Un modèle pour unifier la gestion de ressources linguistiques en contexte multilingue

Frederik Cailliau

Sinequa Labs – Sinequa / LIPN – Université Paris 13
cailliau@sinequa.com / cailliau@lipn.univ-paris13.fr

Résumé

Le bon fonctionnement d'*Intuition*, plate-forme de recherche d'information, repose sur le développement et l'intégration d'un grand nombre de ressources linguistiques. Dans un souci de cohérence et de meilleure gestion, l'unification de ressources contenant des connaissances hétérogènes s'impose. Comme *Intuition* est disponible dans la plupart des langues européennes, cette unification se heurte au facteur multilingue. Pour surmonter les problèmes causés par les différences structurelles entre les langues, une nouvelle architecture linguistique a été conçue et exprimée en UML. Ce méta-modèle est le point de départ pour la nouvelle base de données qui sera le noyau d'un nouvel environnement de travail centré sur son utilisateur, l'expert linguistique. Cet environnement centralisera la gestion de toutes les ressources linguistiques d'*Intuition*.

Mots-clés : gestion de lexiques, base de données linguistique, ressources linguistiques, multilinguisme, architecture linguistique.

Abstract

The proper functioning of *Intuition*, Sinequa's platform for information retrieval, depends on the development and integration of a large number of linguistic resources. To improve their coherence and management, these resources which contain very diverse knowledge, need to be unified. As *Intuition* is available in most European languages, the multilingual factor complicates this unification. To overcome the problems caused by the structural differences among languages, a new linguistic architecture has been designed and modelled using UML. This meta-model is the starting point for a new database that will be the core component of a new workbench, centred on its user, the linguistic expert. This workbench will provide a centralised management of all the linguistic resources within *Intuition*.

Keywords: lexicon management, linguistic database, linguistic resources, multilinguism, linguistic architecture.

1. Introduction : *Intuition*, plate-forme de recherche d'information

Intuition est une plate-forme de recherche d'information développée par Sinequa[1]. Elle comporte un moteur de recherche et des interfaces de navigation qui intègrent des traitements linguistiques et/ou statistiques, comme l'étiquetage morphosyntaxique, la désambiguïsation sémantique et l'extraction d'entités nommées. Ces technologies permettent à l'utilisateur d'augmenter la pertinence des documents trouvés et d'accélérer sa recherche (Loupy et Crestan, 2004). Les ressources sur lesquelles reposent ces traitements comportent des connaissances linguistiques hétérogènes dans des formats différents. Dans un souci de cohérence et de meilleure gestion, leur unification s'impose. *Intuition* étant disponible dans la

[1] Voir http://www.sinequa.com/

plupart des langues européennes[2], il faut relever en même temps le défi du multilinguisme. Cet article exposera pourquoi et comment Sinequa va se doter d'un environnement de travail linguistique qui unifie la gestion des ressources linguistiques utilisées dans *Intuition*.

2. Des ressources linguistiques hétérogènes

Le fonctionnement linguistique d'*Intuition* repose sur la bonne intégration de trois types de ressources linguistiques : les lexiques, les règles et les corpus. Elles interviennent à différents niveaux dans les chaînes de traitement.

2.1. Lexiques

Les lexiques morphosyntaxiques sont des fichiers texte qui contiennent les mots de la langue avec leurs lemmes et leurs descriptions morphosyntaxiques. Ces dernières sont exprimées par un jeu d'étiquettes commun à toutes les langues qui s'enrichit au fur et à mesure que de nouvelles langues sont développées. Ce type de lexiques contient aussi bien des mots fréquents que des mots de langue spécialisée. Pour des raisons de gestion, le lexique peut s'étaler sur plusieurs fichiers, introduisant la possibilité qu'une même entrée lexicale soit présente plusieurs fois. Les lexiques sémantiques sont organisés différemment. Ces fichiers texte contiennent des lemmes assortis de chiffres représentant des descripteurs. Ceux-ci sont utilisés pour calculer le vecteur sémantique d'un document dans un espace vectoriel de Salton à 800 dimensions comme le décrivent Manigot et Pelletier (1997). Pour être pris en compte, ces lemmes doivent être codés dans les lexiques morphosyntaxiques. Il existe finalement d'autres lexiques, sous forme de fichiers texte, qui contiennent des abréviations, des synonymes, des variantes orthographiques et des dérivations.

