DiET : Diagnostic et Évaluation des Systèmes de Traitement de la Langue Naturelle

Ludovic Tanguy, Susan Armstrong, Pierrette Bouillon, Sabine Lehmann

DiET : Diagnostic et Évaluation des Systèmes de Traitement de la Langue Naturelle

DiET : Diagnostic and Evaluation Tools for Natural Language Processing

L. Tanguy[1] , S. Armstrong, P. Bouillon et S. Lehmann
ISSCO, Université de Genève

Mots-clés : Évaluation, phrases-test, annotations, ressources linguistiques
Keywords : Evaluation, test-items, annotations, linguistic resources

Résumé

Cet article présente en détail le projet DiET[1] (Diagnostic and Evaluation Tools for Natural Language Processing). Ce projet propose une plate-forme complète d'évaluation des systèmes de traitements automatiques de la langue naturelle, sous leurs aspects morphologiques, syntaxiques et sémantiques. En plus d'un ensemble structuré de données de test pour le français, l'anglais et l'allemand, cette plate-forme comprend des outils destinés à configurer et construire des bancs de test spécifiques pour un outil particulier, et pour automatiser ces tâches de configuration.

Abstract

This paper presents the DiET project (Diagnostic and Evaluation tools for Natural Language Processing) [1] . This project proposes a complete workstation for natural language processing software evaluation. It includes a complete database of test-items for morphological, syntactic and semantic phenomenons in three languages (English, French and German). It provides several tools for building test-suites and automatically customizing these suites for a specific software and/or linguistic domain.

[1]ISSCO, 54 Route des Acacias, 1227 Genève Suisse
Tél :+41/22/7057112
Fax : +41/22/300 1086
Ludovic.Tanguy@issco.unige.ch
Susan.Armstrong@issco.unige.ch
Pierrette.Bouillon@issco.unige.ch
Sabine.Lehmann@issco.unige.ch

[1]DiET est un projet financé par la Communauté Européenne (LE 4204), et l'Office Fédéral suisse de l'Education et de la Science. Informations sur le site WWW : http://dylan.ucd.ie/DiET/

1.Introduction[2]

Au vu du nombre toujours croissant de produits commerciaux liés au traitement automatique de la langue naturelle, sous tous les aspects de cette discipline, l'évaluation devient un véritable champ d'activités (King, 1996). Celle-ci concerne tant les développeurs de ce type d'outils que leurs utilisateurs potentiels. Cependant, alors que les procédures d'évaluation générales des produits informatiques doivent prendre en considération la convivialité de l'interface, la documentation et le support technique, un aspect crucial spécifique aux logiciels de traitement de la langue est bien entendu la performance linguistique. Une telle évaluation nécessite des données de référence, mais en l'absence de telles données et d'outils d'évaluation généraux, les utilisateurs et développeurs en sont souvent réduits à construire leur propre base de test.

Les ressources langagières utilisées à des fins d'évaluation consistent généralement en un ensemble de données extraites d'un corpus et censées constituer un échantillon représentatif pour une application particulière. Certains développeurs disposent ainsi de données qui s'accroissent au long des procédures d'évaluation et qui doivent être gérées et adaptées pour tenir compte de l'évolution des systèmes. Quoiqu'il en soit, il n'existe que très peu de données pour des évaluations systématiques à grande échelle.

Les raisons de cette absence de systématicité sont évidentes : les échantillons sont choisis de façon plus ou moins arbitraire[3] , et ne permettent que rarement un contrôle total des performances d'entrée/sortie du système. L'évaluateur est souvent confronté au dilemme de la profondeur et de la couverture des données : un échantillon réduit permet d'envisager une analyse et un diagnostic détaillés. Inversement, choisir des données représentatives conduit souvent à un volume trop important pour permettre un contrôle et une interprétation précis des résultats.

De plus, la mise en place de bancs de test systématiques et fiables est extrêmement coûteuse, comme le montrent les grandes conférences américaines sur l'évaluation (MUC, TREC, etc.) Un des facteurs de ce coût est sans doute la construction et la préparation des données, c'est-à-dire l'obtention d'une quantité suffisante de ressources langagières annotées pour établir une base de test. En conséquence, il est pratiquement impossible pour une entreprise ou une institution de développer ses propres procédures d'évaluation.

Enfin, le partage de données de référence est souvent freiné non seulement par le manque de transparence des formats d'échange, mais aussi par le fait que les données elles-mêmes sont destinées à des domaines et des applications très spécifiques, et ne sont donc pas interchangeables. Ainsi, le coût de développement de ces procédures ne peut se justifier que pour des applications indépendantes du domaine, comme la recherche documentaire en texte intégral.

Bien qu'il ne vise pas à résoudre tous les problèmes des procédures d'évaluation, le projet DiET se veut une contribution à l'avancée vers des réponses à ces questions. Développé à la suite de projets comme TSNLP (Estival & Lehmann, 1997), TEMAA (Maegaard *et al*, 1996) et FraCas (Cooper *et al,* 1997), DiET propose des outils pour construire des données systématiques et contrôlées, ainsi que pour affiner ces données en vue de domaines et d'applications précis. Plus précisément :

- DiET propose une architecture souple intégrant différents outils pour la construction, le stockage et la modification de données de test.

- Une boîte à outils permet à l'utilisateur de développer manuellement des données structurées représentant des phénomènes linguistiques de différents niveaux. Cependant, ces outils ne

[2] Note : cette introduction reprend en partie la description de DiET présentée dans (Netter *et al*, 1998).

[3]Une exception est l'utilisation de corpus représentatifs en traduction automatique, cf (Rayner *et al*, 1995).

sont pas limités à de telles données de référence, mais peuvent servir au traitement et à l'étiquetage de corpus non structurés.

- Le système propose différents types d'annotations, permettant à l'utilisateur de rassembler, construire et sélectionner les types d'annotations les plus pertinents pour une application donnée. Ces annotations peuvent être entrées manuellement ou par le biais d'une application serveur.

- DiET propose des outils et des méthodes pour l'analyse de corpus, qui permettent à l'utilisateur de relier systématiquement les bancs de test aux corpus réels représentatifs du domaine de l'application. Ainsi, les données de test annotées et structurées peuvent être utilisées comme une forme condensée de corpus plus importants, ou comme entrées d'index vers le corpus où des exemples réels peuvent être extraits.

- Le système offre des fonctionnalités permettant de modifier des données de test en fonction des domaines et des applications. L'utilisateur peut ainsi sélectionner un jeu de phrases-test et utiliser l'outil de remplacement lexical pour y insérer un vocabulaire spécifique (*cf.* section. 5.2).

- Enfin, DiET comprend un très grand volume de données. Celles-ci couvrent les phénomènes syntaxiques (elles étendent les données développées dans le projet TSNLP, *cf.* Estival & Lehmann 1997), morphologiques et certains phénomènes sémantiques comme l'ellipse et l'anaphore, le tout dans trois langues différentes (anglais, français et allemand).

2.Architecture du système

Dans un but de généricité, l'architecture de DiET a été définie en évitant de restreindre le type des données et des annotations, ainsi que les modules externes (pour la modification des données). Il s'agit donc d'une architecture client-serveur, avec comme client central une interface graphique pour la construction, l'annotation et la configuration des données, et des modules serveurs, comme la gestion de la base de données et des outils d'annotation automatique.

