre listes des mots du texte, triées selon sa *fréquence*, relation *alphabétique*, relation *alphabétique inverse* et *taille*; une cinquième liste apparaît lors de l'activation de l'option *Distancias* [*Distances*] du sous-menu reliée à l'option *Analizar* [*Analyser*] de la barre d'outils, ce qui montre les mots triés par distance croissante à celle qu'on ait sélectionnée dans les autres listes - on peut sélectionner le type de distance entre la distance de Levenshtein (*DL*) et celle qu'on obtient en fonction du calcul de la sous-séquence commune la plus longue ou contiguë (*SCML*).

L'onglet *Resultados*[*Résultats*] montre des renseignements de caractère général sur le nombre de mots du document, combien parmi eux sont différents et comment ils se distribuent numériquement par catégories grammaticales - ces chiffres apparaissent aussi sous chacune des listes de l'onglet [*Catégories*]. À droite on montre un profil graphique illustrant le rythme d'apparition des nouveaux mots dans le document.

L'onglet *Distribución* [*Distribution*], figure 8, montre les facteurs de distribution des mots quand on active *Distribución* dans le sous-menu d'analyse; cette information est présentée tant par les mots présents dans le texte comme par ses formes canoniques. Dans les deux cas apparaissent : 1) le mot ou forme canoniques accompagnée de sa fréquence absolue, 2) la fréquence relative d'apparition, 3) l'uniformité de la distribution, 4) le parcours, 5) l'équilibre interne du parcours et 6) la centralité du parcours dans le texte. L'information se montre triée par fréquences croissantes d'apparition des mots à partir d'un seuil - configurable avec les boutons en pointes de flèche qui se trouvent entre les boîtes de distribution des mots et des formes canoniques.

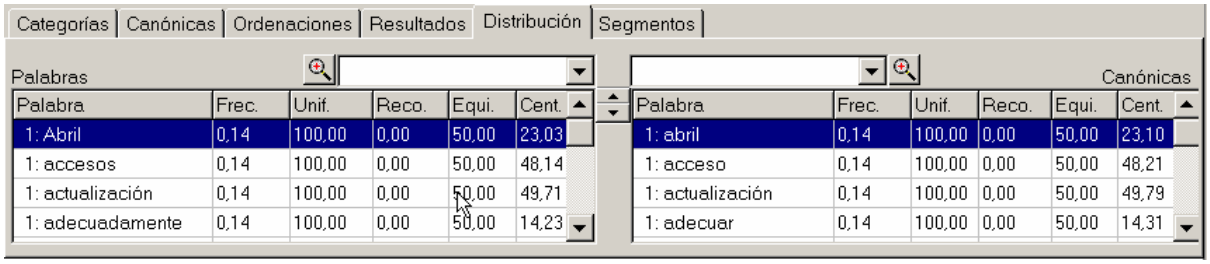

Figure 8 : vue de la distribution

Au-dessus de chacune des boîtes de distribution apparaît une liste combinée. Pousser le bouton *loupe* implique trouver dans le texte une liste de mots ou de formes canoniques avec une distribution similaire à celle qui soit en ce moment-là sélectionnée dans la boîte de distribution correspondante. La similitude de la distribution entre deux mots est calculée en fonction des minimums carrés des différences dans chacun des quatre facteurs de distribution considérées et des fréquences relatives d'apparition de chacune - ce qui permet d'obtenir des pistes intéressantes sur la corrélation et la relevance avec lesquelles apparaissent dans le texte certains mots. On l'interprète comme une approximation « au vol » à l'affinité entre mots, sur laquelle on peut obtenir des méditions quantitatives en fonction de la corrélation des co-occurrences des mots dans le texte [RODR99]; dans ce cas-ci, l'information fournie est plus qualitative que quantitative : ce qui évite le problème de définir qu'est-ce qu'on considère co-occurrence - on ne définit pas des fenêtres ou d'autre classe de portée d'apparition, mais on travaille en combinant les méditions de distribution des mots obtenues de façon indépendante pour chacune - et fait la différence entre mots et classes de mots - qui sont clairement séparés dans l'interface.