Les différents lexiques partagent certaines connaissances. Cette duplication d'information est un problème récurrent qui complique la gestion des ressources. C'est l'une des principales motivations de ce projet.

2.2. Règles

Intuition repose sur un ensemble de règles, qui s'appliquent à plusieurs niveaux (orthographique, phonétique, morphologique, lexicale et syntaxique). Leur degré de formalisation dépend en grande partie de l'existence d'un éditeur dédié. Les plus formalisées sont les règles syntaxiques utilisées dans l'extraction d'information. Ce sont des transducteurs codés en XML et édités à l'aide d'une interface graphique faite maison. Le degré de formalisation des autres règles est variable. Elles sont écrites par un expert linguistique et testées après implémentation par un informaticien qui les a interprétées.

Une formalisation accrue facilite et détend la communication entre expert linguistique et programmeur. Elle fluidifie le cycle d'implémentation, de test et de correction.

2.3. Corpus

Au cours des années, Sinequa s'est constitué une collection de corpus impressionnante. La plus grande partie provient de journaux clients et sont donc monolingues, mais il existe quelques corpus multilingues alignés. Fonctionnellement trois types se distinguent : les

[2] *Intuition* intègre aussi quelques langues asiatiques. Leur intégration étant basée sur une autre architecture informatique, la gestion de ces langues ne sera pas traitée dans ce papier.

corpus non étiquetés, les corpus étiquetés automatiquement et les corpus étiquetés manuellement. Ces derniers ont été étiquetés automatiquement pour être ensuite corrigés par un expert linguistique.

Pour assurer leur gestion au quotidien, l'introduction d'un système de méta-données IMDI[3] à l'aide des outils IMDI est envisagée. Cet ensemble d'outils, disponibles sur le site du projet même, se décline principalement en un éditeur de méta-données et un navigateur performant. Les données sont exprimées dans des fichiers XML conformes à la norme IMDI et indépendants des ressources elles-mêmes. À l'aide d'arbres de navigation les corpus se laissent classifier selon plusieurs critères qui restent à définir. La création d'une base de données optimise la recherche dans les méta-données elles-mêmes et complète ainsi l'interface de consultation.

Ces outils, utilisés dans le projet Intera[4], visent tout d'abord la création d'un espace européen de mise en commun de méta-données. Si Sinequa peut envisager de rejoindre ce domaine, il est peu probable que ses corpus deviennent un jour accessibles à la communauté. Les droits d'auteur ou bien le caractère souvent confidentiel des données y font obstacle. Il n'en va pas de même pour ses lexiques qui sont déjà commercialisés et distribués par le biais d'ELDA[5].

3. Le défi du multilinguisme

Intuition couvre un grand éventail de langues qui sont structurellement très différentes. Ces dernières années, le français et l'anglais ont été rejoints par une grande partie des langues européennes. L'ajout d'une langue demande le développement des ressources décrites ci-dessus au moins jusqu'au niveau morphologique. Les remarques suivantes sur les propriétés des langues ne concernent que les langues qui ont été intégrées dans *Intuition*.

À cause de leurs caractéristiques linguistiques différentes, le développement de chaque nouvelle langue constitue un défi pour les structures et traitements existants. Certains problèmes demandent surtout un effort d'ingénierie, comme le codage en Unicode des caractères russes et grecs. D'autres sont conceptuels : quel genre d'étiqueteur faut-il pour désambiguïser grammaticalement des langues à cas dont l'ordre de mots est presque libre, comme le russe et le finnois ? La plupart des problèmes se trouvent quelque part au milieu. Pour l'intégration du finnois avec ses 14 cas et ses phénomènes d'harmonie vocalique[6] et d'alternance consonantique[7], un nouveau fléchisseur a été écrit. Il contient près de 20 000 règles et permet de gérer les lexiques simplement à partir des lemmes assortis de leur type de flexion. La gestion du finnois se fait donc en intension, tandis que les lexiques des autres langues sont gérés en extension. En portugais, on peut infixer la forme pronominalisée de l'objet direct ou indirect entre le radical du verbe et les marques de la déclinaison, avec une possible transformation du radical. C'est le cas dans *levá-lo-iam* : le pronom *o* a été inséré dans le verbe *levariam* entre le radical *levar* et la terminaison *iam*, entraînant des opérations