L'outil central de construction et d'annotation permet d'entrer de nouvelles données et de les décrire à l'aide d'attributs librement définis et configurés. Les données de test elles-mêmes sont soit des phrases-test, des suites ordonnées de phrases-test, ou des segments de ces phrases (mots, syntagmes, morphèmes)[4]. Ces données peuvent être entrées manuellement ou extraites d'une source externe, de façon à ce que le système puisse, en principe, annoter des phrases isolées construites manuellement ainsi que des corpus plus volumineux choisis par l'utilisateur.

L'utilisateur est libre de choisir le type de données à développer, ainsi que le type d'annotations associées à ces données. De façon très modulaire, il se voit proposer un grand choix d'annotations de base, parmi un nombre fixé de types de données comme des arbres, des chaînes de caractères, des valeurs numériques, des booléens, etc. Ces différents modes d'annotation sont directement associés à des fonctions d'affichage, d'édition et de stockage. Ceci implique par exemple qu'une fois une annotation de type arbre déclarée, toutes les fonctions permettant de construire, d'éditer, de stocker et d'effectuer des recherches dans des arbres sont accessibles à l'utilisateur.

L'utilisateur peut définir de nouveaux types d'annotation, mettant en relation des types de données et des modes d'annotation, qui sont alors insérés dans une hiérarchie. Ceux-ci peuvent s'appliquer à

[4]On continuera cependant dans la suite de l'article à employer le terme générique de phrase-test.

différents types d'objets linguistiques (phrases-test, groupes de phrases, etc.) et leurs valeurs associées (si elles existent) peuvent être entrées manuellement ou bien fournies par un service externe (serveur). Parmi les instances de ces types d'annotations, on trouve par exemple des structures syntagmatiques pour le type arbre, des relations anaphoriques pour le type arc, des jugements de grammaticalité pour le type booléen, le nombre d'éléments identifiés par un programme de segmentation pour le type numérique, etc.
Une fois les types d'annotation définis, l'utilisateur peut en sélectionner un sous-ensemble, les attribuer à un élément de test, et affecter des valeurs aux différents attributs. Bien qu'en principe chaque phrase-test puisse se voir attribuer un type d'annotation différent, on définira bien entendu des groupes d'annotations pour des bancs de test destinés à un certain type d'application. Ceci est supporté par un ensemble de fonctionnalités pour mémoriser, copier et éditer des configurations et des annotations spécifiques.
La hiérarchie ainsi formée correspond à celle décrite dans la figure 1 :

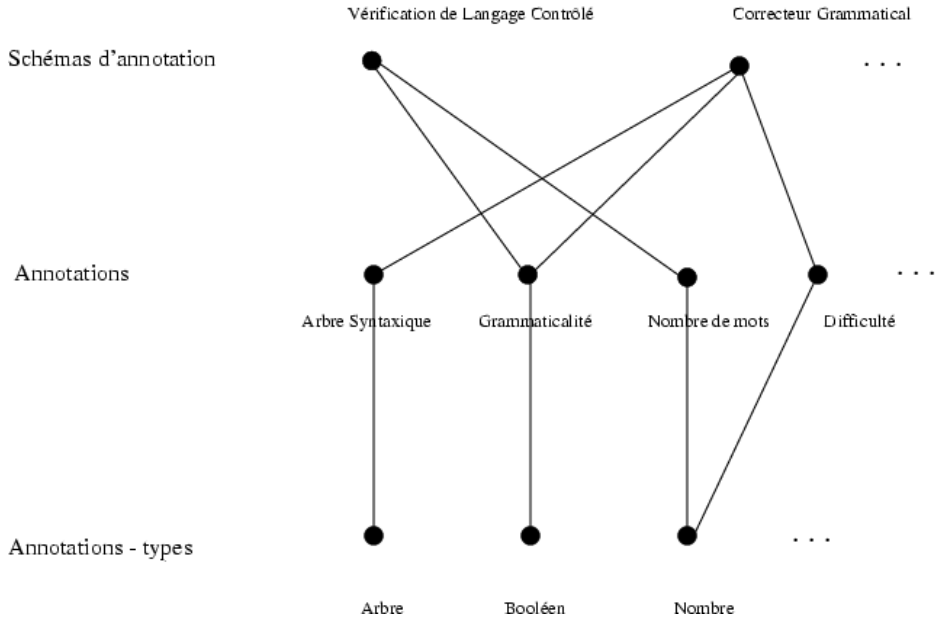

Figure 1: Hiérarchie d'annotations

Alors que la plupart des fonctions de déclaration, de sélection et d'entrée des données sont effectuées sur le client central, un certain nombre de serveurs spécialisés et potentiellement décentralisés pour la construction et l'annotation automatique des données sont disponibles.

3.Construction et annotation des données

Bien que l'objectif principal de DiET soit de proposer une série d'outils pour la construction des bancs de test, il met également à la disposition de l'utilisateur une quantité importante de données pour les trois langues du projet. Ces données sont essentielles pour valider au sein du projet les différents outils dans leur adéquation et leur facilité d'utilisation. De plus, en l'absence de source commerciale suffisante pour de telles données, une extension continue de celles-ci reposera certainement sur les contributions de l'ensemble de la communauté scientifique du TALN.

La construction des données dans DiET suit la définition des jeux de phrases-test développée dans TSNLP (Estival & Lehmann 1997). Les données stockées dans la base sont constituées d'une phrase (d'un syntagme ou d'une suite de phrases) associée à différents types d'informations, à savoir les annotations. La méthodologie de TSNLP consiste en la variation progressive et systématique de phrases-test de base. Ces phrases-test de base sont elles-mêmes construites de manière systématique et aussi complète que possible à partir d'un répertoire de phénomènes, et leur variation s'effectue en modifiant la valeur des attributs considérés comme pertinents pour ce phénomène. De telles variations peuvent conduire à la définition de phrases-test négatives, qui n'apparaissent normalement pas dans les corpus, mais qui demeurent indispensables pour un diagnostic. En fonction de l'application à évaluer, une phrase-test négative ne se limite pas à une phrase agrammaticale, mais peut par exemple contenir une tournure interdite dans le cadre d'un langage contrôlé (comme les constructions passives ou interrogatives). De plus amples détails seront données plus loin sur les données spécifiques au français.
Le rôle de ces annotations est multiple. En plus de décrire les phénomènes linguistiques, la plupart d'entre elles sont essentielles pour rechercher une phrase-test ou un ensemble de phrases dans la base de données. Certaines permettent également d'établir un lien entre les phrases-test et les corpus, par exemple pour estimer la fréquence d'un phénomène dans un texte (*cf*. la section sur l'analyse de corpus). De telles relations peuvent être établies en comparant les annotations structurales (comme les séquences de catégories grammaticales) d'une phrase-test aux informations présentes dans un corpus étiqueté. La séquence d'étiquettes de la phrase-test peut alors être considérée comme une requête sur le corpus. Enfin, pour la procédure d'évaluation elle-même, les phrases-test peuvent être annotées en précisant les résultats attendus à la suite de leur traitement par un système particulier dans la phase de test (traduction souhaitée pour un système de traduction par exemple). Des mesures de performance et des statistiques sur les bancs de test peuvent ainsi être ajoutées à des fins de comparaison entre différents systèmes.
Les annotations ont généralement différents buts : il est donc pertinent de considérer aussi leur contenu. Vues sous cet angle, les annotations de DiET peuvent être réparties en quatre catégories :

