Construction collaborative d'une base lexicale multilingue

Le projet Papillon

Mathieu Mangeot-Lerebours[*] — Gilles Sérasset[**] — Mathieu Lafourcade[***]

[*] *National Institute of Informatics*
Hitotsubashi 2-1-2-1913 Chiyoda-ku Tokyo 101-8430 Japan
mangeot@nii.ac.jp

[**] *GETA-CLIPS, IMAG, Université Joseph Fourier*
BP 53, 38041 Grenoble cedex 9
Gilles.Serasset@imag.fr

[***] *TAL-LIRMM, Université de Montpellier II*
161, rue Ada, 34392 Montpellier cedex 5
lafourcade@lirmm.fr

RÉSUMÉ. *Nous présentons le projet Papillon dédié la construction d'une base lexicale multilingue linguistiquement riche. Ce projet s'appuie sur le principe de construction collaborative, qui permet à chacun, professionel ou amateur, institution ou individu, de contribuer, dans la mesure de ses moyens, à ce grand chantier. Pour qu'un tel travail collaboratif puisse s'amorcer, il est nécessaire de fournir un ensemble conséquent d'informations lexicales multilingues, sur lesquels les contributeurs pourront s'appuyer. Après avoir présenté l'architectures linguistique, lexicale et informatique du projet Papillon, nous détaillons la méthode utilisée pour créer les informations initiales mises à disposition des contributeurs.*

ABSTRACT. *This paper presents the Papillon project dedicated to the building of a linguistically rich multilingual lexical database. This project is based on collaborative construction principle, which allows each one, professional or amateur, institution or individual, to contribute, with its own means, to this building task. For such a collaboratif work to be effective, it is necessary to provide a important set of multilingual lexical information, that will be the base of the contributors' work. After a presentation of the linguistic, lexical and software architectures of the Papillon project, we detail the method used to create the initial lexical information.*

MOTS-CLÉS : *Base lexicale multilingue, Dictionnaire, travail collaboratif.*

KEYWORDS: *Multilingual lexical database, dictionary, collaborative work.*

2e soumission à *Traitement Automatique des Langues*, le 30 mars 2003.

1. Introduction

Qu'elle soit implicite ou explicite, la connaissance linguistique reste un constituant fondamental des systèmes de traitement des langues. Le coût généralement constaté de création d'une connaissance lexicale explicite (un dictionnaire) est l'un des freins majeurs dans le développement d'un système de traitement des langues (TAL).

De la même manière, malgré le nombre et la diversité des dictionnaires à usage humain, il reste de nombreux trous à combler. Ainsi, un francophone ne peut actuellement trouver de dictionnaire bilingue français-japonais lui donnant une transcription utilisable des traductions en kanji (idéogrammes japonais) et lui fournissant des informations qui lui sont nécessaires (les spécificateurs numériques du japonais par exemple). Ces besoins sont encore plus flagrants pour des locuteurs de langues moins représentées au niveau lexical.

Dans cet article, nous présentons tout d'abord les motivations du projet Papillon dont l'objectif est de combler ce manque en construisant une base lexicale fortement multilingue offrant des informations linguistiquement riches. Les coûts de construction d'une telle base sont réduits par l'adoption d'une stratégie (présentée en 2.2) basée sur le modèle « open source » où les données disponibles se voient constamment enrichies par des contributions d'utilisateurs aux compétences diverses. Enfin, les coûts restants sont rendus acceptables par l'adoption d'une structure linguistique et lexicale (détaillées en 2.3) favorisant la réutilisation des données construites.

Nous décrivons ensuite l'implémentation du serveur de communauté au travers duquel se fait le travail de construction de cette base. Après avoir donné une vue d'ensemble du serveur Papillon (en 3.1), et présenté les principes de représentation des différentes structures de données manipulées (en 3.2), nous détaillons les méthodes utilisées pour offrir un service d'accès unifié aux diverses données disponibles sur le site (en 3.3).

La stratégie adoptée implique un travail initial de construction d'une amorce de base lexicale contenant un ensemble d'entrées initiales non détaillées, qui servira de base aux contributions des utilisateurs. L'architecture interlingue de la base rend cette construction relativement difficile. Nous présentons donc les outils (en 4.1) et méthodes (en 4.2) mises en œuvre pour cette étape d'amorçage.

2. Le projet Papillon

2.1. *Motivations du projet*

Le projet Papillon a été initié suite à différents constats :

– Il n'existe pas à l'heure actuelle de dictionnaire français-japonais électroniques et gratuits. De plus, les dictionnaires existants sont en général conçus pour les Japonais. La transcription des kanjis (idéogrammes japonais) est, dans la plupart des cas, omise. Les francophones ne peuvent donc pas se servir de ces dictionnaires à moins de

savoir lire le japonais. De plus, d'autres informations nécessaires pour s'exprimer en japonais font aussi défaut. Il existe par exemple, une grande variété de spécificateurs numériques en japonais. Certains échappent à toute logique. Il est donc indispensable que cette information soit accessible.

– Pour un francophone, il est beaucoup plus difficile d'obtenir des informations lexicales sur le malais ou le thaï que sur l'anglais.

Les besoins en données lexicales restent donc importants, non seulement pour un utilisateur humain, mais aussi pour les systèmes de traitements des langues, non seulement pour un francophone, mais pour tout utilisateur humain quelle que soit sa langue.

La principale difficulté réside dans les coûts prohibitifs de construction de grandes quantités de données. Par exemple, le projet Electronic Dictionary Research (EDR) de construction d'un dictionnaire japonais-anglais a nécessité plus de 1200 hommes années de travail. Son prix de vente, 14 000 €environ, est très inférieur aux coûts réels de construction qui ne seront probablement jamais rentabilisés. Il est cependant encore trop élevé pour un particulier. De ce fait, seules des institutions peuvent l'acquérir. De plus, les données fournies à ce prix sont utilisables principalement par certains systèmes de traduction automatique fondés sur des techniques particulières.