[3] IMDI : ISLE Meta Data Initiative, voir http://www.mpi.nl/IMDI/. Wittenburg *et al.* (2002) donnent une bonne introduction des principes de base, ainsi qu'une comparaison avec d'autres jeux de méta-données.

[4] Voir http://www.mpi.nl/INTERA/ et http://www.elda.org/rubrique22.html

[5] ELDA : Evaluations and Language Resources Distribution Agency (http://www.elda.org/)

[6] Par harmonie vocalique on entend le fait que certaines voyelles ne peuvent pas coexister à l'intérieur d'un mot simple. En finnois, a, o et u ne se combinent pas avec ä, ö ou y dans un même mot ; i et e sont des voyelles neutres.

[7] En finnois, l'alternance consonantique fait que tout radical dispose d'un degré fort et faible, son utilisation dépendant de l'ouverture ou fermeture de la syllabe qui suit. *Rampa* (nominatif singulier) est au degré fort, *ramman* (génitif singulier) au degré faible ; l'alternance est mp > mm.

d'origine phonétique sur le radical et sur le pronom. En suédois, l'article défini s'agglutine au substantif. En finnois, *dans ma maison* est un seul mot : *talossani*, qu'il faut décomposer en *talo* (maison), *-ssa* (dans) et *-ni* (ma). L'agglutination comme moyen de construire des mots composés existe en néerlandais, allemand, suédois, danois et finnois. Les règles de décomposition dans ces langues sont très similaires, mais pas au point d'être identiques dans toutes les langues. Il existe des règles de préfixation et de suffixation dans toutes les langues, mais c'est seulement en grec qu'un simple préfixe peut prendre un suffixe pour construire un nouveau mot. L'ampleur prise par l'affixation en grec est d'ailleurs bien illustrée par le grand nombre de lemmes d'affixes : il y en a plus de 1 500. Certaines propriétés demandent d'intervenir dans des niveaux aussi bas que la segmentation en mots. En néerlandais, une apostrophe peut représenter l'élision d'une ou de plusieurs lettres. Une phrase peut donc commencer par le pronom *'t* (*het* > *'t*), laissant au mot suivant le privilège de la capitalisation du début de phrase. Dans les sigles finnois, les deux points servent aussi de séparateur morphologique entre le radical et la terminaison.

Les structures initiales qui étaient conçues pour le traitement optimal du français et de l'anglais ne sont pas très adaptées pour coder les ressources linguistiques des autres langues. Chaque développement d'une nouvelle langue a apporté sa portion d'ajustements, et le modèle de langue initial a souffert à chaque nouvelle intégration. La refonte de l'architecture linguistique et des ressources d'*Intuition* repose sur l'hypothèse qu'il est possible d'élaborer un modèle unique et cohérent pour toutes ces langues.

4. L'architecture linguistique d'*Intuition*

Implicitement tous les lexiques reposent sur un modèle de données, mais seuls quelques-uns ont vraiment fait l'objet d'une conception préalable. C'était le cas pour certains projets comme Genelex et Multilex[8] qui ont conçu un modèle de données lexical avant de développer les lexiques. La création d'un modèle générique et indépendant d'une langue en particulier faisait partie des recherches menées.

Dans son travail sur l'architecture du lexique, Eagles (1996) considère que ce genre de modèle, appelé « architecture linguistique », « définit les objets de base du modèle et leurs relations »[9]. Les directives de Eagles ont été prises comme points de départ par la plupart des projets de lexiques depuis. En ce moment, une norme qui spécifie une telle architecture linguistique est en préparation à l'ISO dans le comité TC34 SC4[10]. Sinequa suit ces évolutions de près comme membre du groupe de consultation industrielle du consortium *Lirics*[11] qui prépare cette norme, ainsi qu'en participant à *Normalangue-RNIL*, le groupe miroir français.