3.1Caractéristiques linguistiques

Parmi les annotations purement linguistiques, on peut distinguer celles qui concernent l'ensemble de la phrase-test, comme la langue et la grammaticalité.
De plus, toutes les phrases-test peuvent être classées en fonction du phénomène linguistique qu'elles illustrent. Les annotations comprennent donc le nom et une description de ce phénomène au sein d'une classification hiérarchique, et les propriétés caractéristiques, comme les valeurs de nombre et de personne pour un phénomènesyntaxique comme accord entre sujet-verbe.
D'autres types d'annotations ne concernent que des sous-parties de la phrase-test, comme les informations morphologiques, syntaxiques ou sémantiques. Les annotations morphologiques comprennent la définition d'une classe d'ambiguïté grammaticale unique à une unité lexicale. Les informations syntaxiques suivent le schéma défini au cours de TSNLP, et prennent la forme d'un arbre ou les nœuds non-terminaux correspondent à des constituants (syntagmes) et les arcs à des fonctions grammaticales (sujet, objet, etc.). Enfin, les annotations sémantiques servent à préciser la direction et le type des relations entre les segments de phrase-test (anaphore, cataphore, etc.)

3.2Informations liées à une application spécifique

Les attributs liés à une application spécifique peuvent, de façon triviale, associer à une phrase-test le nom d'une application, permettant ainsi à l'utilisateur de rechercher rapidement dans la base un banc de test pertinent. Ils peuvent aussi indiquer le résultat attendu par le traitement de cette

phrase :par exemple, pour les outils de vérifications de grammaire ou de langage contrôlé, on peut définir le type d'erreur associé à une phrase-test agrammaticale ou mal structurée, ou bien encore la phrase grammaticale ou acceptable correspondante (qui sera comparée aux suggestions proposées par le système testé). Pour les outils d'analyse syntaxique, une annotation peut contenir le nombre attendu d'ambiguïtés grammaticales d'une phrase. Pour les outils de traduction, on peut annoter les phrases-test en précisant leur(s) traduction(s).

3.3Informations liées à un corpus particulier

Les annotations liées à un corpus sont celles qui établissent un lien entre les phrases-test et des corpus liés à un domaine et/ou une application précise. Les informations associées peuvent indiquer le nombre d'occurrences (et donc la pertinence) d'un phénomène associé à une phrase-test dans un texte, ou encore pointer vers les phrases elles-mêmes, et ainsi contenir des exemples « réels » de ce phénomène. Ces informations seront détaillées dans la section 5.

3.4Informations liées à une procédure de test particulière

Les annotations liées à une procédure d'évaluation aident l'utilisateur à garder une trace des résultats obtenus : elles décrivent le scénario d'évaluation, le type d'utilisateur (client, développeur), le nom et le type du système testé, le but de l'évaluation (diagnostic, comparaison, progression), les conditions (boîte noire, boîte blanche), les critères et les mesures de l'évaluation, etc. De plus, seront ainsi stockés les réponses réelles du système, ainsi que les résultats attendus, et une comparaison sera effectuée afin de définir de façon générale les performance du système.
En plus de ces annotations spécifiques, la base de données contient des méta-informations à propos du banc de test dans son intégralité, comme les listes de vocabulaire contenu dans les phrases-test, les listes d'étiquettes ou encore la terminologie générale des annotations.

3.5Exemple

La figure 2 donne un exemple d'annotation linguistique d'une phrase-test pour le cas des phénomènes syntaxiques. Aux phrases-test sont associées des annotations diverses, concernant l'origine de la phrase, les phénomènes qu'elle représente, son statut grammatical/agrammatical, le nombre de mots qu'elle comprend, etc. Une description plus complète des différents types d'annotations peut être trouvée dans (Estival & Lehmann 1997).

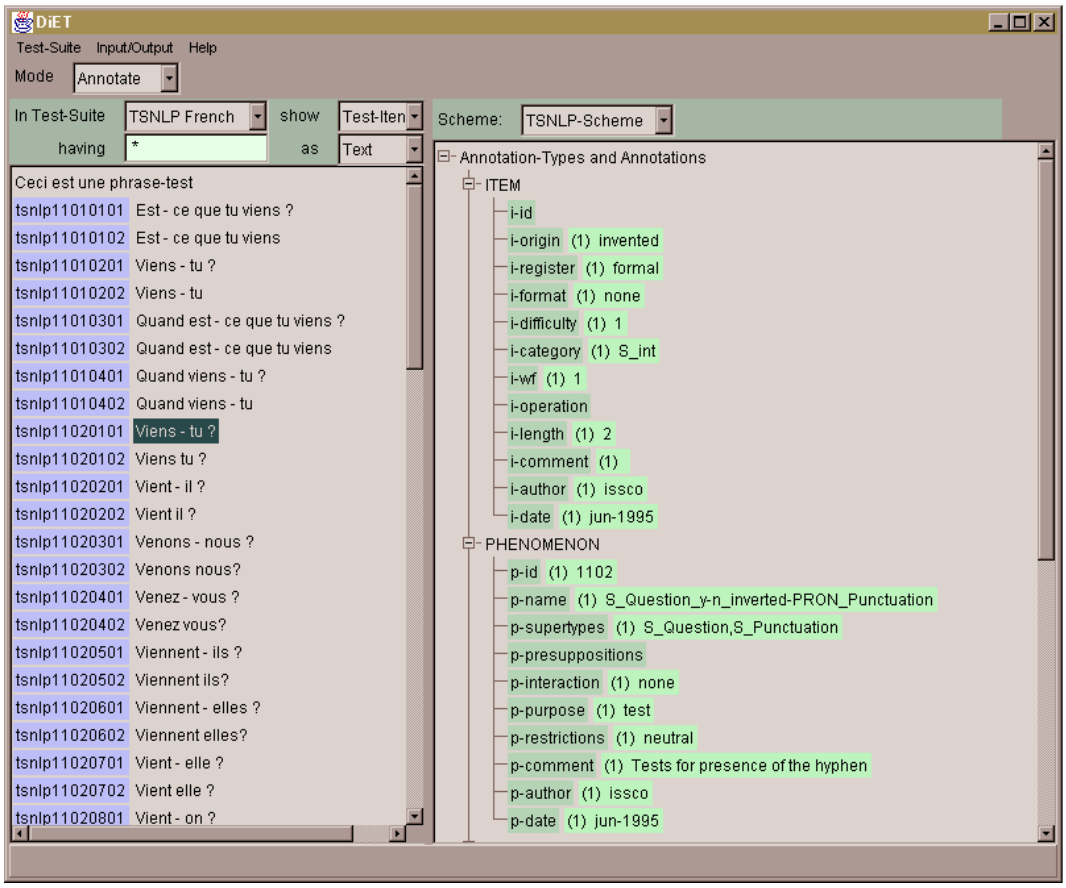

Figure 2: Exemple d'annotation de phrase-test

4.Données de test pour le français

Les phrases-test du projet TSNLP se concentraient dans les phénomènes purement syntaxiques. Ces dernières ont été reprises et étendues dans DiET, et de nouvelles données ont été définies à d'autres niveaux de l'analyse linguistique : la morphologie et la sémantique.