Le projet Papillon met en œuvre plusieurs stratégies pour réduire ces coûts et les rendre acceptables :

– En utilisant une structure lexicale suffisamment générale et complète pour que la plupart des applications du TAL y trouvent (de manière directe ou indirecte) les données dont elles ont besoin.

– En offrant des outils simples permettant à de non-spécialistes de partager leur connaissance naturelle de leur langue maternelle. La compétence des spécialistes étant utilisée afin de nettoyer et valider les informations ainsi obtenues.

– En construisant une base multilingue fondée sur une approche interlingue par acceptions, qui permet, en factorisant l'ensemble des connaissances bilingues disponibles, de s'appuyer sur les langues bien dotées pour avancer sur les langues moins représentées.

– Enfin, en appliquant le paradigme de construction « open-source » à la construction de données lexicales : chaque utilisateur contribue bénévolement à la base lexicale et les ressources sont ensuite disponibles gratuitement pour tous.

L'utilisation du paradigme « open-source » a déjà été utilisée dans des projets similaires de construction collaboratives de données lexicales sur le Web, parfois depuis plusieurs années. Le projet EDICT de construction de dictionnaire japonais-anglais dirigé par Jim Breen, professeur à l'université Monash en Australie, a démarré, il y a plus de 10 ans. De plus, des projets parallèles d'adaptation de ce dictionnaire à d'autres langues, comme le français conduit par Jean-Marc Desperrier ([DES 02]), ont démarré avec succès. D'autres projets de construction bilingue de dictionnaires incluant le japonais ont été lancés plus récemment comme SAIKAM, japonais-thaï et WaDoKuJiten, allemand-japonais.

C'est l'utilisation conjointe de l'ensemble des stratégies énoncées qui est novatrice. Nous pensons en effet que chacune des trois premières stratégies renforce l'impact du paradigme « open-source ». La première, en couvrant de nombreux besoins, permet d'impliquer des spécialistes du TAL qui apporteront leur pierre à l'édifice. La seconde, en permettant à des utilisateurs non-spécialistes de s'impliquer dans le projet, élargie le nombre de contributeurs potentiels. La troisième, en proposant une approche multilingue dès le début du projet, nous permet d'impliquer des partenaires de nombreux pays.

Lancé en 2000 par Emmanuel Planas, François Brown de Colstoun et Mutsuko Tomokiyo, le projet Papillon a été lancé en partenariat avec le National Institute of Informatics à Tokyo (Frédéric Andrès). Après trois séminaires (dont 2 à Tokyo et 1 à Grenoble), de nombreux partenaires se sont manifestés et ont souhaité rejoindre le projet : Jim Breen, auteur du dictionnaire EDICT (Université Monash, Australie), Francis Bond (NTT, Keihanna), Yves Lepage (ATR, Keihanna), Ulrich Appel, auteur du dictionnaire allemand-japonais WaDoKuJiten, Jean-Marc Desperrier, responsable de l'adaptation au français du dictionnaire EDICT, l'université Kasetsart et le NECTEC (Bangkok, Thaïlande), l'Universiti Sains Malaysia (Penang, Malaisie), les universités de Da Nang et de Hanoi (Vietnam), etc. Actuellement, les langues couvertes sont l'allemand, l'anglais, le français, le japonais, le lao, le malais, le thaï, le vietnamien et, très récemment, le chinois. Des contacts sont en cours concernant les langues indiennes.

2.2. *Stratégie de construction de la base lexicale multilingue*

Le succès du projet Papillon dépend de sa capacité à intégrer des informations fragmentaires de toutes natures dans un modèle unique. Ces informations peuvent provenir de dictionnaires existants ou d'utilisateurs contributeurs. Dans le premier cas, il s'agit d'informations cohérentes, disponibles dans un modèle propre, duquel nous extrayons les informations que nous souhaitons représenter dans le modèle de la base lexicale multilingue Papillon. Dans le second cas, il s'agit d'informations parcellaires exprimées dans le modèle de la base Papillon, sous forme de modification de données existantes.

Les contributions ne peuvent donc exister que s'il existe un ensemble minimal d'informations lexicales sur lesquelles les contributeurs apporterons des modifications (ajout, correction ou suppression). Il est donc primordial d'adopter une stratégie « en largeur » qui commence par une étape d'amorçage dont le but est d'obtenir automatiquement, à partir de dictionnaires existants, une première base lexicale contenant de nombreuses entrées associées à des informations minimales. Cette étape d'amorçage est complexe du fait de l'utilisation d'une architecture lexicale interlingue. Nous la détaillons dans la partie 4.

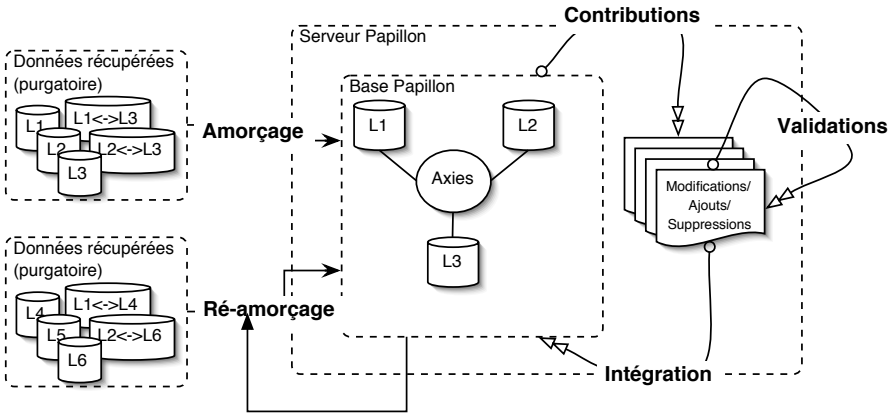

Figure 1. *Stratégie de construction de la base Papillon*

Une fois ce premier travail effectué, la base obtenue est installée sur le serveur Papillon. Les contributions sont effectuées sur cette base qui entre dans une phase de constante évolution.

