TiLT : plate-forme pour le traitement automatique des langues naturelles

Johannes Heinecke[*] — Grégory Smits[**] — Christine Chardenon[*] — Emilie Guimier De Neef[*] — Estelle Maillebuau[*] — Malek Boualem[*]

[*] *Orange Labs*
2, avenue Pierre Marzin
F-22307 Lannion cedex

{johannes.heinecke, christine.chardenon, emilie.guimierdeneef, estelle.maillebuau, malek.boualem}@orange-ftgroup.com

[**] *ENSSAT*
6, rue Kerampont
F-22300 Lannion

gregory.smits@univ-rennes1.fr

RÉSUMÉ. *Cet article décrit une plate-forme de TALN, modulaire et multilingue, enrichie d'un système de contrôle basé sur l'aide multicritère à la décision. La présentation est complétée par une description des données linguistiques utilisées ainsi que des applications basées sur cette technologie.*

ABSTRACT. *This article describes a modular and multilingual NLP platform, which is enriched by a system of multicriteria decision-aid. Further we describe the linguistic data used by this platform as well as the applications based on its technology.*

MOTS-CLÉS : *boîte à outils TALN, architecture modulaire et symbolique, analyse lexicale, syntaxique et sémantique, ontologies, multilinguisme, aide multicritère à la décision, ressources linguistiques riches, indépendance données/traitements.*

KEYWORDS: *NLP toolbox, modular and data-driven architecture, lexical, syntactic and semantical analysis, ontologies, multilinguism, multicriteria decision-aid, rich linguistic ressources, separation of processing and data.*

1. Introduction

France Télécom mène des activités de R & D sur le traitement automatique des langues depuis près de deux décennies. Les travaux sur le traitement du langage naturel écrit sont menés notamment au sein de l'équipe Langues Naturelles de Orange Labs[1]. Compte tenu de la dimension internationale de France Télécom et de la popularisation, de plus en plus croissante, des moyens de communication, faciliter le traitement et l'accès à l'information dans un grand nombre de langues revêt un intérêt particulier dans les activités de R & D. Les solutions liées au traitement automatique des langues naturelles sont ainsi mises à la disposition des utilisateurs (accès aux annuaires et aux bases de données, recherche d'information, etc.). Pour répondre aux besoins en matière d'accès à l'information, une plate-forme industrielle de TALN, baptisée TiLT (traitement linguistique des textes), a été mise en place à Orange Labs (Guimier de Neef *et al.*, 2002). Cet article est orienté vers une approche descriptive de TiLT qui tente d'en souligner un certain nombre de caractéristiques liées aux problématiques architecturales et méthodologiques. Il est à noter que cette description concerne essentiellement l'analyse linguistique et n'aborde que très sobrement la génération. Après avoir présenté les choix architecturaux, les différents modules de la plate-forme et les ressources linguistiques nécessaires à son fonctionnement, nous nous intéressons à l'interopérabilité des différents composants à travers le prisme du contrôle des processus d'analyse linguistique réalisables par la plate-forme. Enfin, nous terminons par une présentation des applications opérationnelles utilisant TiLT.

2. Architecture de la plate-forme TiLT

2.1. *Choix architecturaux*

L'interprétation linguistique d'un énoncé écrit est souvent présentée comme reposant sur un ensemble de niveaux d'analyse et de connaissances. L'une des plus grandes difficultés soulevées par le TALN est la gestion de l'indéniable interdépendance et complémentarité de ces niveaux d'analyse. De nombreux systèmes se sont inspirés des modèles cognitifs issus des travaux en psycholinguistique pour concevoir des architectures informatiques de traitement respectant les propriétés de parallélisme, de complémentarité et d'interdépendance des connaissances et des étapes d'analyse.

Cependant, bien que théoriquement justifiables, ces développements se sont en pratique heurtés à de nombreuses difficultés, telles que la gestion des communications entre les niveaux d'analyse, la formalisation des structures de connaissances ou encore l'efficacité des algorithmes de traitement. Cette vision théorique s'oppose ou plutôt se complète par une approche plus pragmatique considérant le TALN comme un ensemble de technologies au service des applications. S'inscrivant dans cette dernière vision et sans doute influencés par des impératifs liés au contexte industriel,

1. Anciennement CNET - Centre National d'Études des Télécommunications et ensuite France Télécom R & D

nous nous sommes basés sur des modèles informatiques « classiques » pour concevoir une architecture plus facile à développer, à maintenir et à étendre avec de nouvelles fonctionnalités.

Ainsi, la plate-forme TiLT a été développée en privilégiant des propriétés d'adaptabilité, d'extensibilité et de maintenance. Ceci a conduit au morcellement du processus d'analyse en modules de traitement. L'organisation de l'application des différents modules de traitement est gérée par une approche séquentielle, où différents modules de traitement spécifiques peuvent être appliqués successivement pour atteindre le niveau d'interprétation souhaité. Afin de pallier les limites de cette approche, des structures de stockage des hypothèses d'interprétation intermédiaires ont été définies de manière centralisée, pour que les modules de traitement puissent exploiter l'ensemble des connaissances générées suite à l'application d'autres modules. La stratégie d'application successive des modules de traitement est définie de manière externalisée dans un fichier de configuration, garantissant l'adaptabilité du système mais permettant également de rompre la séquentialité de l'approche. Ainsi, sous réserve de respecter certaines contraintes de dépendance forte entre modules, il est possible d'injecter dans les structures de stockage centralisées des connaissances issues de l'application d'un module de traitement et de relancer l'application de modules de plus bas niveau afin qu'ils prennent en compte ces connaissances initialement non disponibles.[2]

Le schéma en figure 1 expose d'une façon simplifiée l'architecture globale de la plate-forme TiLT. En fonction d'une stratégie, des données linguistiques, du texte à analyser et de la configuration des modules de la plate-forme, un module nommé *stratège* appelle les différents modules pour un traitement spécifique. Les résultats de chaque module sont entreposés dans un graphe d'analyse (treillis). Afin de pouvoir prendre une décision en cas d'ambiguïtés, un contrôle multicritère (voir section 4) exploite des critères associés aux résultats concurrents.

2.2. *Multilinguisme et ressources linguistiques de la plate-forme TiLT*

L'une des principales propriétés de la plate-forme TiLT est sa conception multilingue, non seulement pour traiter des documents en différentes langues (multi-monolingue, par ex. voir section 5.2) mais aussi pour réaliser des applications interlingues (comme le CLIR[3] ou la traduction automatique). Ainsi, afin de faciliter le passage d'une langue à l'autre, chaque module de traitement a été défini de manière la plus déclarative possible, permettant ainsi une séparation rigoureuse entre les algorithmes de traitement et les connaissances linguistiques exploitées. De ce fait, TiLT est aussi indépendant que possible des langues traitées. Il est utilisable, avec des niveaux de couverture variables, pour des langues européennes (français, anglais, allemand, espagnol, portugais), pour l'arabe ou encore le chinois. Des travaux ont aussi été réa-

2. Par ex., l'application de l'analyse en dépendance dédiée à la reconnaissance d'entitées nommées permet d'ajouter des hypothèses lexicales avant le découpage en constituants (*cf.* 3.4).
3. *Cross language information retrieval*, voir aussi section 5.4

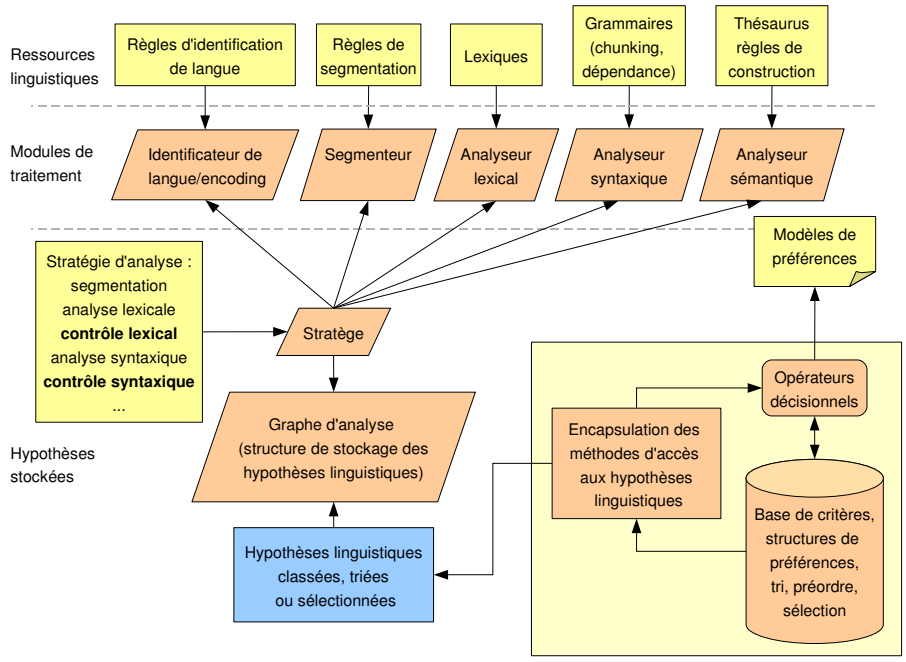

Figure 1. *Schéma de l'architecture fonctionnelle de TiLT*

lisés sur la langue des signes française (LSF, voir fin section 3.6.1). Le lien entre les langues, dans le cas des applications multilingues, est véhiculé à travers le thésaurus sémantique (*cf.* 3.6.2) qui définit les concepts utilisés dans les traitements d'ordre sémantique.