4.1Morphologie

Les données morphologiques dans DIET ont une double fonction : elles permettent, d'une part, d'évaluer les analyseurs morphologiques et, d'autre part, d'associer aux mots les propriétés morphologiques qui déterminent le remplacement lexical (catégorie syntaxique, nombre, genre, personne, temps, mode, etc.).
Ces données se présentent comme un ensemble de classes d'équivalence, qui classifient les formes fléchies en fonction de leur propriétés morphobgiques. En voici quelques exemples :

Verbe (conditionnel présent (première ou deuxième) personne singulier)
Exemples : abaisserais abandonnerais abasourdirais abâtardirais abattrais abcéderais abdiquerais abêtirais abhorrerais abîmerais...

Nom ((masculin ou féminin) pluriel)

Exemple : aborigènes acrobates actionnaires adeptes adultes adversaires aéronautes agnostiques agronomes aides

Il faut noter qu'ici la notion d'équivalence ne couvre pas les ambiguïtés qui sont établies entre deux catégories grammaticales (c'est-à-dire les mots polyfonctionnels), mais uniquement sur les flexions.
Pour le français, ces classes ont été extraites du dictionnaire *Mmorph*, développé dans le cadre du projet Multext (Armstrong, 1996), et sont au nombre de 250. Une telle classification des unités lexicales simples du français constitue en quelque sorte le minimum requis d'une application effectuant ce type d'analyse.

4.2Syntaxe

Les phrases-test de la base de données TSNLP traitaient un nombre limité de phénomènes syntaxiques considérés comme les plus cruciaux pour les applications de TALN dans les trois langues : la complémentation, l'accord, la modification, la diathèse, les flexions verbales, la coordination, la négation, et certains phénomènes extra-grammaticaux comme la ponctuation et les abréviations.
DiET s'est quant à lui fixé les buts suivants :

• Correction et validation des données et des annotations provenant de TSNLP.

• Ajout de nouveaux traits descriptifs. Parmi ceux-ci, notons : l'attribut *i-operation*, qui décrit, pour une phrase-test négative, l'opération qui la relie à sa forme acceptable (remplacement, ajout ou suppression d'une unité, etc.), ainsi que *i-difficulty*, une estimation de la difficulté d'un traitement spécifique de la phrase-test (repérage des constituants, traduction, etc.)

• Élaboration de nouvelles données. Celles-ci ont pour but de couvrir de nouveaux phénomènes (comme les phrases nominales complexes ou les clauses adverbiales) ou d'approfondir des aspects déjà décrits (comme la coordination ou la complémentation).

• Représentation de chaque phrase-test par un arbre syntaxique, comme présenté dans la figure 3.

La construction de ces données a été effectuée directement dans le système DiET afin de tester cet outil de manière plus approfondie.

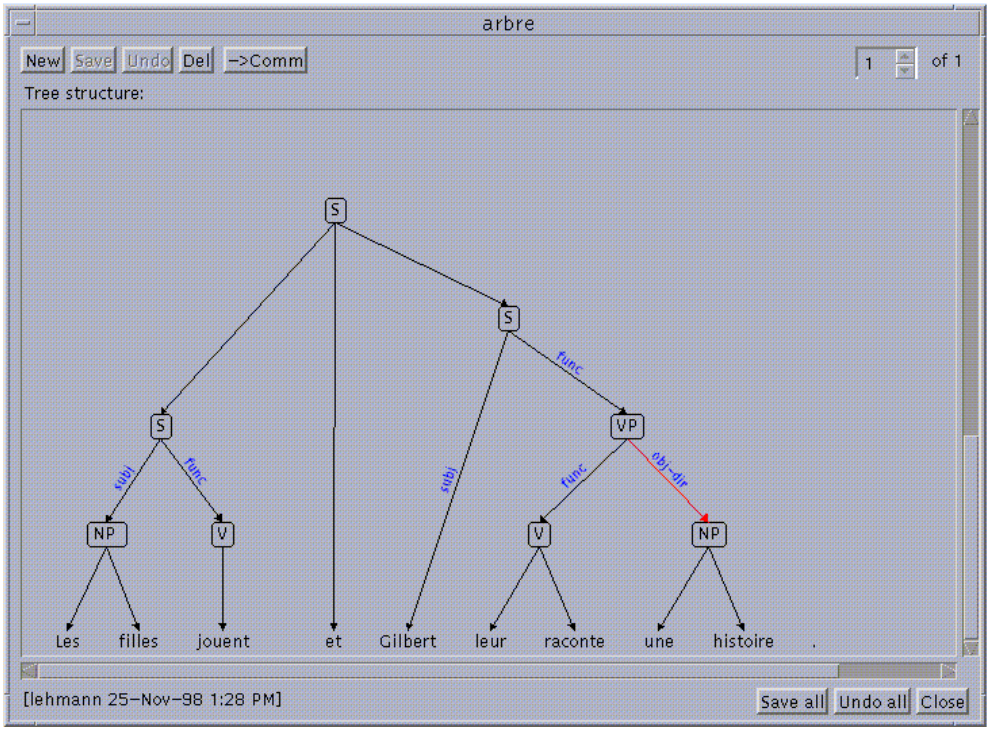

Figure 3: Exemple d'annotation syntaxique de phrase-test

4.3Discours et sémantique

Les phénomènes du discours traités dans DiET ont été choisis tant pour le consensus établi quant à leur représentation linguistique que pour leur traitement effectif dans les systèmes actuels de TALN (Lewin *et al*, 1999). Il s'agit de l'anaphore (1a), de l'ellipse (1b) et de la synonymie (1c) :

(1a) *Une fille chante. Elle regarde un garçon. Il siffle.*
(1b) *Pierre aime les fraises et Marie le chocolat.*
(1c) *Samedi il est parti.*
Il est parti samedi.

Celles-ci se distinguent des propriétés syntaxiques par différents aspects : elles sont inter-phrastiques. Par là-même, elles complexifient le remplacement lexical. Par exemple, 'il' peut être substitué à 'elle' en (1a), mais le contenu du texte en est complètement modifié. Enfin, la résolution des anaphores peut nécessiter des connaissances pragmatiques, comme en (2).

(2) *Je suis rentrée dans le café. Il m'a demandé ce que je voulais boire.*

Ces caractéristiques expliquent les lignes de conduite que nous avons définies pour la construction des phrases-test :

- les propriétés discursives retenues doivent pouvoir être résolues à l'aide de connaissances linguistiques ;

- les phrases de test sont des textes, c'est-à-dire des séquences de phrases ;

- les parties de la phrases intéressantes (antécédent, anaphorique ou cataphorique, etc.) doivent pouvoir être marquées de manière non-ambiguë.

• elles doivent être annotées avec les propriétés linguistiques qui caractérisent le phénomène en question. Celles-ci permettent, en aval, d'élaborer les phrases-test de manière systématique et, en amont, de fournir les informations nécessaires pour leremplacement lexical.