Parallèlement à ces activités de normalisation et avec une approche similaire, nous avons développé une architecture linguistique qui contient tous les objets linguistiques qui existent dans les ressources linguistiques d'*Intuition*. Ce méta-modèle formalise les relations entre les objets dans un diagramme de classes en UML. Ce langage a été choisi pour sa lisibilité visuelle et sa courbe d'apprentissage rapide.

[8] Sérasset (1993) fait une comparaison entre les deux modèles. Wittenburg (2001), à la recherche d'un modèle de lexique abstrait (*Abstract Lexicon Model*), présente plusieurs modèles.

[9] http://www.ilc.cnr.it/EAGLES96/lexarch/node5.html

[10] Voir http://tc37sc4.org/ : Language Resources Management.

[11] Voir http://lirics.loria.fr/ pour *Lirics* et http://www.technolangue.net/article82.html pour *Normalangue-RNIL*.

La future base de données qui centralisera tous les lexiques doit respecter les contraintes exprimées dans ce méta-modèle. Toute duplication de connaissances sera évitée, menant à une solution qui est conceptuellement satisfaisante. Elle intégrera toutes les ressources linguistiques existantes et pourra en intégrer d'autres sans que la base du modèle soit remise en cause. Le pari est fait que ces nouvelles structures de données seront plus robustes quant à l'intégration de nouvelles langues.

Le diagramme en UML facilite la collaboration entre l'expert linguistique et l'ingénieur. Le linguiste comprend mieux les exigences de formalisation d'un ingénieur, et l'ingénieur dispose d'un schéma de référence qui lui donne des définitions formelles des objets linguistiques qu'il manipule constamment.

La figure 1 montre la partie du diagramme de classes qui formalise les objets morphosyntaxiques dans des boîtes rectangulaires et exprime les relations entre eux par des lignes. La plupart des objets sont liés par une flèche de généralisation à l'objet *Mot*. Le diagramme contient quelques exemples d'agrégations, comme la relation entre *Expression Multimot Continue* et *Mot*. Un autre exemple qui illustre bien cette relation est l'objet *Mot Derivé*, qui a besoin au minimum de deux autres objets, *Mot* et *Morphème* pour exister. Les quelques règles qui sont présentes montrent les relations de dépendance, comme par exemple entre *Mot Composé* et *Règle de Composition*. Une autre relation présente dans cette partie du diagramme est l'association, qu'on trouve entre les objets d'abréviation et ceux qui correspondent à leurs formes développées. L'objet *Mot Composé* représente les mots composés qui sont formés par agglutination, contrairement à *Expression Multimot Continue*, dont les mots composés contiennent des espaces. En dehors de ce schéma, ces deux derniers objets sont appelés communément des mots composés.