Un outil interne à l'équipe permet d'adapter les données génériques de chaque langue à une application donnée notamment en limitant ou spécialisant le vocabulaire nécessaire ainsi que les grammaires ou les données sémantiques. Les données utilisées par chaque module sont compilées de manière à en optimiser l'accès en vitesse même en très grande volumétrie (par ex. 1 000 000 entrées lexicales ou relations du thésaurus).

2.3. *Configuration et interopérabilité de la plate-forme TiLT*

Le comportement des modules est entièrement contrôlé par des fichiers de configuration. Ces derniers spécifient les données à utiliser (telles que lexiques, grammaires, thésaurus, voir section 3), des paramètres et stratégies pour le contrôle et l'analyse ainsi que l'ordre des étapes de l'analyse. Certaines données (notamment pour l'iden-

tification de langue) et fichiers de configuration associés sont créés automatiquement. TiLT peut être couplé avec des modules externes en entrée ou en sortie (par ex. la segmentation pour le chinois, (Liu *et al.*, 2006) ou la synthèse vocale).

La plate-forme TiLT est opérationnelle sous Linux et Windows (XP). Elle peut être déployée soit comme serveur HTTP, soit comme bibliothèque dynamique (C/C++, des interfaces pour Java, Python et Perl existent aussi). Pour faciliter le développement et les tests sur les données linguistiques, une interface graphique permet la visualisation de tous les modes d'applications et les résultats intermédiaires des traitements. Afin de pouvoir communiquer avec d'autres outils, TiLT permet de sortir les résultats du traitement en XML. Cette sortie peut être modifiée en spécifiant à TiLT des fichiers XSLT. En entrée, les modules de TiLT acceptent les encodages standard : soit le codage 8-bits en fonction de la langue utilisée, c'est-à-dire ISO-8859-1 ou ISO-8859-15 pour la plupart des langues occidentales, ISO-8859-2 pour les langues d'Europe de l'Est (dans notre cas le polonais), ou encore ISO-8859-6 ou CP-1256 pour l'arabe. Les lexiques chinois sont codés en GB-2312 ou en BIG-5. Dans tous les cas, TiLT permet d'analyser des textes en UTF-8 (unicode). La plate-forme elle-même accepte des textes bruts ou des documents XML (qui doivent être accompagnés par un fichier XSLT afin de pouvoir extraire les parties textuelles à traiter).

3. Modules de la plate-forme

Nous distinguons deux types de modules : ceux qui sont présents dans toutes les applications parce qu'ils fournissent une fonctionnalité fondamentale pour le TALN (comme l'analyse lexicale ou syntaxique) et ceux qui mettent en œuvre une application spécifique comme le résumé automatique ou le traitement des requêtes d'utilisateurs que nous ne décrirons pas fonctionnellement dans le cadre de cet article (voir section 5 pour leurs cas d'utilisation).

3.1. *Identification de langue*

La plate-forme TiLT étant conçue d'une façon multilingue, le premier composant appelé sur un texte est un identificateur de langue[4]. Il est basé sur trois méthodes combinables en fonction de la taille du texte (trigrammes, avec lexiques, avec patrons morphologiques). En plus de la langue identifiée, l'identificateur rend aussi le codage du texte (UTF-8, ISO-8859-1, etc.). Le choix des différentes méthodes se fait en fonction de la taille du texte à analyser : l'identification par trigramme est peu fiable sur des énoncés très courts (< 15 caractères).

4. Nous disposons actuellement des données permettant de distinguer environ 40 langues.

3.2. *Segmentation de la phrase*

Cette étape consiste à découper la phrase en segments, en fonction des données de segmentation décrites sous la forme d'expressions régulières. Chaque segment est constitué d'un type et d'une chaîne de caractères. Le type permet d'orienter les traitements postérieurs effectués sur ce segment. Un exemple de résultat de cette étape pour une phrase comme « L'année 2007 était bonne ! » est « *L'*[MOT] *année*[MOT] ␣[ESPACE] *2007*[ANNÉE] ␣[ESPACE] *était*[MOT] ␣[ESPACE] *bonne*[MOT] ␣[ESPACE] *!*[POINT] ».

3.3. *Analyse lexicale et types de correction*

Cette phase consiste à appliquer des actions à chaque segment identifié, en fonction de son type, afin de lui associer les interprétations lexicales qui lui correspondent. Ces objets, appelés « terminaux », sont stockés dans le graphe d'analyse comme le sont les segments (figure 1). Par rapport à l'exemple de section 3.2 on ne va pas rechercher dans le lexique des chaînes qui ont été typées ANNÉE ou POINT, en revanche toutes celles typées MOT feront l'objet d'un accès au lexique.

Chaque lexique monolingue comporte les informations morphologiques, phonétiques et syntaxiques des unités lexicales de la langue, ainsi que leur découpage en sens. Classiquement, chaque unité lexicale est référencée par un lemme auquel est attaché un code flexionnel correspondant à un paradigme graphique et phonétique. L'alignement entre graphie et phonétique rend possible la fonctionnalité de correction phonétique du logiciel. Des descripteurs morphologiques et syntaxiques (genre, nombre, auxiliaire de conjugaison, valence, etc.) distinguent les formes fléchies deux à deux et encodent les comportements syntaxiques des entrées lexicales.

Les entrées lexicales peuvent être de type mot simple ou locution : « pomme », « rendez-vous », « animal de compagnie », « Banque nationale de Paris » ; un mécanisme spécifique du module d'analyse permet dans ce cas de créer les interprétations correspondant à des locutions connexes si tous les éléments de ces locutions sont présents dans la phrase analysée. Les entrées peuvent être des formes contractées : « desquelles » ou « au » pour le français, « *vom* », « *del* » ou « *gonna* » pour l'allemand, l'espagnol ou l'anglais ou être des clitiques : « *'alkitāb* » ou « *sayaktubuhu* » pour l'arabe.

Si un mot est inconnu du lexique, différentes méthodes de correction peuvent lui être appliquées : correction par réaccentuation, correction phonétique, correction typographique, etc. Un mécanisme d'analyse morphologique peut être aussi appelé pour compléter les analyses d'un mot ou le corriger. Il est important de noter que l'emploi de méthodes de correction sur les mots inconnus peut avoir pour conséquence qu'à un segment unique correspondent des formes lexicales multiples. Par exemple, la correction par réaccentuation de « peche » donne « pêche » (le fruit ou l'acte de prendre du poisson), « péché » (la faute), « pèche » (une des formes conjuguées de « pécher »

l'acte de commettre une faute). Le résultat de l'analyse lexicale de « peche » sera donc l'ensemble des résultats des analyses lexicales des trois formes citées.

3.4. *Analyse syntaxique : découpage en constituants syntaxiques*

Le module de découpage en constituants syntaxiques (*chunking*) a pour rôle principal de construire une analyse syntaxique de surface (*shallow parsing*). Il s'appuie sur les résultats de l'analyse minimale. Le but du découpage en constituants est :

– de regrouper les terminaux de mêmes catégories et correspondant à un mot donné au sein d'un objet unique (GS1). À partir des GS1 on crée des syntagmes minimaux (constituants syntaxiques ou *chunks*) initiaux ;

– d'appliquer des règles hors contexte pour créer des constituants syntaxiques à partir des GS1 ; par exemple, le bloc de règles de la figure 2 (à gauche) décrit, sans exhaustivité, la composition possible d'un groupe nominal en français : « déterminant + adjectif + nom » avec optionnalité de l'adjectif. Avec ces quatre règles et la portion de texte « le petit rouge », un seul constituant syntaxique factorise les trois analyses possibles montrées à droite dans la même figure ;

– de vérifier des contraintes d'accord au sein des constituants par un mécanisme d'unification pour assurer la cohérence grammaticale du groupe ; dans l'exemple de la figure 2 la contrainte d'accord associée aux constituants de type GNN[5] vérifie l'accord en genre et en nombre de l'adjectif, du déterminant et du nom ;

– de sélectionner une suite de constituants syntaxiques respectant deux à deux des contraintes de séquentialité. La stratégie de base consiste à sélectionner les constituants syntaxiques les plus longs possibles de la gauche vers la droite. Précisons que cette stratégie peut être remise en cause par le mécanisme de contrôle (*cf.* 4.6.2).