Nous distinguons trois tâches dans l'évolution de cette base :

– **La contribution** où chaque utilisateur, spécialiste ou non, peut proposer des modifications (ajout d'informations dans des lexies existantes, ajout de lexies, suppression de lexies). Cette tâche pourra également être effectuée par des agents automatiques (acquisition de données à partir de corpus, etc.).

– **La validation** où les contributions seront soumises, directement ou indirectement à l'accord des utilisateurs, spécialistes ou non (des outils spécifiques seront développés afin de recueillir ces validations auprès d'utilisatrices non-spécialistes). Cette tâche pourra également être effectuée par des agents automatiques (confrontation à des corpus, ou à des ressources existantes, etc.).

– **L'intégration** où des utilisateurs de confiance acceptent ou rejètent les contributions (validées ou non) pour qu'elles soient effectivement appliquées à la base.

Cette méthode permet un certain contrôle sur la qualité de la base, mais pose un problème aux contributeurs qui ne souhaitent pas de délai entre le moment où ils contribuent et le moment où la contribution est prise en compte. Pour éviter ce délai, les modifications sont stockées dans l'espace personnel du contributeur et sont appliquées automatiquement à chaque fois qu'il consulte les entrées concernées. Ainsi les contributions sont instantanément visibles par le contributeur qui pourra aussi les partager avec d'autres utilisateurs.

2.3. *Architectures linguistique et lexicale*

2.3.1. *Macrostructure de la base lexicale*

La macrostructure de la base lexicale Papillon a été définie dans [SÉR 94b] et expérimentée à petite échelle pour la construction d'une petite base lexicale multilingue par [BLA 95].

Cette macrostructure est fondée sur la notion d'acceptions. Chaque dictionnaire monolingue est vu comme un ensemble d'acceptions d'une langue. Les liens entre les langues sont établis grâce à un dictionnaire pivot qui contient un ensemble d'acceptions interlingues (que nous appelons axies).

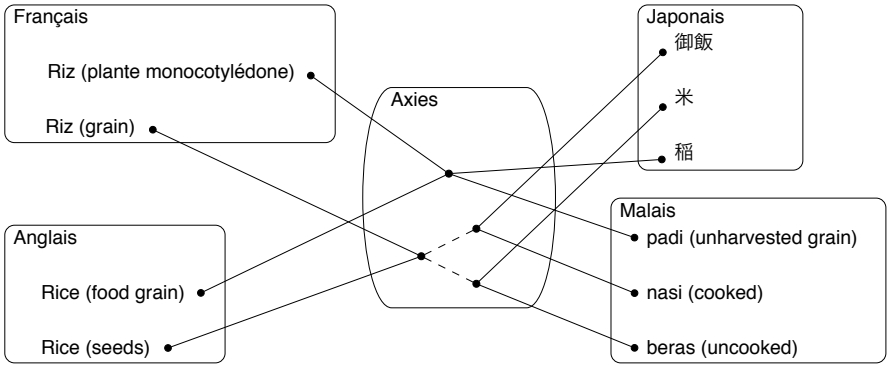

Figure 2. *Utilisation d'un lien inter axies pour représenter les phénomènes contrastifs de l'équivalence lexicale.*

Les vrais problèmes contrastifs de l'équivalence lexicale (qu'il ne faut pas confondre avec la polysémie monolingue, l'homonymie ou la synonymie comme [MEL 01] l'explique clairement) sont représentés grâce à un lien entre axies. Ce phénomène se retrouve dans l'exemple de la traduction du mot « Riz » dans 4 langues. Pour cet exemple, nous avons utilisé les acceptions définies dans des dictionnaires monolingues du commerce.

Chaque langue définit une acception correspondant au sens français désignant le riz comme une plante céréalière. Il est donc aisé de relier chacune de ces acceptions monolingues à une axie unique.

Par contre, ni le français, ni l'anglais ne définissent d'acceptions différentes suivant que le « riz » (considéré comme aliment) est cuit ou non alors que le japonais et le malais font cette distinction. Une axie ne peut être reliée à la fois aux acceptions malaises de « nasi » et « berak », à moins qu'on ne souhaite les considérer comme des synonymes (ce qui serait une erreur ici). Il faut donc créer 3 axies différentes pour relier ces acceptions monolingues : la première correspond aux acceptions de « nasi » et de « 御飯 » (gohan) ; la seconde correspond aux acceptions de « beras » et de

« 米 » (kome) et la dernière correspond aux acceptions de « riz » et de « rice ». Les traductions sont ensuite établies par l'ajout de deux liens entre la dernière axie et les deux premières.

Il faut noter que ce lien inter-axie ne représente que le fait que les deux acceptions reliées peuvent être utilisées comme traduction l'une de l'autre. Il ne porte pas de sémantique particulière et ne doit pas être confondu avec un lien ontologique.

2.3.2. *Microstructure des articles*

La structure de chacune des unités du lexique est issue de la théorie sens-texte ([MEL 84, MEL 88, MEL 92, MEL 96]). En effet, cette théorie étant indépendante des langues, elle nous permet de manipuler une structure unique pour toutes les langues de la base (à certaines nuances prêt, exposées plus bas).

Dans la théorie sens-texte, et dans le dictionnaire explicatif et combinatoire (DEC) qui en est sa composante lexicale, les articles sont nommés des lexies. Nous reprenons ici la définition d'une lexie de [POL 02] :

> Une lexie, aussi appelée unité lexicale, est un regroupement 1) de mots-formes ou 2) de constructions linguistiques qui ne se distinguent que par la flexion.
>
> Dans le premier cas, il s'agit de lexèmes et dans le second cas, de locutions.
>
> Chaque lexie (lexème ou locution) est associée à un sens donné. Que l'on retrouve dans le signifié de chacun des signes (mots-formes ou constructions linguistiques) auxquels elle correspond.