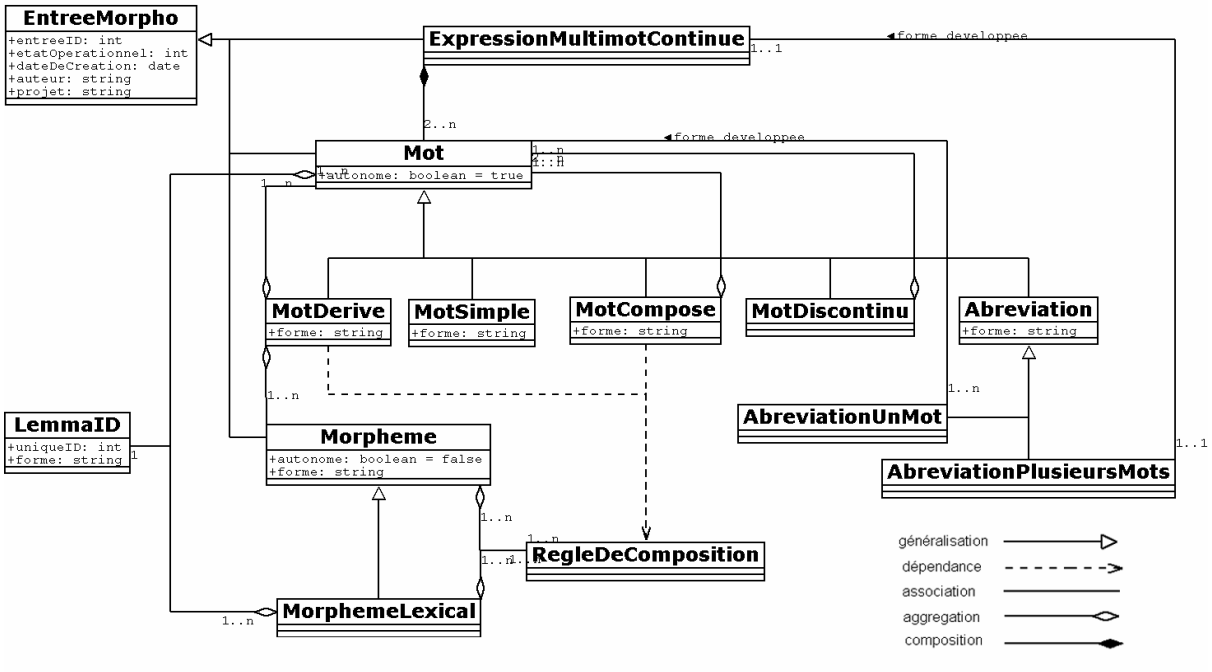

Figure 1. L'objet Mot et les objets liés

Les définitions de ces objets correspondent rarement exactement à celles qui sont usuelles dans le monde linguistique. Une expression multi-mot a été définie dans le document de travail de l'ISO[12] comme étant continue ou discontinue. Dans ce modèle, cette notion est absente car elle ne fait pas partie de l'architecture linguistique d'Intuition. Il existe cependant un autre objet plus précis qui représente l'expression continue. Dans le même document ISO, un morphème est également défini comme la plus petite sous-partie d'une forme porteuse de sens. *Intuition* n'utilise pas ce genre de connaissances. L'objet qui porte ce nom dans le diagramme est défini par un critère opérationnel : c'est un élément sans autonomie qui ne peut donc pas apparaître sans s'adosser morphologiquement à un autre élément. Ce sont donc entre autres les préfixes, suffixes et terminaisons de verbe qui sont visés par cet objet.

Ces problèmes terminologiques ne devraient pas poser de souci tant que le modèle reste intérieur à Sinequa et que les objets sont bien définis. En même temps, cela montre que le bon choix terminologique reste difficile quand on travaille en linguistique. En l'absence d'une terminologie uniforme qui est généralement acceptée, tous les termes doivent être redéfinis. Le registre des catégories de données tel qu'il est prévu dans les travaux ISO devrait mettre un peu d'ordre dans la terminologie linguistique.

L'objet *Mot Discontinu* correspond au phénomène des verbes à particule séparable, qu'on retrouve entre autres en néerlandais, en allemand et en suédois. C'est le seul objet qui a comme généralisation *Mot* et qui n'a pas de propre forme, car il est constitué d'au moins deux autres mots déjà existants.

Le diagramme de la figure 1 indique que le *Mot Composé* est fait d'au moins deux *Mots* sous respect d'une *Règle de Composition* qui dépend de la langue. Cette définition est moins triviale qu'elle ne le semble. Pour les mots composés, le schéma impose que la future base de données comporte des sortes de pointeurs qui réfèrent directement aux mots composants. Dans les structures existantes, les composants sont calculés à partir d'une étiquette descriptive. L'étiquette NN marque une possible décomposition en deux noms, comme pour *Aufsichtsrat*, allemand pour *conseil d'administration*. En appliquant les règles de décomposition, on calcule les composants *Aufsicht* et *Rat*. Le futur système de pointeurs évitera tout calcul et devrait permette de gagner en temps de traitement, ce qui est très important pour une application industrielle. Ceci permet aussi d'éliminer toute ambiguïté de décomposition qui peut exister si l'étiquette permet théoriquement plusieurs décompositions. C'est le cas par exemple pour le mot néerlandais *spiegelei* (œuf au plat) pourvu de l'étiquette NN, dont la bonne décomposition est *spiegel* (miroir) + *ei* (œuf), la mauvaise serait *spie* (clavette) + *gelei* (gelée).