GND + GNA → GNA	(le/GND petit_rouge/GNN)
GNA + GNN → GNN	(le/GND petit/GNA rouge/GNN)
GND + GNN → GNN	(le/GND petit/GNN rouge/GNA)
GNN + GNA → GNN	

Figure 2. *Exemple de règles de* chunking *(à gauche) et le résultat de leur application*

Le résultat de cette étape est le découpage en constituants de la phrase. Chaque constituant peut être ambigu en termes de suites de GS1 et donc de terminaux. Des mécanismes de pondération permettent de restituer une solution unique pour obtenir une désambiguïsation morphosyntaxique des segments de la phrase (*part-of-speech tagging*). Les grammaires de découpage en constituants syntaxiques des différentes

5. GNN : nom commun/groupe nominal, GNA : adjectif/groupe adjectival, GND : déterminant. Les parties à gauche de la première partie des règles ainsi que la partie droite (après la flèche) sont des constituants, les parties à droite de la première partie des règles sont des GS1.

langues ont été constituées manuellement d'après observations sur corpus. Le nombre de règles pour chacune des langues couvertes varie entre 1 000 et 2 000. L'analyseur a été évalué sur le français dans le cadre de la campagne GRACE (Adda *et al.*, 1999) avec une précision supérieure à 95 %.

3.5. *Analyse syntaxique : dépendance*

Dans cette étape, on utilise une grammaire de dépendance (Tesnière, 1959) pour construire une analyse syntaxique arborescente dans laquelle les relations fonctionnelles de la phrase sont exprimées. Ces relations syntaxiques sont construites entre les différents groupes de premier niveau (GS1) d'une phrase.

Pour y parvenir, on utilise des contraintes non locales telles que les compléments régis par une tête lexicale ou l'accord entre les groupes syntaxiques. La grammaire de dépendance se compose de règles de création de sous-arbres telles que la règle GV-5 (voir figure 3). Celle-ci permet l'attachement d'un pronom sujet à une tête verbale : SUJ est le type de relation créé. GV-PT est la catégorie de la tête (principal ; P), PRN-S celle du dépendant (D). Le symbole « >> » exprime l'ordre linéaire entre la tête et le dépendant, ici le dépendant précède la tête. Les *ConditionsPrincipales* expriment les contraintes sur la tête : « ¬IMPERS ¬SUJ_REMPLI » le verbe tête ne doit pas être impersonnel et ne pas déjà avoir un sujet. Des contraintes sur le dépendant peuvent être précisées si besoin. Les clauses « P/NOMBRE *unifier* D/NOMBRE » et « P/PERSONNE *unifier* D/PERSONNE » des *AutresConditions* s'assurent de l'accord entre le pronom (ici le dépendant) et le verbe (ici le principal) par unification des traits de nombre et de personne. Le trait « SUJ_REMPLI/+ » s'ajoute aux traits de la tête pour bloquer l'attachement de plusieurs sujets sur un même verbe.

IdentifiantUnique	GV-5
RelationSyntaxique	SUJ
Schéma	GV-PT >> PRN-s
ConditionsPrincipales	¬SUJ_REMPLI ¬IMPERS
AutresConditions	P *rajouter* SUJ_REMPLI/+
	P/NOMBRE *unifier* D/NOMBRE
	P/PERSONNE *unifier* D/PERSONNE

Figure 3. *Exemple d'une règle de dépendance*

On ne cherche pas forcément à traiter tous les phénomènes syntaxiques d'une langue donnée ; par exemple pour le français (la langue la mieux couverte), on vise à traiter :(a) la syntaxe des principaux groupes : groupes nominaux (avec noms propres et noms communs), groupes verbaux et groupe adjectivaux ; (b) les phénomènes de sous-catégorisation ; (c) les phénomènes d'alternances syntaxiques les plus saillantes ; (d) les principaux types de circonstants ; (e) les principaux types de subordonnées ; (f) les principales tournures interrogatives ; (g) les principaux cas de coordination ;

(h) la grammaire des dates et des heures. L'arbre de dépendance en figure 4 montre le résultat d'une analyse. L'analyse est montante, et se fait par îlots, ce qui permet une certaine robustesse. Dans le cas où la grammaire ne permet pas de produire un arbre syntaxique pour toute une phrase, TiLT produit un ensemble d'arbres syntaxiques, où chaque arbre représente un tronçon de la phrase.

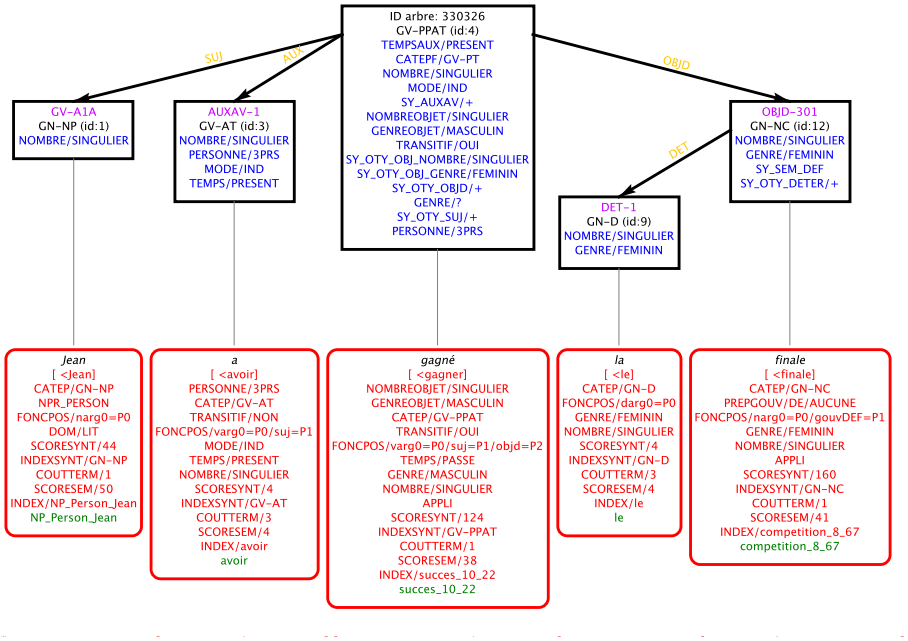

Figure 4. *Arbre de dépendance pour « Jean a gagné la finale »*

L'analyseur a été évalué dans le cadre de la campagne EASY (Paroubek *et al.*, 2007) (système P1) et a montré sa robustesse par la stabilité de ses résultats sur les différents types de corpus testés (oral, presse, Web ...).

3.6. *Analyse sémantique*

3.6.1. *Graphes sémantiques*

L'analyse sémantique construit une représentation du sens de la phrase, sous forme d'un graphe conceptuel (Sowa, 1984) le plus indépendant possible de la structure syntaxique. Ces graphes sont orientés et acycliques ; par exemple la phrase anglaise « *Jean won the final* » ou le passif « Cette finale a été gagnée par Jean » produisent le même graphe. Il est à noter qu'il ne s'agit pas d'une logique du premier ordre mais plutôt d'une structure prédicat-argument : la quantification, par exemple, n'est pas modéli-

sée, pas plus que la portée de la négation, la sémantique du discours ou la modélisation du focus et du thème.

Pour parvenir au calcul du graphe, plusieurs types de données et représentations sont mobilisés, en particulier les données de sémantique lexicale regroupées dans le thésaurus (voir section 3.6.2). Le graphe sémantique est calculé à partir de l'analyse en dépendance. Chaque nœud de l'arbre est parcouru et associé à un ou plusieurs prédicats. Pour l'énoncé « Jean gagne la finale », *REL_EVENT.gagner(agent=*x_1*, situation=*x_2*, thème=*x_3*)* est le prédicat associé à *gagner*, *player˜Jean*(x_1) celui de *Jean*, et *EVENT.finale_match* celui de *finale*. Les différents prédicats sont liés en fonction des relations syntaxiques de l'arbre de dépendance. Par exemple, la relation de dépendance SUJ entre *Jean* et *gagne* est exploitée par la règle : « SUJ : P/suj = D/narg$_0$ ». afin de lier les prédicats associés (*player˜Jean* et *REL_EVENT.gagner*). Cette règle peut s'interpréter comme « dans la relation SUJ, les variables doivent être unifiées entre l'item qui a la fonction sujet dans le principal et l'item qui a la fonction narg$_0$ du dépendant ». Le graphe sémantique de *Jean gagne la finale* est donné en figure 5.