Pour l'anaphore par exemple, les annotations sont résumées dans le tableau suivant :

Attribut	**Valeurs**
phenomenon_name	anaphore
antecedent_category	phrase, syntagme verbal, nominal, …
antecedent_subcategory	nom, nom-propre, ...
antecedent_function	sujet, objet, ...
antecedent_animateness	animé, non-animé, ...
antecedent_type	défini, indéfini
antecedent_gender	masculin, féminin
antecedent_number	singulier, pluriel
anaphor_category	pronom
anaphor_subcategory	personnels, démonstratifs, ...
anaphor_function	sujet, objet, ...
anaphor_genre	masculin, féminin
anaphor_number	singulier, pluriel
grammaticality	correct, incorrect, douteux
distance	nombre de phrases séparant l'antécédent du marqueur anaphorique
valid_antecedent	nombre d'antécédents valides dans la phrase
link	repérage graphique de l'antécédent et du marqueur anaphorique
type_of_link	spécifie si le marqueur anaphorique peut être remplacé par l'antécédent sans modifierle sens ou l'acceptabilité de la phrase
competitor_antecedent	nombre d'autres candidats pour l'antécédent
competitor_via	trait grammatical (genre, nombre) expliquant l'invalidité des *competitor antecedent*

Par exemple, pour la phrase (3),

(3) LES FILLES jouent et Gilbert LEUR raconte une histoire.

nous avons :

Attribut	**Valeurs**
phenomenon_name	anaphore
antecedent_category	syntagme nominal
antecedent_subcategory	nom
antecedent_function	sujet
antecedent_animateness	animé
antecedent_type	défini
antecedent_gender	féminin
antecedent_number	pluriel
anaphor_category	pronom

anaphor_subcategory	personnel
anaphor_function	objet indirect
anaphor_genre	féminin
anaphor_number	pluriel
grammaticality	correct
distance	0
valid_antecedent	1
link	*cf.* figure 4
type_of_link	*String* (l'anaphore réfère à l'intégralité de l'antécédent)
competitor_antecedent	0
competitor_via	

Le lien (*link*) entre le marqueur anaphorique (*leur*) et l'antécédent (*les filles*) est représenté dans l'interface graphique de DiET par un code de couleur, comme indiqué dans la figure 4.

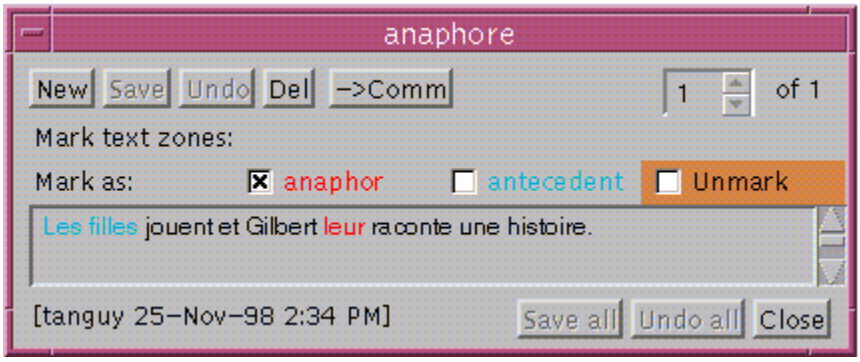

Figure 4: Exemple d'annotation sémantique de phrase-test anaphorique

La base de données pour l'anaphore en français contient actuellement 1500 entrées. Celles-ci testent systématiquement les différentes valeurs des attributs et leur combinaison.

5.Adaptation des données

Les données construites systématiquement peuvent être utiles pour un diagnostic, mais ne constituent pas une base suffisante pour évaluer l'adéquation d'un système face à un but et un corpus précis. Un système doit être testé dans le contexte où il aura à évoluer. Afin d'aider l'utilisateur à définir des bancs de test sur mesure, un certain nombre d'outils servent à adapter et modifier les phrases-test. Le but est ici de combler le fossé entre des données de test artificielles et isolées et des données empiriques de corpus.

Pour une application donnée, seules les phrases-test déclarées pertinentes pour son domaine d'utilisation doivent être utilisées. Si l'utilisateur dispose d'un corpus sur lequel un système de traitement de la langue naturelle doit être évalué, la sélection des phrases-test doit se faire en fonction des phénomènes repérés dans ce corpus, ou du moins en fonction de ceux qui sont susceptibles de s'y trouver. A cette fin, des outils d'extraction des propriétés caractéristiques d'un corpus sont proposés.

De plus, une fois la sélection des phrases-test pertinentes effectuées, l'utilisateur doit avoir la possibilité de les adapter afin qu'elles reflètent la réalité des données que le système aura à traiter.

Un module de remplacement lexical est donc disponible, pour adapter les phrases à un vocabulaire précis.

5.1 Caractéristiques de corpus

Les outils utilisés pour établir les propriétés caractéristiques d'un corpus se basent sur les technologies actuelles d'analyse de la langue naturelle : analyse morphologique, étiquetage morpho-syntaxique, mesures statistiques, et recherche de séquences. La qualité de cette analyse dépend bien entendu de celle des premières phases (analyse morphologique notamment) : cet outil bénéficiera donc beaucoup d'un étiquetage préexistant du corpus, que celui-ci soit effectué manuellement ou automatiquement, à l'aide d'un outil spécifiquement entraîné pour ce corpus. Dans ce but, DiET propose une chaîne de traitement complète, et entièrement automatique, basée pour la langue française sur les outils développés lors du projet *Multext* (Armstrong, 1996).
Au fur et à mesure que les différents modules d'analyse sont appliqués à un corpus, de nouvelles informations sont disponibles, alors que la fiabilité de ces informations décroît. On peut distinguer trois différents niveaux d'analyse :

5.1.1 Analyse de surface

Informations :
L'étape de base de l'analyse consiste à établir une liste de caractéristiques concernant la taille du corpus et des phrases, les caractéristiques typographiques (mots en majuscule, fréquence des valeurs numériques, etc.), la ponctuation. De même, une table de fréquence des occurrences de mots dans le corpus est établie, ainsi que les caractéristiques qui en résultent sur la richesse du vocabulaire employé dans le texte (rapport type/occurrence, indices stylométriques, etc). Cette première analyse, pour pauvre qu'elle soit, est toutefois parfaitement objective. Un aperçu de ces données pour un cours texte français est présenté dans la figure 5. On peut y trouver, dans l'ordre, les informations suivantes :

- Taille des paragraphes en nombre de phrases (minimal, maximal et moyen)
- Taille des phrases en nombre de mots ou de signes (*tokens*) (mots et signes de ponctuation)
- Nombre et type des signes de ponctuation
- Nombre de signes
- Types et fréquences des catégories morphosyntaxiques
- Vocabulaire : types et fréquences