Les lexies sont ensuite regroupées en vocables. Nous reprenons ici la définition d'un vocable de [POL 02] :

> Un vocable est un regroupement de lexies qui sont associées aux mêmes signifiants et qui ont un lien sémantique évident.

Dans les dictionnaires monolingues de la base Papillon, nous reprenons la notion de lexie, qui constitue l'unité du lexique. Ces lexies correspondent à la notion d'acceptions monolingues évoquée dans le paragraphe précédent. Par contre, la notion de vocable n'est pas explicitement exprimée dans la base lexicale Papillon, cette notion se retrouvera au niveau de l'interface entre la base et ses utilisateurs.

Le DEC est actuellement constitué de 4 volumes regroupant 558 vocables en tout. C'est un dictionnaire expérimental avec une structure assez complexe et qui ne peut (encore) servir à un usage général. C'est pourquoi un projet de simplification du DEC (le projet DiCo, [POL 00]) a été lancé récemment par Alain Polguère et Igor Mel'čuk avec l'aide des étudiants de l'Observatoire de Linguistique Sens-Texte de l'université de Montréal au Canada.

Néanmoins, nous avons souhaité formaliser de manière plus détaillée certaines parties présentes implicitement dans la structure Dico.

Ainsi, dans l'exemple de la figure 3, nous représentons explicitement le fait que la fonction « Qsyn » (quasi synonymes) a trois valeurs, réparties dans deux groupes distincts : « assassinat » et « homicide » d'une part, « crime » d'autre part. Cette

Nom de l'unité lexicale	Meurtre.1
Propriétés grammaticales	nom, masc
Formule sémantique	action de tuer : PAR L'individu X DE L'individu Y
Régime	X = I = de N, A-poss Y = II = de N, A-poss
Fonctions lexicales	QSyn assassinat, homicide#1 ; crime V0 tuer A0 meurtrier-adj S1 auteur [de ART]//meurtrier-n /* Nom pour X*/
Exemples	La mésentente pourrait être le mobile du meurtre.
Idiomes	_appel au meurtre_, _crier au meurtre_

Figure 3. *Exemple d'une entrée du DiCo.*

répartition permet de noter qu'il existe une plus grande distance sémantique entre « meurtre » et « crime » qu'entre « meurtre » et « homicide ». Enfin, nous explicitons le lien qui est établi entre le sens 1 de « homicide » et cette lexie.

De plus, nous avons adopté un mécanisme qui nous permet d'adapter légèrement cette structure aux particularités des différentes langues de la base. Ainsi, les valeurs possibles comme propriétés grammaticales du thaï et du français sont différentes. Enfin, nous avons rajouté au dictionnaire du japonais les notions particulières de niveau de politesse, niveau d'usage et niveau de référence. Cette structure lexicale, exprimée en XML, est détaillée § 3.2.2.

3. Le serveur contributif Papillon

3.1. *Vue d'ensemble*

La construction, la gestion, la maintenance et une partie de l'exploitation de la base lexicale multilingue Papillon se fait par l'intermédiaire d'un site de communauté entièrement dynamique. Ce site a été construit entièrement en java, avec le serveur d'application « Enhydra » et la base de donnée « PostgreSQL ». L'ensemble des données utilisées dans ce site sont au format XML et sont exprimées en Unicode (UTF-8). Tous les outils utilisés sont des outils « open source ». Ce site est accessible à l'URL : `http ://www.papillon-dictionary.org/`.

3.1.1. *Services disponibles*

Outre les services inhérents à la création collaborative d'une base lexicale multilingue (interface de consultation/modification de la base Papillon, gestion des contributions, des utilisateurs, de l'historique, etc.), nous avons souhaité fournir 3 autres services pour rendre le site à la fois plus pratique et plus attractif :

– **Un service d'archivage de la liste de discussion des utilisateurs**. Ce service,

quasi systématique dans les sites de communauté, présente ici une particularité. En effet, les discussions se font dans différentes langues et les messages échangés arrivent dans des encodages divers (ISO-LATIN1, SJIS, EUC, UTF8, etc.). Ces encodages doivent être correctement reconnus et transcris en UTF8 afin d'être archives. Nous avons dû adapter les outils standard pour cela.

– **Un service de partage d'informations entre les utilisateurs**. Ce service permet à un utilisateur du site Papillon d'y ajouter des informations qu'il trouve pertinentes pour les autres utilisateurs. L'interface de ce service a été particulièrement soignée, afin que tout utilisateur, quel que soit son niveau en informatique, puisse obtenir un résultat très satisfaisant. L'utilisateur se contente de rédiger un document en HTML avec n'importe quel éditeur du commerce. Ce document peut contenir divers fichier HTML (avec des liens internes ou externes), des images ou d'autres données. Il se contente ensuite de transférer ce document sur le site (en utilisant son client http préféré). Ce document sera analysé, le code html sera corrigé automatiquement, et il sera intégré au site Papillon. Il prendra la même forme que les autres pages (mêmes entêtes, mêmes fonctionnalités, même comportement général). Ce service permet de plus de gérer des documents multilingues : le document est disponible dans plusieurs langues et les lecteurs verront automatiquement la version qui leur convient.

– **Un service d'accès unifié à de nombreux dictionnaires**. Ce service permet à tout utilisateur d'accéder, par une interface unique, à de nombreux dictionnaires monolingues et bilingues. Un utilisateur cherchant un mot dans une langue pourra obtenir les entrées correspondant à ce mot dans l'ensemble des dictionnaires disponibles. Avec ce service, nous espérons attirer des utilisateurs qui ne sont, dans un premier temps que « consommateurs » de données lexicales. Avec le temps, nous pensons que certains d'entre eux deviendront, à leur tour, « producteurs » de données.