L'onglet *Segmentos* [*Segments*] sert à trouver toutes les séquences de mots - segments - respectant certaines restrictions - configurables - par rapport au nombre de mots qui l'intègrent et fréquence minimale d'apparition dans le texte.

3.4.-Synchronisation de l'information affichée.

Autant que possible, l'information des différents zones de l'interface de NAWeb reste synchronisée quand l'utilisateur « se déplace » par elles, de façon à qu'il puisse observer la corrélation depuis des différentes « vues » d'un phénomène linguistique particulier à la place de des données isolées ; par exemple, à mesure

qu'on change la sélection des mots dans la liste alphabétique, on mettra à jour les autres montrant les mots selon des différents ordres ou catégories, on ressortira le mot dans la vue du texte, et on synchronisera la vue de la lemmatisation.

Excepté dans la liste montrant le trie alphabétique, la sélection explicite d'un mot dans toute liste de tries ou de catégories déclenche un événement de sélection dans la liste de trie alphabétique - centre coordinateur de la synchronisation. La sélection d'un mot dans la liste de trie alphabétique provoque : 1) la sélection du même mot dans toutes les autres listes de tries et de catégories, 2) la recherche et la mise en relief de toutes les apparitions du mot dans la vue du texte et que le pointeur soit positionné dans la première apparition, 3) la sélection de la première apparition du mot dans la vue de la lemmatisation et 4) la sélection du mot dans la vue de la distribution.

La sélection explicite d'un mot dans la vue de la distribution provoque un événement de sélection du mot dans la liste du trie alphabétique. La sélection d'un mot dans la vue de la lemmatisation lance aussi un événement de sélection du mot dans la liste de trie alphabétique, mais accompagné d'un autre qui interfère dans la sélection de la première apparition du mot dans la vue de la lemmatisation qui se produirait comme réponse à l'événement de sélection du mot dans la liste de trie alphabétique - si ce n'était par cette interférence, on ne pourrait jamais sélectionner un mot directement dans la vue de la lemmatisation, puisque ceci déclencherait la sélection de la première apparition de celui-ci à la place de celui qu'on veut. La sélection d'un mot dans la vue du texte se transforme en un événement de sélection dans la vue de la lemmatisation et en un événement d'interférence s'appliquant sur la réaction de la liste de trie alphabétique de façon à permettre la mise en relief de toutes les apparitions du mot dans la vue du texte, mais empêchant le repositionnement du pointeur - ce qui détournerait l'utilisateur de la zone dans laquelle il a centré son attention.

On établit un deuxième circuit de coordination en relation avec les formes canoniques: entrent dans le jeu les listes alphabétique et par fréquences des formes canoniques, la vue du texte, la vue de la lemmatisation et la vue de la distribution des formes canoniques.

La différence avec le premier circuit est que dans le cas des formes canoniques on ne peut pas synchroniser depuis la vue du texte ou depuis la vue de la lemmatisation envers les listes de trie, à moins de disposer d'une complète désambiguation du texte - un mot pourrait correspondre avec plus d'une forme canonique.

4.- Conclusions.

Ce travail s'inscrit dans l'intérêt de l'informatique vers tout ce qui est en rapport avec le langage. Une telle ouverture a produit et continue à produire des techniques et des outils d'aide importants dans des différents aspects du travail du linguiste. Aussi, le champ de l'informatique en rapport avec le développement des techniques de traitement du langage naturel s'est bénéficié et continue à le faire de l'attraction que ces outils produisent dans beaucoup de philologues et de la résultante amélioration dans la connaissance des langues que des telles activités impliquent.

La disponibilité des outils adéquats constitue le chemin nécessaire pour obtenir le rendement dû du méta-réseau comme source linguistique à débit jamais imaginé. Il existe le précédent de la propre expansion de l'Internet : même si les éléments physiques étaient disponibles, elle n'a pas commencé à devenir un phénomène de masses que jusqu'à ce qu'on a conçu le World Wide Web et que les navigateurs adéquates sont parus pour que les utilisateurs sans qualification informatique ou technologique spécifique puissent accéder avec facilité à la information. Il ne s'agit plus d'élaborer des outils basiques, mais ceux à un caractère plus complexe et spécialisé permettant de découvrir des espaces d'utilité dans l'exploitation de l'information que les premières approches ont ouverts. Précisément, les outils en rapport avec le langage permettent prévoir la plus grande projection possible, puisqu'ils travaillent dans la base de la communication de tout genre d'information.