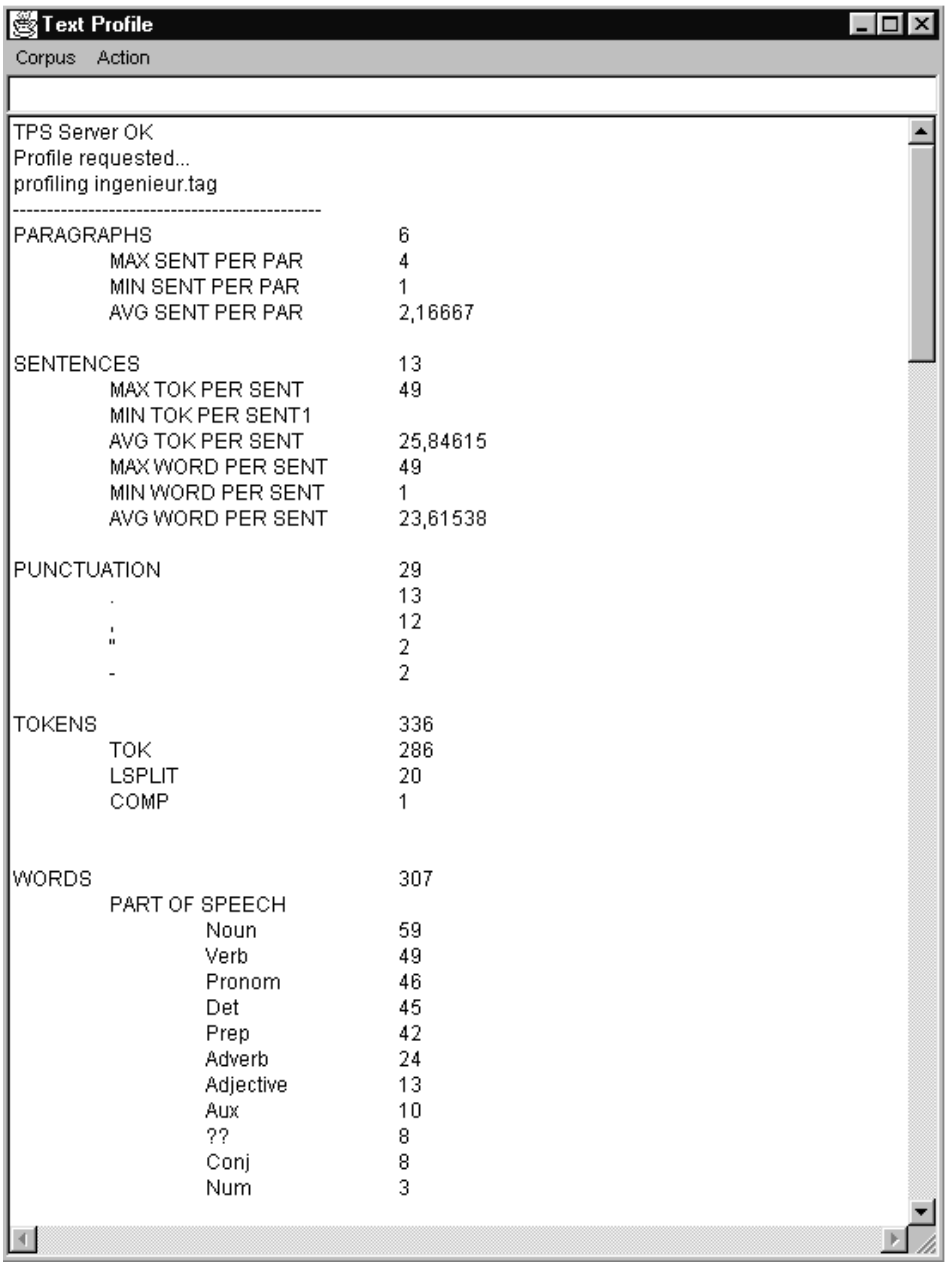

Figure 5: Exemple de caractéristiques d'un corpus : données générales

Liens avec l'évaluation :
Certains phénomènes syntaxiques ont des indices de surface qui sont repérés à ce niveau. Ainsi, l'absence ou la présence de certains signes de ponctuation (? ou !) dans le corpus permet de juger de la pertinence de certaines phrases-test, comme les phrases interrogatives ou exclamatives. La longueur des phrases du corpus est également un indice de la complexité des tournures syntaxiques employées.
L'inventaire du vocabulaire du texte permet d'établir un coefficient de corrélation entre ce corpus et le vocabulaire servant à définir les phrases-test. Ces données sont également utilisées dans la suite pour le remplacement lexical.

Enfin, certaines unités lexicales, notamment celles appartenant aux classes fermées (conjonction, préposition, certains adverbes) permettent également d'estimer la présence et la fréquence de certaines constructions syntaxiques.

5.1.2Analyse morphologique

Informations :
Chaque unité lexicale est analysée pour en identifier le lemme, la catégorie syntaxique et les propriétés morphologiques (genre, nombre, temps verbal, etc.). Plusieurs caractéristiques importantes du corpus sont extraites à ce stade, notamment la répartition des différentes catégories syntaxiques (noms, verbes, adverbes), la présence de flexions particulières (subjonctifs, impératifs, participes présents, etc.). Une seconde table de fréquence résume le vocabulaire du corpus, en ne prenant en compte que les lemmes (ou formes non fléchies).

Liens avec l'évaluation :
Ici également, le simple dénombrement de certaines catégories grammaticales permet d'estimer la pertinence de certaines phrases-test. Ceci est surtout vrai pour les différentes formes verbales, certains modes (subjonctif, impératif) correspondant à des ensembles complets de phrases-test dont l'utilisation peut ainsi être évitée.
Mais l'information la plus importante extraite du corpus durant cette étape de l'analyse est la suite de marqueurs morpho-syntaxiques pour chaque phrase, comme présentée dans la figure 6.Le chiffre de la colonne de gauche indique le nombre d'occurrences dans le corpus (ici un manuel technique) de phrases ayant la structure décrite par la suite de catégories morpho-syntaxiques (colonne de droite).