3.1.2. *Organisation des données*

Les données lexicales disponibles sur le site Papillon sont diverses. Il peut s'agir de données de la base lexicale Papillon en cours de construction ou de données provenant de dictionnaires existants libres de droits ou qui nous ont été transmis par leurs auteurs. Ces données lexicales sont réparties dans 3 « zones » :

– *Les limbes* contiennent les données lexicales dans leur format propriétaire d'origine. Lorsqu'un dictionnaire nous est fourni, il est disponible dans cette zone en attendant d'être « récupéré ». Chaque dictionnaire est associé à un fichier de méta données contenant toutes les informations disponibles à son propos (son nom, les languages qu'il contient, la date de création, sa taille, ses auteurs, son domaine éventuel, etc.). Les dictionnaires présents dans cette zone peuvent être téléchargés tels-quels, mais les entrées qu'ils contiennent ne peuvent être obtenues individuellement.

– Après récupération, les données lexicales des limbes sont disponibles dans *le purgatoire*. Cette récupération consiste à définir en XML la structure du dictionnaire original (qui n'est pas modifiée). Les données sont ensuite transformées en XML en encodées en UTF8.
Les éléments de la structure XML obtenue (ou de la structure originale si le diction-

naire était déjà disponible en XML) sont ensuite identifiés grâce au mécanisme de pointeurs CDM détaillé § 3.3.2. Cette identification est stockée dans le fichier de méta données du dictionnaire. Elle permet d'accéder individuellement aux entrées de ce dictionnaire par des requêtes émises via l'interface unifiée disponible sur le site Papillon.

– *Le paradis* contient les entrées de la base lexicale multilingue Papillon. Ce dictionnaire est lui aussi accessible par l'interface unifiée disponible sur le site. Par contre, il est le seul dictionnaire qui puisse être modifié par l'interface d'édition.

3.1.3. *Implémentation*

L'application Papillon est organisée en trois couches (figure 4). La première couche prend en charge l'interface vers les utilisateurs. La seconde couche contient l'ensemble des services fournis aux utilisateurs. La dernière couche gère la persistance des données XML manipulées. Cette architecture rend facile l'ajout et la modification des interfaces de l'application vers ses clients (une interface WML est en cours de développement pour permettre l'accès via des téléphones portables). Elle permet aussi de s'abstraire de la manière dont les données sont stockées. Dans l'implémentation actuelle, les données XML sont stockées dans une base relationnelle « open source », via l'API java standard JDBC. Cette application peut être dupliquée sur de nombreux serveurs afin de répartir la charge.

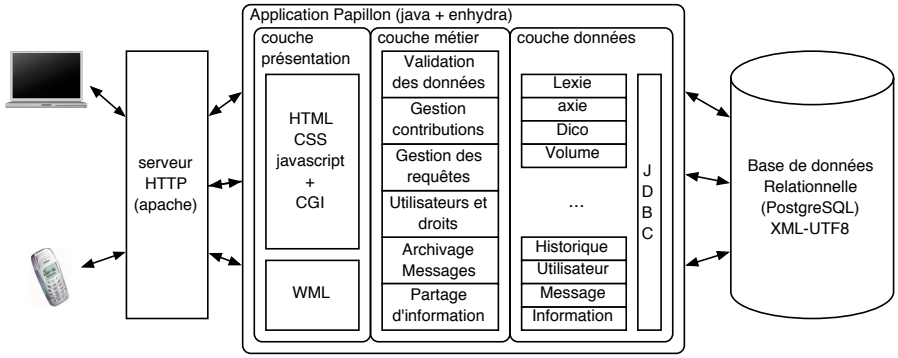

Figure 4. *Architecture du serveur de communauté Papillon.*

3.2. *Représentation des données de la base lexicale multilingue Papillon*

Dans le cadre de travaux antérieurs ([SÉR 94a, SÉR 94b] puis [MAN 01] et [MAN 02]), nous avons défini un cadre de représentation de données lexicales hétérogène. Pour le projet Papillon, nous utilisons ce cadre implémenté en XML (figure 5).

L'implémentation de notre microstructure s'appuie sur 3 niveaux de représentations : un cadre générique de représentation de dictionnaires (Dictionary Markup Lan-

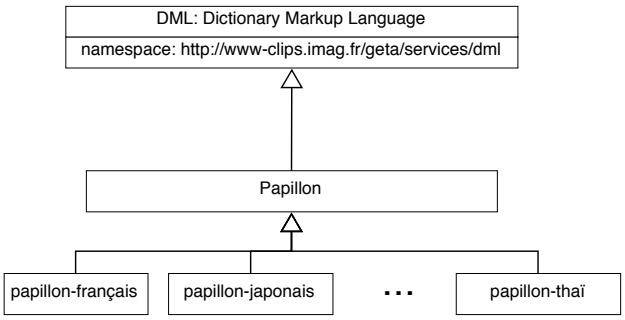

Figure 5. *Définition XML des structures de dictionnaires monolingues de la base Papillon.*

guage), un niveau de représentation de la structure commune à tous les dictionnaires de la base Papillon et un niveau de représentations représentant les spécificités de chaque langue.

L'ensemble des structures est implémenté en XML. Ces trois niveaux sont implémentés à l'aide de schémas XML qui incluent un mécanisme d'héritage simple.

3.2.1. *DML : description de dictionnaires en XML*

Le premier niveau de représentation est un cadre général, qui défini les concepts généraux, qui peuvent être utilisés pour représenter n'importe quel dictionnaire. Ce cadre est générique et permet la représentation aisée de dictionnaires hétérogènes.

Ce cadre est un espace de nom (namespace) XML nommé DML (Dictionary Markup Language, figure 6). Toute donnée d'une base lexicale peut être décrite en utilisant des éléments DML. Ce cadre défini non seulement les objets de base permettant de représenter des structures lexicales complexes (arbre, graphe, automate, etc.), mais aussi des types généraux utiles (booléen, date, langue, etc.) ainsi que les APIs permettant à des clients d'utiliser les données décrites ou à des fournisseurs de rajouter des services.