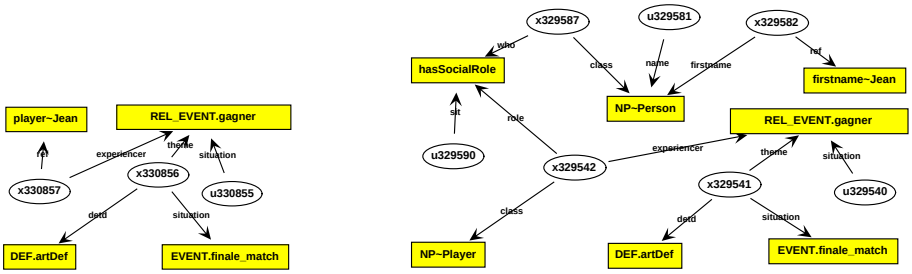

Figure 5. *Graphe sémantique pour* Jean gagne la finale *(avant et après l'application des règles de transformation)*

En fonction de l'application, des règles de transformation peuvent être utilisées afin de transformer des parties du graphe, rajouter ou supprimer des prédicats, notamment pour faire émerger les prédicats à partir des traits morphologiques (comme temps, aspect ou encore nombre). Le graphe de gauche en figure 5 est transformé en celui de droite afin de créer des représentations ontologiques (*cf.* section 3.7). Des perspectives pour améliorer cette approche sont décrites dans (Amblard *et al.*, 2008).

Tous les arbres issus de l'analyse en dépendance sont *a priori* utilisables pour la création des graphes, ce qui entraîne une explosion combinatoire. En effet, l'ambiguïté des mots et des structures morphosyntaxiques multiplie le nombre de graphes obtenus. Pour limiter l'explosion, et afin d'obtenir le graphe le plus adéquat, nous utilisons des heuristiques simples : limitation du nombre d'arbres en dépendance en fonction de leur pertinence syntaxique ; suppression des doublons de graphes, tri et sélection en fonction de leur connectivité. Bien entendu, la construction des graphes dépend de la qualité de l'analyse en dépendance et de la couverture du thésaurus. À cause de la limitation des phénomènes traités par la dépendance on ne peut pas traiter des phrases

syntaxiquement complexes. En revanche, pour les phrases de type requêtes utilisateur (*je cherche un train pour Londres lundi prochain*) dans un domaine sémantiquement limité, nous avons montré qu'il était possible de mettre au point les données linguistiques nécessaires pour la construction des graphes (Heinecke et Toumani, 2003).

La représentation sémantique associée à chaque énoncé peut faire office de pivot interlingue pour un système de traduction automatique. Un module de génération peut y être adjoint et permettre une reformulation de l'énoncé de départ dans une autre langue. Des travaux sur la traduction automatique symbolique avec pivot interlingue ont été menés dans l'équipe pour la traduction du français vers l'anglais et inversement ainsi que la traduction du français vers la langue des signes avec, dans ce dernier cas, restitution par un avatar signeur. Nous renvoyons le lecteur intéressé aux travaux de (Iheddadene, 2006) et (Kervajan *et al.*, 2007).

3.6.2. *Thésaurus multilingue*

Les informations sémantiques pour les différentes langues sont répertoriées dans un thésaurus multilingue. Il s'agit d'un catalogue qui accumule de l'information sur les différents sens des mots et sur les relations qu'ils entretiennent entre eux. L'information sémantique y est structurée afin de permettre l'exploitation de son contenu par les différents modules.

Il est structuré en hiérarchie thématique ; quatre niveaux sont représentés : (a) 26 macrodomaines (ensembles de domaines), exemple : *nature* ; (b) 175 domaines (ensembles de thèmes), exemple : *mammifères* ; (c) 880 thèmes (ensembles de synsets), exemple : *chien* ; (d) 100 000 synsets (inspirés de WordNet (Fellbaum, 1998), groupes multilingues de lexicalisations) exemple : [*caniche* | *poodle* | *Pudel*] (français, anglais, allemand).

3.6.3. *Modèle sémantique*

Un modèle sémantique explicitant les relations des synsets entre eux s'applique de manière transversale au thésaurus. Ce modèle est en partie inspiré de (Mel'čuk et Polguère, 1984)[6] ; l'idée de certaines fonctions lexicales a notamment été reprise (Park *et al.*, 2007). Les éléments centraux qui constituent le modèle sont appelés « tribus ». Ce sont des familles sémantiques qui regroupent des synsets. Chaque tribu est associée à un prédicat décrivant une distribution d'arguments (argument_1, argument_2...). Chaque argument remplit un rôle sémantique (agent, patient...). Les mots présents dans la tribu sont les lexicalisations des prédicats. Par exemple, les mots *marcher* et *marcheur* sont dans la même tribu. Les tribus bénéficient d'un autre élément permettant de structurer leur contenu : les fonctions de lexicalisation (il en existe environ 170 différentes). Ces fonctions permettent de passer d'un sens à un autre, au sein d'une même tribu. Par exemple, *vendange* et *raisin* sont reliés par la fonction $<$*est la récolte de*$>$. Toute fonction peut se composer avec les autres fonctions : exemple *dormir* + *CAUSATIVE/TERMINATIVE* $\Rightarrow$ *réveiller* ; c'est-à-dire : « causer la fin de dormir ».

6. Voir aussi (Mel'čuk, 1992b), (Mel'čuk, 1992a) et (Mel'čuk *et al.*, 1999).

Les informations concernant les tribus sont ajoutées manuellement et semi-automatiquement dans le thésaurus de manière progressive, en fonction des besoins. Un système de tribus créées automatiquement en exploitant les propriétés de morphologie dérivationelle (Pétrier, 2000) a été mis en place afin que le thésaurus soit couvert entièrement lors de son utilisation dans les projets.

3.7. *Génération de représentations ontologiques*

Afin de coupler TiLT avec des applications ou systèmes basés sur des ontologies, nous disposons d'un module qui permet de transformer une phrase, une requête utilisateur en langue naturelle ou des mots-clés en représentation ontologique (Heinecke et Toumani, 2003). Le résultat de l'analyse syntaxique (*cf.* figure 4) et de l'analyse sémantique (*cf.* le graphe à droite de la figure 5) est transformé en une représentation ontologique au format RDFS (Lassila et Swick, 1999) ou OWL (McGuinness et van Harmelen, 2004) (la figure 6 est basée sur des ontologies de domaine (Dasiopoulou *et al.*, 2007) issues du projet européen aceMedia[7]).

```
<?xml version="1.0"  encoding="ISO-8859-1"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dolcel="http://ontology.ip.rm.cnr.it/ontologies/DOLCE-Lite#"
    xmlns:midlevel="http://www.acemedia.org/ontos/midlevel#"
    xmlns:tennis="http://www.acemedia.org/ontos/tennis#">
  <dolcel:Natural-Person rdf:about="#V329587">
    <midlevel:hasFirstName>Jean</midlevel:hasFirstName>
    <midlevel:hasSocialRole rdf:resource="#V329542"/>
  </dolcel:Natural-Person>
  <tennis:Player rdf:about="#V329542"/>
  <tennis:Finale rdf:about="#V329541">
    <tennis:isWonBy rdf:resource="#V329542"/>
  </tennis:Finale>
</rdf:RDF>
```

Figure 6. *Représentation ontologique (*RDFS/XML*) pour « Jean gagne la finale »*

Cette projection nécessite l'alignement préalable des ressources linguistiques (notamment les données sémantiques) avec les classes et propriétés (concepts et relations) des ontologies cibles. Pour le projet MKBEEM (Léger *et al.*, 2000)[8], cet alignement était fait manuellement, en revanche pour le projet aceMedia, un alignement automatique a été expérimenté (Heinecke, 2006 ; Chagnoux et Heinecke, 2007).

7. *http://www.acemedia.org/*
8. *http://mkbeem.elibel.tm.fr/*

4. Système de traitement contrôlé

4.1. *Génération et propagation d'hypothèses concurrentes et erronées*

Comme pour la plupart des systèmes de TALN, notamment ceux basés sur une modélisation formelle des connaissances linguistiques, l'application de TiLT dans certains contextes se heurte au problème récurrent de la gestion d'hypothèses d'interprétations concurrentes, parmi lesquelles certaines sont erronées. En considérant l'architecture de traitements modulaires et quasiment séquentielle (voir section 2) de TiLT, la gestion de ces indéterminations était initialement déléguée aux différents modules de traitement. En effet, chaque module qui compose un processus d'analyse a pour objectif de construire de nouvelles hypothèses d'interprétation correspondant au niveau de traitement concerné, mais également d'attester de la pertinence des hypothèses intermédiaires construites par les modules précédents à l'aide de nouvelles sources de connaissances.

Cependant, le manque de complémentarité entre les différents niveaux de traitement, ainsi que les incomplétudes ou les imprécisions des ressources linguistiques exploitées entraînent fréquemment une propagation des hypothèses concurrentes et erronées, se matérialisant au final par une explosion combinatoire de l'espace des résultats générés. Afin d'augmenter la précision et donc symétriquement de réduire le bruit parmi les résultats générés, il paraît indispensable de compléter les processus d'analyse par des stratégies spécifiques de contrôle. Contrairement aux principaux travaux traitant du contrôle du processus d'analyse linguistique qui ne proposent que des stratégies dédiées à un contexte particulier d'indétermination, nous avons défini une approche globale de contrôle pouvant être appliquée lors des différentes étapes du processus d'analyse réalisé par TiLT.