NAWeb, en plus, se montre utile sans besoin d'accès externe, comme outil d'auto-apprentissage permettant d'étudier et corriger le style de celui qui l'utilise avec ce but ; la possibilité d'analyser des documents dans le format MS-WORD résulte de l'intérêt. Avec connexion externe, il facilite l'étude du style de ceux auteurs dont les textes soient accessibles. Il approche la connaissance de la structure du lexique : d'un texte, auteur, époque, etc., de la langue, avec ultérieures applications, par exemple, dans l'enseignement de l'espagnol aux étrangers.

5.- Bibliographie.

[CORM90]Tomas H. Cormen ; charles E. Leiserson ; Ronald L. Rivest. *Introduction to Algorithms*. The MIT Press, 1990.

[GUSF97]Dan Gusfield. Algorithms on strings, trees, and sequences. Computing Science and Computational Biology. Cambridge University Press, 1997.

[DÍAZ93]Díaz, M. ; Pérez, J. ; Santana, O. *Distancia Dependiente de la Subsecuencia Común Más Larga entre Cadenas de Caracteres*. Anales de las II Jornadas de Ingeniería de Sistemas Informáticos y de Computación, Quito (Ecuador). Abril, 1993. 117/123.

[RODR99]Horacio Rodríguez Hontoria. *Técnicas estadísticas en el tratamiento del lenguaje natural*. Filología e Informática. Seminario de filología e informática de la Universidad Autónoma de Barcelona. 1999. 111/140.

[SANT93]Santana, O. ; Rodríguez del Pino, J. C. ; González Domínguez, J. D. *Frectext: Una Aplicación de Ayuda a la Elaboración de Documentos*. Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Febrero, 1993. Nº 13, 451/462.

[SANT95]Santana, O. ; Hernández, Z. ; Rodríguez, G. ; Rodríguez, J. C. ; González, J. D. *Proyecto SOTA: Sistema de Organización de Texto Abierto*. Procesamiento de Lenguaje Natural, Revista nº 16. Ed. : SEPLN. Avril, 1995. Nº 16, 92/94.

[SANT97]Santana, O. ; Pérez, J. ; Hernández, Z. ; Carreras, F. ; Rodríguez, G. *FLAVER: Flexionador y lematizador automático de formas verbales*. Lingüística Española Actual XIX, 2, 1 997. Ed. Arco/Libros, S.L. 229/282.

[SANT98]Santana, O. ; Pérez, J. ; Carreras, F. ; Duque, J.D. ; Hernández, Z. ; Rodríguez, G. *Reconocedor y generador automático de formas nominales*. Diccionarios e informática, 1998. Publicaciones de la Universidad de Jaén. 57/74.

[SANT99a]Santana, O. ; Pérez, J. ; Carreras, F. ; Hernández, Z. ; Rodríguez, G. ; Duque, J.D. *De un reconocedor y generador morfológico del español en Internet*. Publicado Mayo, 1999, Lexicon Planet Ltd.

[SANT99b]Santana, O. ; Pérez, J. ; Carreras, F. ; Duque, J. ; Hernández, Z. ; Rodríguez, G. *FLANOM: Flexionador y lematizador automático de formas nominales*. Lingüística Española Actual XXI, 2, 1999. Ed. Arco/Libros, S.L. 253/297.

[SANT02]Octavio Santana Suárez, José Rafael Pérez Aguiar, Luis Javier Losada García, Francisco Javier Carreras Riudavets. *Hacía la desambiguación funcional automática en Español*. Procesamiento de Lenguaje Natural, Revista nº 28. Ed. : SEPLN. Mai, 2002. Nº 28, 1/22.

[WEB01]*Levenshtein distance in Three flavors*. http://www.merriampark.com/ld.htm

[WEB02]*Distance betwen strings*. http://www.cut-the-knot/do_you_Know/string.html