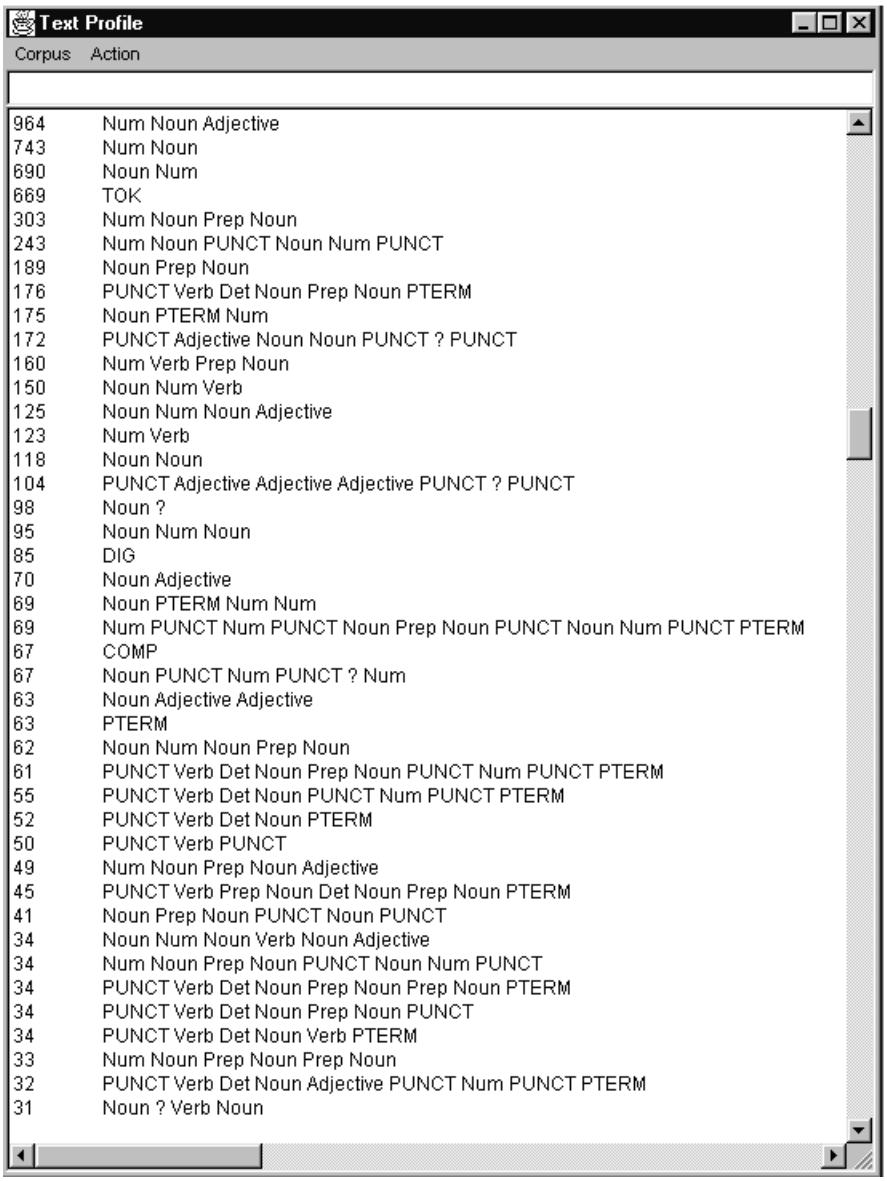

Figure 6 : Exemple de caractéristiques d'un corpus : analyse morphologique

En ajoutant à l'outil d'analyse un moteur de recherche de telles séquences dans le corpus, on peut établir un lien direct entre les phrases-test et le texte lui-même. Pour un grand nombre de phénomènes syntaxiques, il est aisé de définir, par un langage simple d'expressions régulières, des schémas de suite de marqueurs qui, s'ils sont repérés dans le corpus, indiqueront non seulement la présence d'une occurrence de phénomène correspondant, mais permettront également d'obtenir un exemple «réel» de ce phénomène. Par exemple, pour le phénomène correspondant à l'accord entre deux adjectifs coordonnés dans un syntagme nominal, il suffit de rechercher une séquence du type :

Déterminant + Nom + Adjectif + Conjonction de coordination + Adjectif

Une grande partie des phénomènes syntaxiques définis dans la base de données de DiET peuvent se faire associer de telles séquences. Une interface graphique permet à l'utilisateur qui construit des données de test de définir de telles requêtes, basées sur des mots ou des expressions régulières de mots, ou des disjonctions de catégories morpho-syntaxiques, comme présenté dans la figure 7. Le

résultat d'une telle requête peut être directement intégré à la base de données en tant que nouvelle phrase-test. Le moteur de recherche, spécifiquement conçu pour le format d'annotation des outils Multext, s'inspire d'outils de recherche de séquences dans les corpus annotés, comme GSearch (Corley et al, 1997) ou XKwic (Christ, 1993). Plus de détails sur cet aspect de DiET sont présentés dans (Lewin *et al*, 1999).

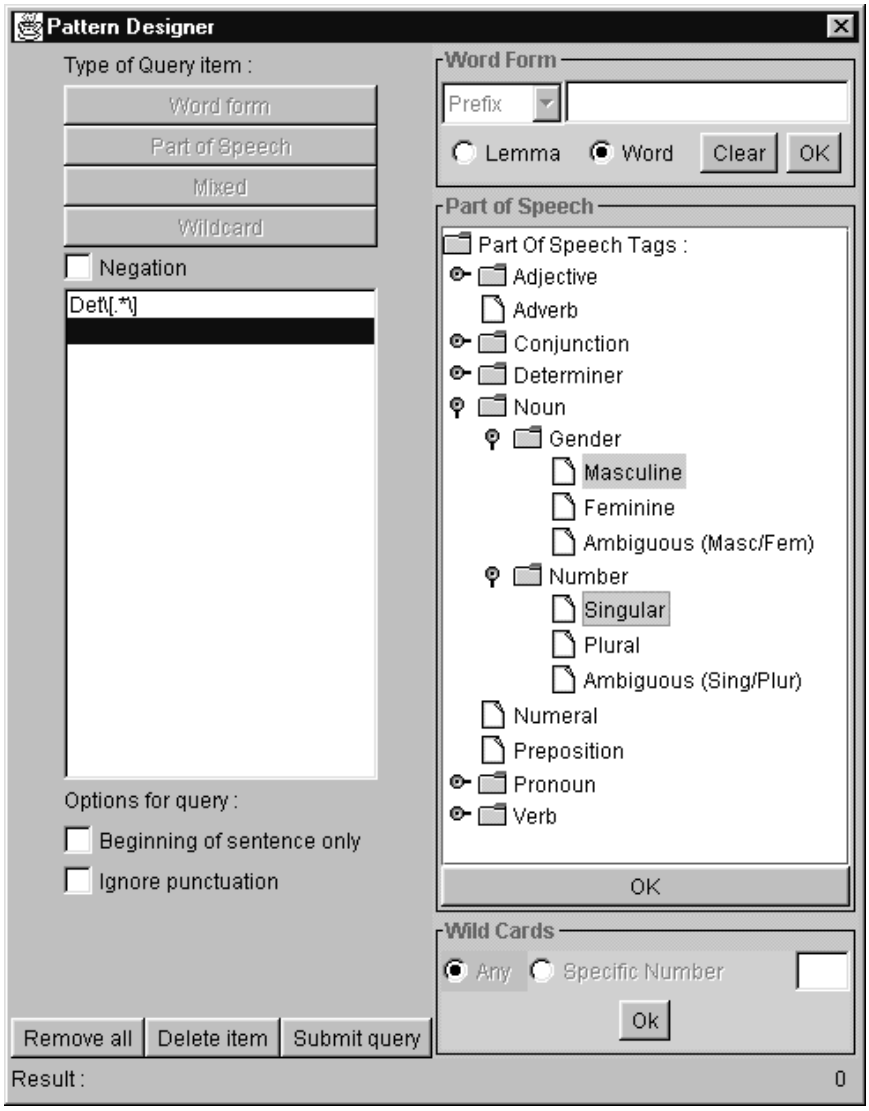

Figure 7 : Interface pour la construction de requêtes

Ainsi, sur un banc de test complet, il est possible, de façon automatique, de parvenir à un sous-ensemble de phrases-test pertinentes pour un texte donné.
Malheureusement, certains phénomènes plus généraux ou plus complexes ne peuvent être efficacement repérés par cette méthode, comme par exemple les phénomènes de coordination entre des syntagmes nominaux complexes.

5.1.3Analyse syntaxique

Pour franchir les limites de l'analyse précédente, il est donc nécessaire d'approfondir encore l'analyse du corpus afin d'y identifier les composants syntaxiques principaux.

Informations :

Le résultat de cette analyse apporte donc un second niveau de structuration sur le corpus, en précisant le découpage des phrases en syntagmes. A ce stade plus encore que dans les deux précédents, les technologies employées ne peuvent garantir de résultats face à un corpus non spécifiquement analysé. Rappelons tout de même que les facilités proposées ici le sont à titre indicatif dans l'établissement d'un banc de test spécifique, et ne sauraient, pour atteindre un bon niveau de précision, se passer d'un certain nombre de corrections par l'utilisateur.

Liens avec l'évaluation :

L'apport de ce type d'informations est d'enrichir la liste des phénomènes pouvant être repérés dans le corpus. Il est en effet possible à ce stade de repérer des structures comme «Harry rencontre Sally et Pierre rencontre Marie» comme étant une coordination de syntagmes verbaux, et non de syntagmes nominaux, comme cela aurait été le cas si l'analyse s'était limitée à l'étape précédente.