4.2. *Le contrôle : un processus décisionnel basé sur la combinaison de critères de comparaison*

L'objectif des stratégies de contrôle est de faire émerger les hypothèses les plus pertinentes parmi toutes celles générées. Cet objectif se matérialise par différentes problématiques en fonction notamment des impératifs du contexte applicatif et du cas d'indétermination traité. Il s'agira soit de ranger les hypothèses selon leur pertinence, soit de sélectionner un sous-ensemble de ces hypothèses, soit d'affecter ces hypothèses dans des classes ordonnées. Pour répondre à l'une de ces problématiques, il est nécessaire de disposer d'une évaluation de la pertinence relative des différentes hypothèses générées. Le contrôle des processus d'analyse, tel que nous le définissons, repose donc sur deux étapes : l'évaluation puis l'élaboration d'une recommandation de décision (rangement, sélection ou affectation).

Cette vision décisionnelle du contrôle que nous utilisons s'intègre évidemment dans une démarche plus générique du contrôle des systèmes d'intelligence artificielle. Les travaux de (Bachimont, 1992) ont permis d'identifier les trois questions cen-

trales soulevées lors de la mise en place d'une stratégie de contrôle : (a) « Quelles connaissances de contrôle utiliser ? », (b) « Comment utiliser ces connaissances ? » et (c) « Comment transmettre l'ensemble de ces connaissances au système de contrôle ? ».

À travers une étude des différentes stratégies de contrôle existantes, nous avons constaté que l'évaluation de la pertinence des hypothèses concurrentes reposait sur l'intégration et la prise en compte de connaissances supplémentaires et initialement indisponibles ou inexploitées. Nous considérons ces connaissances supplémentaires comme des critères de comparaison apportant un jugement sur la pertinence des hypothèses. Ces critères peuvent être de différentes natures :

– empirique : usage de probabilités de patrons syntaxiques pour le contrôle des grammaires de propriétés (Blache et Rauzy, 2006) ;

– heuristique : vérification de propriétés syntaxiques telles que les attachements droits ou minimaux ;

– symbolique : usage de cadres de sous-catégorisation pour le contrôle d'un analyseur syntaxique statistique (Bourigault et Frérot, 2004).

Exploité individuellement, chaque critère apporte des informations distinctives permettant de résoudre une partie des indéterminations identifiées. Pour obtenir un jugement fiable et robuste il est donc indispensable de combiner les différents critères de comparaison disponibles. L'efficacité d'une stratégie de contrôle basée sur la combinaison de critères complémentaires a déjà été démontrée dans des contextes variés, en syntaxe avec notamment (Charniak, 2005) et en désambiguïsation du sens des mots avec (Audibert, 2007).

L'intégration de critères de comparaison matérialisant l'usage de sources de connaissances supplémentaires permet de répondre à la première question soulevée à savoir : « Quelles connaissances de contrôle utiliser ? ».

4.3. *Approche par surclassement pour combiner les critères disponibles*

Afin de déterminer la façon dont ces connaissances spécifiques de contrôle doivent être utilisées et ainsi de répondre à la question « Comment utiliser ces connaissances ? », nous avons cherché une méthodologie adaptée à notre formalisation décisionnelle du contrôle et aux impératifs de notre contexte industriel. Ceci nous a amenés à considérer d'autres méthodes que celles issues de l'apprentissage automatique, dans la mesure où leur application et leur efficacité sont entièrement conditionnées par la disponibilité de corpus d'apprentissage qui ne sont pas toujours libres. Par ailleurs, les méthodes plus classiques d'agrégation de critères, telles que les méthodes lexicographiques ou par critère unique de synthèse, souffrent de limites importantes notamment lorsque l'on cherche à combiner des critères non commensurables et apportant des jugements imprécis et incertains.

Ces constats nous ont conduits à effectuer une intersection avec un domaine spécialisé dans la résolution de problèmes décisionnels basés sur la combinaison de critères hétérogènes, l'aide multicritère à la décision (AMCD) et plus particulièrement les approches par surclassement (Roy et Bouyssou, 1993). L'AMCD se définit comme une extension pragmatique des travaux en théorie de la décision, visant à fournir un ensemble de méthodes permettant de répondre à l'une des problématiques suivantes : le rangement, la sélection ou le tri. Contrairement aux méthodes d'apprentissage automatique, les méthodes d'AMCD par surclassement ne dépendent pas de la disponibilité de corpus d'apprentissage. La façon dont les différents critères de comparaison doivent être exploités est définie par les connaissances et les intuitions d'un expert formulées *a priori*. Par « expert », nous désignons les linguistes et informaticiens en charge de paramétrer TiLT dans différents contextes applicatifs.

Ces connaissances expertes sont formalisées en tant que paramètres préférentiels constituant ainsi ce que nous nommons un modèle de préférences. Ainsi, sur chaque critère utilisé, l'expert peut associer : (a) un poids ; (b) un seuil de préférence ; (c) un seuil d'indifférence ; (d) un seuil veto. Les seuils d'indifférence et de préférence permettent de prendre en compte la nature imprécise des jugements émis sur les critères de comparaison. Le seuil veto permet de filtrer des hypothèses jugées comme trop faibles sur un des critères utilisés.

À partir de cette formalisation des connaissances expertes et des performances qu'elles obtiennent sur les différents critères, les hypothèses concurrentes sont comparées entre elles. Ces comparaisons sont matérialisées par des relations de surclassement. Une hypothèse H_1 surclasse une autre hypothèse H_2 si, d'après les connaissances dont on dispose, on peut déterminer que H_1 est au moins aussi pertinente que H_2. Une situation de surclassement est établie si une majorité suffisante de critères valide cette assertion de surclassement et si la minorité des critères qui refuse cette assertion n'est pas trop importante. Cette majorité est matérialisée par une mesure de concordance qui correspond à la somme pondérée de la propension de chaque critère à valider l'assertion de surclassement. Cette mesure est ensuite diminuée par la propension pondérée des critères qui refusent le surclassement, la discordance. Le lecteur intéressé par la méthode de construction de ces relations pourra lire (Roy et Bouyssou, 1993) ou (Smits, 2008, p. 44-68).

Une relation de surclassement correspond alors à une notion générique de comparaison des hypothèses. Les relations construites peuvent ensuite être interprétées pour ranger, sélectionner ou classer les hypothèses.

4.4. *Vers une méthodologie hybride de contrôle*

Nous avons motivé notre choix pour une approche par surclassement par le fait que cette méthode ne reposait pas sur la disponibilité d'un corpus d'apprentissage, mais sur une formalisation de connaissances expertes en tant que paramètres préférentiels.

Sous réserve de disponibilité d'un corpus représentatif du cas de contrôle concerné, nous avons cependant développé un ensemble d'heuristiques permettant de suggérer des valeurs possibles aux différents paramètres préférentiels qui composent un modèle de préférences (Smits, 2008, p. 88-100). Cette extension de l'usage classique des méthodes par surclassement vise à assister et faciliter le travail de l'expert en ce qui concerne la mise en place d'une stratégie de contrôle efficace.

Ainsi, lorsque nous disposons d'un corpus de référence, c'est-à-dire d'un ensemble d'hypothèses annotées comme valides ou non valides, nous effectuons un alignement entre les hypothèses concurrentes à comparer et les hypothèses du corpus de référence. Cet alignement nous permet de construire des tables de performances qui sont composées des hypothèses à comparer, des performances qu'elles ont obtenues sur les critères de comparaison concernés et d'une annotation en tant qu'hypothèse valide ou non valide.

À partir de ces données supervisées, nous exploitons la méthodes d'apprentissage de métriques RELIEF (Kononenko, 1994) pour évaluer et quantifier l'importance relative des différents attributs/critères et ainsi déterminer un poids pour chaque critère.

Pour estimer et suggérer des valeurs aux autres paramètres préférentiels, nous construisons pour chaque critère les courbes de répartition des hypothèses valides et non valides. Ces courbes sont interprétées par des heuristiques statistiques afin d'identifier des zones de préférence – valeur du critère au-dessus de laquelle on trouve une forte majorité d'hypothèses correctes ; d'indifférence – espace de valeur où la différence de proportion d'hypothèses correctes et incorrectes n'est pas significative ; d'incomparabilité – valeur en dessous de laquelle on ne trouve que des hypothèses incorrectes.

À travers ces premiers travaux, nous avons montré que notre approche de contrôle initialement basée sur une formalisation *a priori* de connaissances expertes pouvait être complétée par des méthodes statistiques. La façon dont les connaissances de contrôle sont utilisées repose soit sur des connaissances expertes soit sur des connaissances empiriques.

4.5. *Le module de contrôle : un élément central dans l'architecture de TiLT*

Pour répondre à la troisième question soulevée par la mise en place d'une stratégie de contrôle, à savoir « Comment transmettre l'ensemble des connaissances au système de contrôle ? », nous avons conçu, implémenté et intégré dans TiLT un module spécifique de contrôle. Ce module de contrôle a pour objectifs : (a) de faciliter l'intégration ou la déclaration de critères de comparaison et de les associer aux hypothèses à comparer ; (b) de stocker et de centraliser ces critères afin de les rendre disponibles tout au long du processus d'analyse ; (c) de permettre l'application d'étapes de contrôle des hypothèses générées ou exploitées par les modules de traitement ; (d) de centraliser et de fournir des méthodes d'accès aux recommandations de décisions émises lors des

étapes de contrôle ; (e) d'obtenir une traçabilité complète des résultats des étapes de contrôle.