3.2.2. *Structure commune des dictionnaires monolingues Papillon*

La structure commune des dictionnaires monolingues de la base Papillon est définie par un schéma XML qui utilise l'espace de nom défini dans le DML. Ce schéma XML décrit la structure générale des lexies.

```
<element name="lexie">
  <complexType>
    <sequence>
      <element ref="d:headword" minOccurs="1" maxOccurs="1" />
      <element ref="d:writing" minOccurs="0" maxOccurs="1" />
      <element ref="d:reading" minOccurs="0" maxOccurs="1" />
```

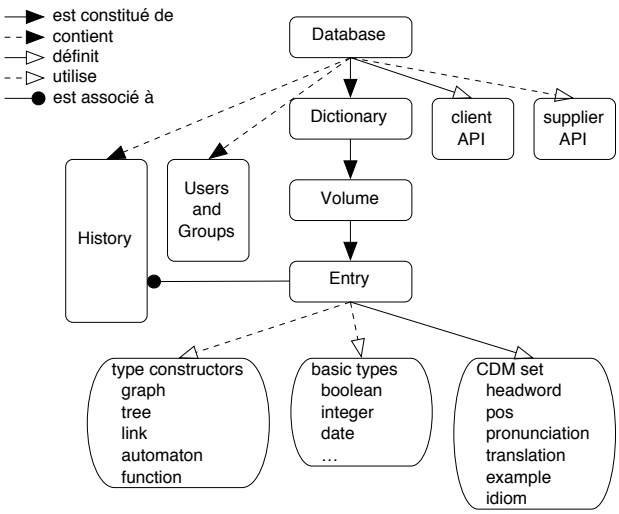

Figure 6. *Les concepts du DML.*

```
      <element ref="d:pronunciation" minOccurs="0" maxOccurs="1" />
      <element ref="d:pos" minOccurs="1" maxOccurs="1" />
      <element ref="d:language-levels" minOccurs="0" maxOccurs="1" />
      <element ref="d:semantic-formula" minOccurs="1" maxOccurs="1" />
      <element ref="d:government-pattern" minOccurs="0" maxOccurs="1" />
      <element ref="d:lexical-functions" minOccurs="0" maxOccurs="1" />
      <element ref="d:examples" minOccurs="0" maxOccurs="1" />
      <element ref="d:full-idioms" minOccurs="0" maxOccurs="1" />
      <element ref="d:more-info" minOccurs="0" maxOccurs="1" />
    </sequence>
    <attribute ref="d:id" use="required" />
  </complexType>
</element>
```

Chaque élément de cette structure est lui aussi défini à ce niveau. Ainsi, à ce niveau, l'élément « pos » (catégorie) est défini comme une chaîne de caractères, tandis que l'élément « lexical-functions » (les fonctions lexicales) est défini comme une liste de « function », qui est un type de base du DML.

```
<element name="pos" type="d:posType" />
<simpleType name="posType">
  <restriction base="string" />
</simpleType>
<element name="lexical-functions">
  <complexType>
    <sequence maxOccurs="unbounded">
```

```
      <element ref="d:function" />
    </sequence>
  </complexType>
</element>
```

3.2.3. *Adaptation de la structure à chaque langue de la base*

Le troisième niveau permet d'adapter légèrement la structure commune définie ci-dessus aux spécificités de chacune des langues de la base lexicale Papillon. Ainsi, chaque langue de la base possède un schéma qui lui est propre et qui redéfinit certains des éléments XML évoqués plus haut. Ainsi, l'élément « pos » (catégorie) du dictionnaire français est redéfini comme suit :

```
<simpleType name="posType">
  <restriction base="d:posType">
    <enumeration value="n.m." />
    <enumeration value="n.m. inv." />
    <enumeration value="n.m. pl." />
    <enumeration value="n.m., f." />
    <enumeration value="n.f." />
    <enumeration value="n.f. pl." />
    ...
  </restriction>
</simpleType>
```

De la même manière, une langue peut redéfinir les valeurs possibles pour les niveaux d'usage et de politesse. Il est souhaitable que chaque langue définisse de manière explicite une liste fermée de valeurs possibles pour ces éléments plutôt que de se fier à la structure générale qui, par défaut, accepte n'importe quelle chaîne de caractère. C'est en effet à partir de ces schémas XML que sont générées les interfaces de saisie.

3.3. *Accès unifié à des dictionnaires existants*

3.3.1. *Interface de consultation*

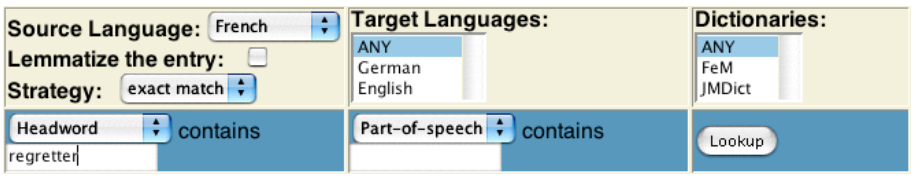

Figure 7. *Interface d'accès unifié aux dictionnaires du purgatoire et du paradis.*

L'interface unifiée permet de faire une recherche sur le lemme, sur la catégorie, sur la prononciation ou sur une traduction d'une entrée de dictionnaire (figure 7). Pour certaine langue, une lemmatisation est disponible pour permettre à l'utilisateur débutant d'entrer une occurrence quelconque. On peut de plus choisir les langues cibles et les dictionnaires dans lesquels se fera la recherche. Le service de lemmatisation est un service externe que nous accédons via Internet.

3.3.2. *Un mécanisme de pointeurs communs : CDM*

Cette interface de requête ne peut fonctionner que s'il existe un moyen d'identifier les éléments sur lesquels portent la recherche (entrée, catégorie, prononciation, traduction) dans des dictionnaires ayant des structures différentes. Cette identification est faite en établissant une correspondance entre un élément particulier d'un dictionnaire et un éléments définis par l'espace de nom DML (cf. § 3.2.1).