5. Une gestion unifiée, centrée sur l'utilisateur

Le grand avantage d'avoir des lexiques sous la simple forme de fichiers texte est qu'ils sont extrêmement faciles à manipuler. Un puissant éditeur de texte suffit pour les consulter et les modifier occasionnellement. Les traitements par batch et scripts peuvent accomplir les opérations plus intensives. À l'inverse, la centralisation des lexiques dans une base de données complique leur édition. Des interfaces spéciales doivent être conçues et des protocoles pour l'importation et exportation de données doivent être définis. Plusieurs formats d'exportation existeront : des fichiers prêts à compiler pour *Intuition*, et des fichiers texte pour distribution par ELDA. La structure de ces derniers est propriétaire, comme c'est le cas

[12] Page 13 du document de travail n° ISO/TC37/SC4 N130 Rev.7.

de la plupart des lexiques commercialisées. Les travaux faits dans le cadre ISO devraient aussi aboutir à la définition d'un format d'échange standardisé. Cela faciliterait énormément la fusion de lexiques, même si cela ne règle pas les différences conceptuelles.

Ces interfaces feront partie d'un environnement de travail qui unifie la gestion de toutes les ressources linguistiques décrites précédemment. L'utilisateur, c'est-à-dire l'expert linguistique, est au centre de toutes les préoccupations. Quatre tâches principales ont été définies et décrites dans un diagramme de cas d'utilisation en UML : modification de lexique, modification de règles, entraînement d'étiqueteur par apprentissage supervisé et étiquetage de corpus brut. Les obligations industrielles font qu'une bonne partie de cette modélisation est dédiée à la suite de tests de régression et à la vérification des résultats. Ces descriptions procédurales assorties de quelques scénarios pris du quotidien serviront de base pour les spécifications de l'environnement de travail.

D'autres outils viendront compléter l'environnement : par exemple, pour la construction semi-automatique de lexiques ou de règles de flexion à partir de corpus brut. L'extraction de suites syntaxiques communes et moins communes à l'aide d'un concordancier aiderait beaucoup pour l'écriture de règles de désambiguïsation ou d'extraction d'entités nommées. L'idée à l'origine de ces outils est que les experts linguistiques valident ou corrigent des propositions faites par l'ordinateur plutôt que de les écrire laborieusement à partir de zéro. Ce n'est pas tant l'optimisation des travaux linguistiques qui est en jeu que le confort de travail de l'expert linguistique, pour qui l'application informatique est souvent une cause de frustration.

6. Conclusion

Cette année, la base de données linguistiques et ses interfaces de gestion vont voir le jour. Elles seront implémentées en respectant l'architecture linguistique proposée dans le méta-modèle exprimé en UML. La solidité de ces éléments centraux est essentielle, même si l'intégration de langues structurellement différentes nécessite une certaine souplesse. Une liste de souhaits et de problèmes associés aux structures de données actuelles sera utilisée pour évaluer l'apport de ce nouvel environnement de travail qui centralise la gestion de toutes les ressources linguistiques.

Références

EAGLES (1996). « Task Group on Lexicon architecture ». In *Draft Report*. http://www.ilc.cnr.it/EAGLES96/lexarch/lexarch.html

LOUPY C. DE, CRESTAN E. (2004). « Browsing Help for Faster Document Retrieval ». In *Actes de Coling*.

MANIGOT L., PELLETIER B. (1997). « Intuition, une approche mathématique et sémantique du traitement d'informations textuelles ». In *Actes de Fractal'1997* : 287-291.

SÉRASSET G. (1993). *Recent Trends of Electronic Dictionary Research and Development in Europe*. Technical Memorandum Electronic Dictionary Research (EDR). Tokyo.

WITTENBURG P. (2001). « Lexical Structures ». In *MPI Technical Report.*

WITTENBURG P., PETERS W., BROEDER D. (2002). « Metadata Proposals for Corpora and Lexica ». In *Actes de LREC*.