Afin de garantir l'applicabilité de ce module aux différents cas d'indéterminations qui peuvent apparaître au cours de processus d'analyse, nous lui avons octroyé une place prépondérante au cœur de l'architecture initiale de traitement (voir figure 1). Cette intégration délicate au sein d'une architecture logicielle existante complexe a pu être réalisée en exploitant les propriétés du paradigme de conception orientée objets. Nous avons notamment procédé à une abstraction de la notion de module de traitement, de laquelle héritent tous les modules présentés au cours de la section 3 et de la notion d'hypothèse d'interprétation, de laquelle héritent tous les objets linguistiques participant à la construction de l'interprétation finale (segments, terminaux, ensembles de traits, constituants, arbres de dépendance, graphes sémantiques, etc.). Ainsi chaque module de traitement dispose de fonctionnalités centralisées liées à la manipulation (instanciation, modification, accès) de critères de comparaison sur les hypothèses d'interprétation qu'il manipule. Différentes informations distinctives associées aux objets linguistiques ou provenant de sources de connaissances supplémentaires peuvent alors être formalisées sous la notion commune de critère de comparaison et être associées aux hypothèses concurrentes pour qualifier leur pertinence relative.

Outre ces fonctionnalités de gestion des critères de comparaison, ce module décisionnel permet à chaque module de traitement de définir des étapes de contrôle. En référençant une configuration externalisée définissant notamment les critères à exploiter et les paramètres décisionnels à utiliser, un module de traitement peut regrouper au sein d'une structure de comparaison des hypothèses concurrentes et appliquer une opération de contrôle. L'aspect déclaratif du module de contrôle se traduit par une externalisation de l'ensemble des éléments de configuration et permet ainsi d'influencer le comportement du processus de traitement sans modifier le code des modules. Ceci contribue à la recherche d'indépendance des modules vis-à-vis de la langue et du contexte applicatif.

L'application de la méthodologie de comparaison des hypothèses concurrentes génère donc une recommandation de contrôle, qui est ensuite exploitée par le module de traitement pour déterminer l'ordre de propagation des hypothèses vers les modules de niveau supérieur ou pour filtrer certaines hypothèses jugées non pertinentes.

On constate à travers la description du fonctionnement du module de contrôle et de son intégration au sein de l'architecture de traitement (voir figure 1), que nous nous rapprochons de la notion de contrôleur présent dans les architectures distribuées de traitement tel que les tableaux noirs (Bachimont, 1992).

4.6. *Illustrations de l'application de la stratégie de contrôle*

La pertinence de notre approche décisionnelle et les fonctionnalités du module dédié à cette tâche ont été évaluées sur différents cas concrets d'indétermination et

notamment : (a) l'identification par tri des liens de coréférence entre expressions (voir 4.6.1) ; (b) la sélection d'une meilleure transcription de SMS (4.6.2).

4.6.1. *Identification des liens de coréférence*

Nous disposons d'un corpus de référence composé de 80 articles du journal *Le Monde* analysé par TiLT et validé manuellement dont les paires d'expressions coréférentes ont été annotées (Smits et Tardif, 2007). Une étude linguistique a permis l'identification de 25 critères apportant des informations distinctives intéressantes pour l'identification des liens de coréférence. Ces critères sont de nature syntaxique ou catégorielle : accords genre et nombre, similarité de fonctions syntaxiques, etc. ; de nature statistique ou contextuelle : distance entre expressions, nombre d'occurences, etc. ; ou bien encore de nature structurelle : similarité graphique des expressions, nombre de mots en commun, etc. Les différentes paires d'expressions extraites candidates, ainsi que leur évaluation sur les 25 critères identifiés constituent donc une table de performances composée à la fois d'hypothèses valides (paires d'expressions coréférentes) et non valides (paires d'expressions non coréférentes). Cette table contient 3 504 paires d'expressions valides et 24 871 paires non valides.

	Modèle expert	Modèle empirique	Modèle hybride	Arbre de décision
F-mesure	0,65	0,65	0,68	0,60

Figure 7. *Résultats obtenus lors du contrôle d'un processus d'identification des liens de coréférence*

La problématique visée par la mise en place d'une stratégie de contrôle est donc le tri des paires extraites en deux classes : celle des paires validées et celle des paires non validées comme coréférentes. Pour obtenir ce tri, nous avons demandé à un expert de mettre en place un modèle de préférences définissant la façon dont les critères doivent être exploités pour identifier les paires d'expressions coréférentes. Nous avons ensuite exploité les heuristiques d'interprétation de corpus de référence pour suggérer automatiquement un modèle de préférences. Dans un troisième temps, nous avons suggéré ces paramètres déterminés de manière empirique à l'expert afin qu'il complète ou révise ses jugements initiaux, ceci permettant d'avoir un troisième modèle de préférences qualifié de mixte. Pour mieux appréhender la qualité du contrôle par tri effectué par notre approche par surclassement, nous avons comparé les résultats avec ceux obtenus à l'aide d'une méthode classique de classification : les arbres de décision (algorithme *C*4.5). La figure 7 illustre en terme de *F-Mesure* les résultats obtenus. On constate que les résultats obtenus à l'aide d'une approche par surclassement sont meilleurs que ceux obtenus avec les arbres de décision. De plus, on remaxrque que l'expert dispose de connaissances *a priori* et d'une compréhension de la sémantique des différents paramètres décisionnels suffisantes pour mettre en place une stratégie de contrôle efficace. De même, les bons résultats obtenus à l'aide du modèle suggéré attestent de la pertinence des heuristiques d'interprétation des corpus de référence. En

bénéficiant à la fois de la capacité des méthodes empiriques à identifier les critères discriminants récurrents et des connaissances plus spécifiques de l'expert permettant notamment d'exploiter des critères peu récurrents mais fortement discriminants, les résultats obtenus à l'aide du modèle mixte sont nettement meilleurs.

4.6.2. *Contrôle empirique d'un processus symbolique de transcription de SMS*

Une instance de la plate-forme TiLT a été déployée afin de transcrire des SMS[9] en français « standard » (cf. 5.3). Le processus initial de transcription est caractérisé par l'usage de ressources linguistiques adaptées aux particularités de ce style atypique d'écriture. Après une phase d'identification des différents segments qui composent un SMS, une analyse lexicale et différentes stratégies de corrections sont appliquées sur chacun des segments, formant ainsi un treillis d'unités lexicales. L'application du module d'analyse syntaxique de surface (voir section 3.4) permet ensuite de déterminer une succession valide syntaxiquement d'unités lexicales qui formeront la transcription du SMS initial. Pour plus de détails sur ce processus, le lecteur pourra consulter (Guimier de Neef *et al.*, 2007).

Une évaluation de ce processus initial de transcription a permis de mettre en avant le caractère encourageant des résultats obtenus tout en soulignant également les faiblesses de cette approche symbolique (Guimier de Neef et Fessard, 2007). Au cours de cette évaluation, nous avons notamment constaté l'impact négatif des indéterminations et de leur propagation. En effet, on observe que, malgré une couverture lexicale complète, 25 % des 29 000 SMS utilisés lors de l'évaluation sont mal transcrits. Ainsi, parmi l'ensemble des unités lexicales concurrentes générées par le module d'analyse lexicale et les différentes méthodes correctives, l'hypothèse valide n'émerge pas malgré la validation syntaxique effectuée. Une analyse des erreurs commises par TiLT a également mis en évidence l'origine multiple de ces mauvaises décisions. Parmi les 25 % de SMS couverts lexicalement mais mal transcrits, nous avons notamment constaté que les sources d'erreurs de décision provenaient soit du découpage en constituants syntaxiques, soit de la sélection d'une distribution fonctionnelle, soit de la sélection finale des unités lexicales concurrentes validées syntaxiquement.

Afin de faire émerger les unités lexicales valides parmi toutes celles générées, nous avons mis en place une stratégie globale de contrôle du processus initial de transcription. Nous avons notamment exploité la disponibilité de corpus de SMS et de leurs transcriptions (corpus de *Louvain* (Fairon et Paumier, 2006) et un corpus réalisé par l'université de Provence) pour intégrer et combiner des critères de nature empirique lors du processus de transcription. Ainsi, pour chaque unité lexicale générée, nous lui associons les fréquences observées de sa forme fléchie et de sa forme lemmatisée. Nous calculons également un meilleur chemin de bigrammes de mots sur le treillis des unités lexicales et nous marquons à l'aide d'un critère binaire chaque unité qui appartient à ce chemin. Nous utilisons également un critère heuristique, matérialisé par un score numérique défini *a priori*, qui est associé aux unités lexicales en fonction

9. Acronyme de « *Short Message Service* ».

de leur catégorie morphosyntaxique permettant de privilégier certaines catégories par rapport à d'autres.