Le sous-ensemble du DML avec lesquels une telle correspondance peut être faite est nommé CDM (Common Dictionary Markup). Il est en constante évolution. La figure 8 en donne un exemple d'utilisation.

Element CDM	**équivalent TEI**	**FeM**	**OHD**	**NODE**
<entry>	(entry)	<fem-entry>	<se>	<se>
<headword>	(hom)(orth)	<entry>	<hw>	<hw>
<pronunciation>	(pron)	<french_pron>	<pr><ph>	<pr><ph>
<etymology>	(etym)			<etym>
<syntactic-sense>	(sense level="1")		<sense n=1>	<s1>
<pos>	(pos)(subc)	<french_cat>	<pos>	<ps>
<lexie>	(sense level="2")		<sense n=2>	<s2>
<indicator>	(usg)	<gloss>	<id>	
<label>	(lbl)	<label>		<la>
<example>	(def)	<french_sentence>	<ex>	<ex>
<definition>	(eg)			<df>
<translation>	(trans)(tr)	<english_equ> <malay_equ>		<tr>
<collocate>	(colloc)		<co>	
<link>	(xr)	<cross_ref_entry>	<xr>	<xg> <vg>
<note>	(note)		<ann>	

Figure 8. *Correspondance entre les éléments CDM et des éléments des dictionnaires TEI, FeM (français-anglais-malais), Oxford-Hachette (français-anglais) et Oxford (anglais)*

3.3.3. *Présentation des résultats*

Le résultat d'une requête est un ensemble d'unités codées en XML. Leurs structures sont variées car elles proviennent de dictionnaires différents.

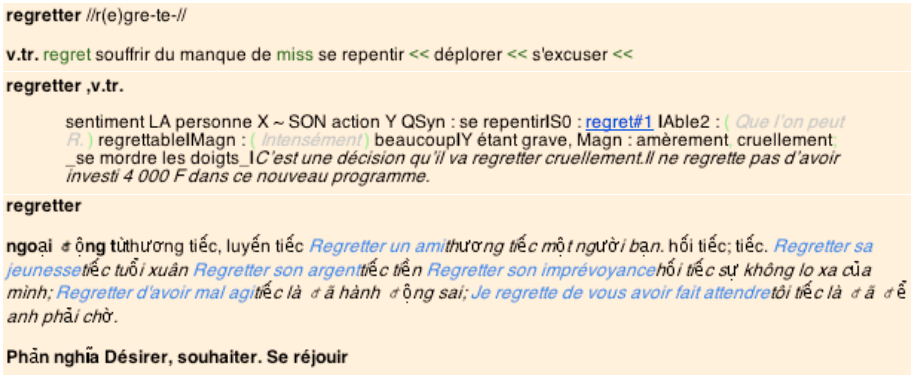

Figure 9. *Trois résultats de la requête « regretter ». Le premier résultat provient du dictionnaire français-anglais-malais (FeM), le suivant de la base Papillon (entrée récupérée du DiCo), le dernier provient du dictionnaire français-vietnamien (vietDict).*

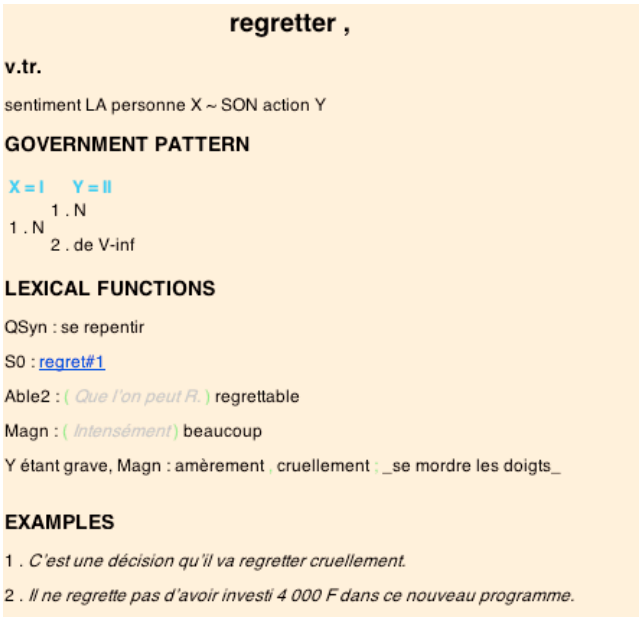

Figure 10. *Forme inspirée du DEC pour la lexie « regretter.1 » du dictionnaire Papillon.*

Ces structures sont transformées en des « formes » qui dépendent des possibilités de l'interface utilisateur. Ces formes sont obtenues en appliquant des transformations XSL aux résultats de la requête. Ces transformations sont elles-mêmes des données du

serveur et plusieurs transformations peuvent être disponibles pour une même structure. Un utilisateur peut ajouter sa propre forme au serveur.

L'utilisateur peut facilement choisir la forme qu'il souhaite. Celle-ci peut mettre en valeur certaines informations lexicales ou en cacher d'autres. La figure 10 présente la lexie « regretter » (sentiment : personne X regrette son action Y) présentée selon une forme inspirée du DEC.

4. Construction d'une base lexicale initiale

Notre stratégie de construction commence la construction automatique d'une première base lexicale multilingue qui sert de base au travail contributif (l'amorçage). Cette construction se fait en deux étapes. Dans un premier temps, nous construisons les dictionnaires monolingues. Dans un second temps, nous construisons la base d'acceptions qui fera le lien entre les différents dictionnaires monolingues.

Ces deux étapes se font en combinant un certain nombre de ressources lexicales monolingues et bilingues. Un exemple de tels croisements est largement illustré par [QUA 01]. Cependant, l'architecture lexicale choisie nous impose d'établir les liens de traduction au niveau des lexies (acceptions monolingues), ce qui pose le problème de la sélection correcte des différents sens de termes polysémiques.