5.2Remplacement lexical

Une autre méthode pour resserrer les liens entre les bancs de test et une application précise est le remplacement lexical au sein des phrases-test. Cette opération sur les données de la base de DiET est envisagée pour deux raisons.

Tout d'abord, une phrase-test peut contenir un vocabulaire inconnu à l'application qui va être testée. Ceci n'est valable que pour certaines applications définies sur un champ lexical restreint, technique par exemple. Dans ce cas, la comparaison des listes de vocabulaires (celle de la base de donnée de DiET et celle extraite, par exemple, lors de l'analyse de corpus) peut permettre d'identifier les phrases-test qui contiennent des mots inconnus et de remplacer ceux-ci par des unités équivalentes, mais identifiables par l'application.

Ensuite, le remplacement lexical peut servir à étendre volontairement les données de la base de test, en fabriquant de nouvelles phrases-test pour le même phénomène. DiET se concentre toutefois sur la première utilisation de cet outil.

Le principe de cette méthode, développée dans (Kiss & Steinbrecher, 1998), est que les classes d'équivalence utilisées pour le remplacement prennent la forme d'un lexique structuré, dans lequel les caractéristiques morphologiques et grammaticales entretiennent des relations d'équivalence et de généralisation. Il est donc possible de remplacer un terme par un terme qui possède la même description, ou une description plus générale dans ce lexique. Dans le cas d'un remplacement au sein d'une phrase agrammaticale, toutefois, une généralisation peut induire une modification de la grammaticalité. Par exemple, «enfants» est moins général que «fils», ce dernier pouvant être un nom pluriel ou singulier. Donc, le remplacement de «enfants» dans «mon enfants est venu» (qui est une phrase agrammaticale) par «fils» entraîne la création d'une phrase correcte. Bloquer ce genre de remplacement se fait par la prise en compte, dans le segment initial, de l'identité de l'élément à la source de l'agrammaticalité (ce qui peut être contenu dans l'annotation).

6.Évaluation

Une procédure d'évaluation à l'aide de DiET peut se résumer à l'aide des grandes étapessuivantes :

• Configuration du banc de test.

  • Identification des phénomènes pertinents pour l'outil : Cette configuration peut se faire sur la base d'informations générales de l'utilisateur de l'outil, qui peut dès lors filtrer les phénomènes linguistiques en fonction de ses connaissances. Dans le cas où un corpus significatif est disponible, la procédure peut s'automatiser en analysant celui-ci à l'aide des outils de DiET (*cf* section 3), qui proposeront alors une sélection des phrases-tests les plus pertinentes. À ce stade DiET peut également donner un certain nombre d'indices quant aux phénomènes identifiés dans le corpus mais non représentés explicitement dans la base de phrases-test. Ceci peut donc conduire à une phase intermédiaire de définition de nouvelles phrases-test.

  • Affinement des phrases-tests : Une fois les phrases-test identifiées, il est possible d'adapter leur vocabulaire à l'aide de l'outil de remplacement lexical.

  • Pondération des phrases-test : En fonction de l'importance des phénomènes sur la base desquels l'outil va être testé, l'utilisateur peut pondérer les futurs résultats de la phase de test.

• Phase de test : Les phrases-test sont passées en séquence à l'outil, et ses réponses sont stockées dans la base de données de DiET. Dans le cas où un jugement extérieur est requis, l'utilisateur aura à préciser certaines valeurs pour qualifier le résultat.

• Synthèse : L'ensemble des résultats est analysé en tenant compte des pondérations définies par l'utilisateur, et un bilan général de la phase de test peut lui être présenté.

Les outils de DiET présentés ici ont été utilisés pour l'évaluation d'outils d'analyse de corpus, notamment un correcteur grammatical et un vérificateur de langage contrôlé. Une description de ces procédures d'évaluation est présentée dans (Regnier et al, 1999).

7.Conclusion

Un des buts ultimes de DiET est de proposer aux professionnels (développeurs, clients, consultants) une méthodologie d'évaluation et un ensemble d'outils aisément reconfigurables vers d'autres domaines ou d'autres applications. En organisant les données sur un serveur central, capable d'emmagasiner et de restituer des données de test, nous espérons permettre l'échange d'informations entre les différents intéressés. Toutefois, DiET ne peut proposer que les bases techniques pour une telle entreprise, dont le succès dépendra fortement de l'intérêt exprimé par l'ensemble de la communauté.

8.Remerciements

Nous tenons à remercier l'ensemble des participants du projet Diet : Klaus Netter, Judith Klein et Tillmann Wegst de DFKI (Saarbrücken), Tibor Kiss de IBM (Heidelberg), David Milward et Ian Lewin de SRI (Cambridge), Reinhard Schäler et Bernice McDonagh de LRC (Dublin), Sylvie Regnier-Prost, Frédéric Joffroy et Nathalie Briot de Aérospatiale (Paris), Laurence Vandenbroucke de ISSCO (Genève).

9.Références

• Armstrong S. 1996. *MULTEXT : Multilingual Text Tools and Corpora*. In : H. Feldweg et E. W. Hinrichs (eds). Lekicon und Text Max Niemeyer, pp 107-120.

• Bouillon P, *et al* 1997. *Développement de lexiques à Grande Échelle*. Actes des journées LTT, Tunis, 1997.

• Christ O. 1993. *The XKwic User Manual*. Institut für maschinelle Sprachverarbeitung. Unisersität Stuttgart.

• Cooper *et al.* 1996. *Using the Framework FraCaS.* Deliverable D16. Edinburgh, 1996. http://www.cogsci.ed.ac.uk/~fracas/

• Corley *et al*. 1997. *Corset II User Manual*. University of Edinburgh. 1997.

• Estival D & Lehmann S. 1997. *TSNLP : Des jeux de phrases-test pour le TALN*. TAL, Volume 38, pp. 155-171.

• King M. 1996. *Evaluating Natural Language Processing Systems*. in: Special edition of Communications of the ACM on Natural Language Processing, Vol. 39, No 1, Jan. 1996, pp. 73-79.

• Kiss T, *et al* 1997. *The DiET User Requirements Analysis*. D1.1 IBM Heidelberg.

• Kiss T & Steinbrecher D. 1998. *Lexical Replacement in Test Suites for the Evaluation of Natural Language Applications.* In : Proceedings of the First International Conference on Language Ressources and Evaluation (LREC), Granada, May 1998

• Lewin I *et al*. 1999. *Discourse Data in DiET.* Proceedings of the EACL workshop on Linguistically Interpreted Corpora (LINC). Bergen, 1999.

• Maegaard B *et al*. 1996. *TEMAA - A Testbed Study of Evaluation Methodologies : Authoring Aids*. Rapport Final, 1996.

• Netter K *et al*. *DiET : Diagnostic and Evaluation Tools for Natural Language Processing Applications.* In : Proceedings of the First International Conference on Language Ressources and Evaluation (LREC), Granada, May 1998, pp. 573-579.

• Rayner M, Bouillon P & Carter D. 1995. *Using Corpora to develop Limited-Domain Speech Translation Systems*. Proceedings of Translating and the Computer, 17 (ASLIP), Novembre 1995.

• Regnier S *et al*. 1999. *DiET report on Evaluation and Verification*, DiET D4, 1999.

• Skut W *et al*. 1997. *An Annotation Scheme for Free Word Order Languages.* Proceedings of ANLP, 88-96, 1997.