Différentes étapes de contrôle ont été ajoutées au processus initial de transcription, afin notamment de valider le découpage en constituants effectué vis-à-vis des critères associés aux unités lexicales. Cette évaluation s'appuie sur un tri en deux classes, valide et non valide, des constituants en cours de construction. Une fois ce découpage syntaxique effectué, nous sélectionnons une meilleure distribution fonctionnelle parmi en moyenne trois hypothèses concurrentes. Finalement, nous sélectionnons une meilleure succession d'unités lexicales qui formera la transcription parmi 2,7 unités lexicales concurrentes en moyenne par segment.

Cette stratégie de contrôle a conduit à une réduction de 20 % des unités lexicales erronées présentes dans les transcription pour les SMS couverts lexicalement. Cette première expérimentation autour du contrôle du processus de transcription de SMS a surtout permis d'ouvrir d'intéressantes perpectives, notamment l'usage de critères complémentaires tels que l'appartenance à un meilleur chemin de trigrammes de catégories morphosyntaxiques ou encore un critère de confiance de correction évaluant à la fois la fréquence de la forme initiale du segment et la distance par rapport aux formes suggérées par les méthodes correctives, ceci nous permettra d'éviter le phénomène de « surcorrection ». Cette expérimentation a mis en évidence la facilité d'utilisation et l'apport du module de contrôle et plus particulièrement la centralisation et la propagation des différents critères de comparaison, les rendant ainsi disponibles et exploitables tout au long du processus d'analyse.

5. Applications opérationnelles de la technologies TiLT

Un des aspects remarquables de la plate-forme TiLT est qu'un certain nombre d'applications qui en sont issues sont déployées à travers des services opérationnels, soit à destination du grand public, soit pour des communautés spécifiques d'utilisateurs. Bien entendu, ces déploiements exigent que la technologie soit :(a) robuste pour traiter des données hétérogènes et non préalablement formatées ; (b) optimisée en termes de performance et de tenue de la charge ; (c) multilingue et adaptable à de nouvelles langues à des coûts raisonnables ; (d) exhaustive en terme de couverture des phénomènes linguistiques concernant le service visé par l'application ; (d) paramétrable pour répondre à des adaptations spécifiques au sein d'un même service ; (e) portable sur différents systèmes et différentes plates-formes ; (f) intégrable sous divers modes d'intégration et (g) bien documentée pour l'accompagnement des intégrateurs, des installateurs et des utilisateurs.

Il est à noter que toutes les applications de la plate-forme intègrent l'identification des langues et des encodages afin de déterminer l'ensemble des données linguistiques adaptées à la langue du texte à traiter (voir section 3.1).

5.1. *Correction et interprétation des requêtes*

Les modules d'analyse de base et de découpage en constituants ont été utilisés pour des applications opérationnelles de correction orthographique et d'interprétation de requêtes. La correction orthographique du site Alapage[10] et 118 712[11] est en partie faite par TiLT ; correction et interprétation de requêtes sont utilisées pour l'analyse du champ « Qui Quoi » du site annuaire des Pages Jaunes[12].

Pour Alapage, l'ensemble des ressources lexicales a été adapté au catalogue produit. TiLT réalise la correction typographique (« *laurent vouzly* » ⇒ « *laurent voulzy* ») et phonétique (« *ouellebec* » ⇒ « *houellebecq* ») mais aussi le décollement de préfixes (« *aly mac beal* » ⇒ « *ally mcbeal* »), les abréviations et variantes orthographiques régulières (« *dictionnaire des filles* » ⇒ « *dico des filles* »). Pour le 118 712, les ressources lexicales et grammaticales générales de TiLT sont adaptées au contexte annuaire (localités, activités des professionnels). Le périmètre des corrections s'apparente à celui offert sur Alapage.

Dans le cas des Pages Jaunes, intervient également l'interprétation de la requête qui permet de rapprocher la demande de l'utilisateur avec une ou plusieurs rubriques de l'annuaire. Un traitement hors ligne (*back office*) fait avec TiLT permet de transformer les descriptifs de rubriques annuaire en concepts TiLT (issus du thésaurus) : TiLT désambiguïse le vocabulaire et l'enrichit. Ainsi le mot « avocat » sera associé au concept *PROFESSION.avocat* dans l'indexation de l'activité professionnelle des avocats, mais il sera associé au concept *PLANT.avocatier* dans l'indexation des maraîchers. Ces concepts permettent également de ramener du vocabulaire supplémentaire (« barreau », « avocassier » ; « avocatier »). Les données issues de cette analyse faite hors ligne sont utilisées en ligne pour analyser la requête utilisateur. Cette analyse se fait en plusieurs étapes : désambiguïsation, correction orthographique, rapprochement au contenu analysé hors ligne et reconnaissance des noms de professionnels (sur la base de patrons syntaxiques issus de la grammaire de dépendance). La requête « *avocat prudhome Maître Dupuis* » sera corrigée en « *avocat prud'hommes Maître Dupuis* ». La requête « *avocat prud'hommes* » permettra de ramener l'activité des avocats dans la réponse et « *Maître Dupuis* » sera isolé comme nom de professionnel.

Pour cette application, la plate-forme est très performante : sur une machine 64 bits, avec 4 Go de mémoire et une vitesse de 2,80 GHz, équipée de Linux, environ 120 requêtes sont traitées par seconde en mode serveur. En production chaque serveur traite environ 731 000 requêtes par jour.

---

10. *http://www.alapage.com/*. Il est à noter que les fonctionnalités optimales de TiLT ne sont possibles qu'avec une synchronisation régulière des données de TiLT et des bases Alapage.
11. *http://www.118712.fr/*, le site Web correspondant au numéro des renseignements d'Orange 118 712.
12. *http://www.pagesjaunes.fr/*

5.2. *Abrégeur*

Pour un texte donné (page Web, document), l'abrégeur fournit un résumé et une liste de mots-clés. Le résumé est généré en identifiant les phrases représentatives du texte source (Renouf et Collier, 1995). Contrairement aux méthodes de résumé purement statistiques, l'abrégeur TiLT s'appuie sur une méthode mixte, statistique et linguistique qui permet de discerner plus efficacement les informations pertinentes dans un texte. L'analyse linguistique permet, par exemple, de reconnaître les différentes variantes d'un mot (conjugaison des verbes, formes fléchies des noms et adjectifs, etc.), d'identifier les séquences de mots correspondant à un concept précis (mots composés, noms de personnes), de filtrer les mots faiblement informatifs (mots-outils ou supports), etc. La taille du résumé peut être choisie en nombre de phrases ou en pourcentage du texte source. L'abrégeur est multilingue et couvre, dans sa version actuelle, sept langues : français, anglais, espagnol, allemand, polonais, portugais et arabe. Des prétraitements documentaires permettent de prendre en compte un certain nombre de formats de documents (texte simple, Word, PDF, Postscript et HTML). L'abrégeur TiLT est déjà intégré dans plusieurs services ou prototypes internes à France Télécom. Il a également été intégré dans une importante plate-forme documentaire de veille.

En terme de performances, la version actuelle de l'abrégeur TiLT, implémentée sur une machine similaire à celle mentionnée précédemment, est capable de traiter, en une minute, près de 54 documents de 8 pages[13] environ (ou approximativement 19 documents de 32 pages ou encore 3 documents de 128 pages), ce qui traduit une performance pondérée de près de 500 pages par minute. Dans une version spécifique optimisée pour des flux documentaires de veille, l'abrégeur est sollicité pour traiter plusieurs milliers de documents par jour.

5.3. *Vocalisation des SMS*

Une autre instance de TiLT est utilisée dans un service opérationnel d'Orange pour la vocalisation des SMS. Ce service permet de recevoir un SMS sur un téléphone fixe quelles que soient les caractéristiques du terminal fixe récepteur. Si le téléphone ne permet pas la lecture de messages, les SMS sont vocalisés. TiLT transcrit si besoin le SMS en français standard en amont de la synthèse vocale. Le trafic actuel du service est de l'ordre de 20 000 à 30 000 SMS par jour bien que les tests de charge montrent une capacité de traitement d'au moins 240 000 SMS par jour.

5.4. *Interface avec des systèmes à base d'ontologie*

Deux prototypes[14] exploitent l'interface de TiLT qui permet aux utilisateurs de communiquer avec un système à base d'ontologie (*cf.* 3.7). Le caractère multilingue

13. Une page contient environ 2 600 à 3 100 caractères.
14. Ils ont été développés dans le cadre des projets MKBEEM et aceMedia.

de TiLT s'illustre parfaitement dans ce contexte : les expressions ontologiques, indépendantes de la langue source, sont indifféremment générées à partir des requêtes ou annotations textuelles françaises, anglaises ou espagnoles (*cf.* synsets du thésaurus 3.6.2). Par exemple, dans aceMedia, les annotations textuelles en plusieurs langues des contenus multimédias sont transformées en représentations ontologiques qui sont stockées (avec d'autres métadonnées) dans une base de connaissances. Ensuite les utilisateurs peuvent formuler leurs requêtes qui sont transformées en expressions ontologiques (SPARQL[15]), afin de les rechercher dans cette base.

6. Conclusion

Cet article présente, de manière succincte et descriptive, la plate-forme de traitement automatique des langues naturelles développée à France Télécom. À travers cette présentation, un certain nombre de caractéristiques de TiLT ont été soulignées. Ainsi, sa dimension multilingue, soutenue par l'indépendance des données linguistiques vis-à-vis des traitements, facilite grandement l'adaptation à de nouvelles langues, et même à des familles de langues différentes (langues indo-européennes, sémitiques ou sino-tibétaines ou encore langues signées). Par ailleurs, l'architecture très modulaire de la plate-forme, enrichie par un système de contrôle global, permet de décliner les traitements sur un grand nombre d'applications différentes utilisant des modules linguistiques communs ou spécifiques. Ainsi TiLT offre une grande interopérabilité tant au niveau des modules linguistiques de base qu'au niveau des modules d'application. Un point particulièrement remarquable est la capacité de TiLT à traiter des types de textes de styles très divers : textes journalistiques, requêtes utilisateurs, transcription de l'oral (Bové *et al.*, 2006), SMS ou résultats de la reconnaissance optique de caractères. Ceci permet à TiLT d'être un composant efficace dans des applications multimodales (Boualem *et al.*, 2002). Ces caractéristiques rendent TiLT particulièrement exploitable pour l'accès à l'information multimédia disponible sur l'Internet et ce à partir de terminaux fixes ou mobiles.

Remerciements

Nous remercions les nombreux collègues qui, depuis des années, ont contribué à la conception et au développement des modules et des données linguistiques de TiLT : Frédérique Arga, Olivier Collin, Arnaud Debeurme, Pascal Filoche, Michel Gilloux, Edmond Lassalle, Gilles Le Calvez, Patrick Le Dévédec, Jean-Michel Ombrouck, Frédérique Pinson, Gilles Prigent, Olivier Tardif et Jérôme Vinesse. Nous remercions aussi tous les doctorants, post-doctorants, stagiaires et intervenants externes pour leurs contributions.

15. *http://www.w3.org/TR/rdf-sparql-query/*

7. Bibliographie

Adda G., Mariani J., Paroubek P., Rajman M., Lecomte J., « Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morpho-syntaxiques pour le français », *TALN*, p. 15-24, 1999.

Amblard M., Heinecke J., Maillebuau E., « Discourse Representation Theory et graphes sémantiques. formalisation sémantique en contexte industriel », *TALN*, p. 350-359, 2008.

Audibert L., « Désambiguïsation lexicale automatique : sélection automatique d'indices », *TALN*, p. 13-23, 2007.

Bachimont B., *Le contrôle dans les systèmes à base de connaissances*, Hermes, 1992.

Blache P., Rauzy S., « Mécanismes de contrôle pour l'analyse en grammaires de propriétés », *TALN*, p. 415-424, 2006.

Boualem M., Almeida L., Amdal I., Beires N., Boves L., den Os E., Filoche P., Gomes R., Knudsen J. E., Kvale K., Rugelbak J., Tallec C., Warakagoda N., « Multimodal, multilingual information services for small mobile terminals, Eurescom MUST project », *TALN, workshop NLP techniques for speech analysis*, Nancy, p. 113-118, 2002.

Bourigault D., Frérot C., « Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène », *TALN*, 2004.

Bové R., Chardenon C., Jean V., « Impact des disfluences sur l'analyse syntaxique automatique de l'oral », *TALN*, p. 103-111, 2006.

Chagnoux M., Heinecke J., « Aligner ontologies et langues naturelles. gérer la synonymie », *Plateforme AFIA. Atelier thématique : Ontologies et Gestion de l'Hétérogénéité Sémantique*, Grenoble, p. 87-94, 2007.

Charniak E., « Coarse-to-fine n-best parsing and MaxEnt discriminative reranking », *43rd ACL*, p. 233-240, 2005.

Dasiopoulou S., Heinecke J., Saathoff C., Strintzis M. G., « Multimedia reasoning with natural language support », *IEEE-Int. Conference on Semantic Computing*, p. 413-420, 2007.

Fairon C., Paumier S., « A translated corpus of 30 000 French SMS », *LREC*, 2006.

Fellbaum C., *WordNet. An Electronic Lexical Database*, MIT Press, Cambridge, MA., 1998.

Guimier de Neef E., Boualem M., Chardenon C., Filoche P., Vinesse J., « Natural language processing software tools and linguistic data developed by France Télécom R&D », *Indo European Conference on Multilingual Technologies*, Pune, India, 2002.

Guimier de Neef E., Debeurme A., Park J., « TiLT correcteur de SMS : évaluation et bilan quantitatif », *TALN 2007*, Toulouse, p. 123-132, 2007.

Guimier de Neef E., Fessard S., « Évaluation d'un système de transcription de SMS », *Lexique et Grammaire 2007*, Bonifaccio, p. 217-224, 2007.

Heinecke J., « Génération automatique des représentations ontologiques », *TALN*, Presses universitaires de Louvain, Louvain, p. 502-511, 2006.

Heinecke J., Toumani F., « A Natural Language Mediation System for E-Commerce applications. An ontology-based approach », *Workshop Human Language Technology for the Semantic Web and Web Services. ISWC*, p. 39-50, 2003.

Iheddadene M., Traduction automatique. Étude et réalisation d'un module de génération à partir d'une représentation sémantique interlingue, PhD thesis, Université de Provence, 2006.

Kervajan L., Guimier de Neef E., Breton G., « Vers un système de traduction automatique français/langue des signes française », *T.A.L.*, 2007.

Kononenko I., « Estimating attributes : Analysis and extensions of RELIEF », *European Conference on Machine Learning*, 1994.

Lassila O., Swick R., « Resource Description framework (RDF) Model and Syntax Specification », 1999. *http://www.w3.org/TR/REC/rdfsyntax*.

Léger A., Michel G., Gitton S., Barrett P., Gómez-Pérez A., Lehtola A., Mokkila K., Rodrigez S., Sallentin J., Varvarigou T., Vinesse J., « Ontology domain modeling support for multi-lingual services in E-Commerce : MKBEEM », *ECAI. Workshop on Applications of Ontologies and Problem-Solving Methods*, 2000.

Liu W., Li H., Dong Y., He N., Luo H., Wang H., « France Telecom R&D Beijing Word Segmenter for Sighan Bakeoff 2006 », *5th SIGHAN Workshop on Chinese Language Processing*, ACL, Sidney, p. 122-125, 2006.

McGuinness D. L., van Harmelen F., « OWL Web Ontology Language Overview », 2004. *http://www.w3.org/TR/2004/REC-owl-features-20040210/*.

Mel'čuk I. A., *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III*, Presses de l'Université de Montréal, Montréal, 1992a.

Mel'čuk I. A., « Paraphrase et lexique : La théorie sens-texte et le dictionnaire explicatif et combinatoire », *in* I. A. Mel'čuk (ed.), *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III*, Presses de l'Université de Montréal, Montréal, p. 9-58, 1992b.

Mel'čuk I. A., Arbatchewsky-Jumarie N., Iordanskaja L., Mantha S., Polguère A., *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV*, Presses de l'Université de Montréal, Montréal, 1999.

Mel'čuk I. A., Polguère A., *Introduction à la lexicologie explicative et combinatoire*, Editions Duculot, 1984.

Park J., Maillebuau E., Guimier De Neef E., Vinesse J., Heinecke J., « Evaluating an Interlingual Semantic Representation », *in* K. Gerdes, T. Reuther, L. Wanner (eds), *Meaning - Text Theory 2007*, München - Wien, 2007.

Paroubek P., Vilnat A., Robba I., Ayache C., « Les résultats de la campagne EASY d'évaluation des analyseurs syntaxiques du français », *TALN*, Toulouse, p. 243-252, 2007.

Pétrier E., « Construction automatique d'un lexique dérivationnel par l'exemple », *RECITAL*, Lausanne, 2000.

Renouf A., Collier A., « A System of Automatic Textual Abridgement », *15th AI*, p. 395-407, 1995.

Roy B., Bouyssou D., *Aide Multicritère à la Décision : Méthodes et Cas*, Economica, 1993.

Smits G., Une approche par surclassement pour le contrôle d'un processus d'analyse linguistique, PhD thesis, Université de Caen, 2008.

Smits G., Tardif O., « Resolving coreference using an outranking approach », *Recent Advances in Natural Language Processing (RANLP)*, 2007.

Sowa J. F., *Conceptual Structures. Information Processing in Mind and Machine*, Addison-Wesley, Reading, MA., 1984.

Tesnière L., *Éléments de syntaxe structurale*, Klincksieck, Paris, 1959.