Pour pouvoir faire cette sélection de manière automatique, nous avons choisi d'utiliser le modèle des vecteurs conceptuels ([CHA 96], [LAF 99]) qui défini une notion de distance que nous interprétons comme une distance sémantique.

4.1. *Le modèle des vecteurs conceptuels*

4.1.1. *Définition et notion de distance*

Le modèle des vecteurs conceptuels a été présenté dans [LAF 02]. On associe à tout segment textuel (mot, syntagme, texte), une association thématique qui prend la forme d'une vecteur de concepts. Le jeux de concepts est prédéfini et constitue un espace générateur sur lequel les sens peuvent se projeter (la figure 11 donne une représentation graphique de deux de ces vecteurs). Par exemple, les sens de ‹*barrage*› peuvent être projetés sur les concepts suivant (les CONCEPT[*intensité*] étant ordonnés par intensité décroissante) : $V_{barrage} =$ (BARRIÈRE[*0.84*], OBSTACLE[*0.83*], ÉLECTRICITÉ[*0.82*], SPORT[*0.77*], FLOT[*0.76*], GUERRE[*0.76*], ...).

Dans ce modèle vectoriel, nous disposons des notions de *similarité* (utilisée habituellement en recherche d'information) et de *distance angulaire* $D_A(V_1, V_2)$. Cette dernière est une vrai mesure de distance (contrairement à la notion de similarité) et elle vérifie les propriétés de réflexivité, symétrie et inégalité triangulaire. Nous l'interprétons comme une évaluation de la *proximité thématique* entre sens de mots.

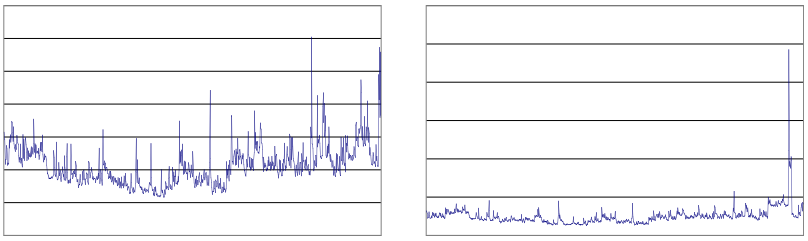

Figure 11. *Représentation graphique des vecteurs des termes* échange *(très polysémique) et* cession.

4.1.2. *Notion de contextualisation faible*

L'opération de contextualisation faible, notée $\Gamma(V_1, V_2)$ (elle aussi définie dans [LAF 02]), nous est aussi très utile dans le travail d'amorçage. En effet, les composantes communes à V_1 et V_2 seront fortement présentes dans $\Gamma(V_1, V_2)$. Cette notion permet d'amplifier les propriétés saillantes d'un vecteur dans un contexte donné.

Pour exploiter cette notion, nous associons à chaque vocable un vecteur égal à la somme normée des vecteurs de ses acceptions. En appliquant l'opération de contextualisation faible à un vecteur de vocable polysémique et à un vecteur contexte (de vocable polysémique ou non), nous obtenons un vecteur qui se rapprochera de l'un des sens du vocable considéré.

Par exemple, le terme *bank* est ambigu et son vecteur est globalement la moyenne entre les vecteurs des sens *river bank* et *money institution*. Si le vecteur de *bank* est contextualisé par celui de *river*, alors les concepts du champs sémantique lié à la finance seront considérablement inhibés.

4.2. *Construction de la base*

4.2.1. *Construction des dictionnaires monolingues*

En compilant les informations extraites de dictionnaires variés, (Hachette, Thésaurus Larousse, dictionnaire de synonymes de l'université de Caen, Wordnet, dictionnaire Oxford...) nous avons construit les lexies des dictionnaires monolingues français et anglais. Ces lexies, sont codées selon le schéma XML Papillon, mais contiennent très peu d'information (mot forme, catégorie et définition en langue naturelle).

Notre premier travail consiste à calculer le vecteur conceptuel associé à chacune de ces lexies. Le jeu de concept (les dimensions de l'espace) est prédéfini à l'aide des concepts présents dans le thésaurus Larousse.

Un indexage manuel de 5000 termes dans chaque langue, nous permet de connaître un premier ensemble de vecteurs. Ensuite, la définition de chaque lexie est analysée avec l'analyseur morphosyntaxique SYGMART. À partir des vecteurs des mots

connus de la définition, et de l'arbre d'analyse produit, nous calculons les vecteurs associés à chaque lexie et à chaque mot forme. Ce processus est itéré jusqu'à stabilité.

Nous disposons ainsi de dictionnaires monolingues vectorisés que nous définissons comme suit :

$$D_a = \{Lex_i\} \quad \text{et} \quad Lex_i = (w_i, \text{cat}_i, \text{def}_i, V_i) \tag{1}$$

Dans un dictionnaire monolingue vectorisé D_a, chaque vocable v correspond à n lexies $Lex_i = (v, \text{cat}_i, \text{def}_i, V_i)(n \geq 1)$. Si $n = 1$ le vocable v est strictement monosémique.

Nous donnons ci dessous un exemple des lexies du vocable *exiger* (dans cet exemple, les définitions sont issues du dictionnaire Hachette).

> **exiger.1** V, #s=1# Réclamer, en vertu d'un droit réel ou que l'on s'arroge. (Exiger le paiement de réparations) - (Exiger que (+subj)) (Il exige qu'on vienne), V_1
> **exiger.2** V, #s=2# Imposer comme obligation. (Allez-y, le devoir l'exige) (Les circonstances exigent que vous refusiez), V_2
> **exiger.3** V, Nécessiter. (Construction qui exige une main-d'oeuvre abondante), V_3

4.2.2. *Construction du dictionnaire interlingue d'acception*

Initialement, nous créons une acception interlingue (axie) pour chaque lexie des dictionnaires monolingues. Chaque acception est associée à un vecteur conceptuel. Initialement, ce vecteur est égal au vecteur de la lexie correspondante. La figure 12 présente l'état de la base lexicale après cette étape.

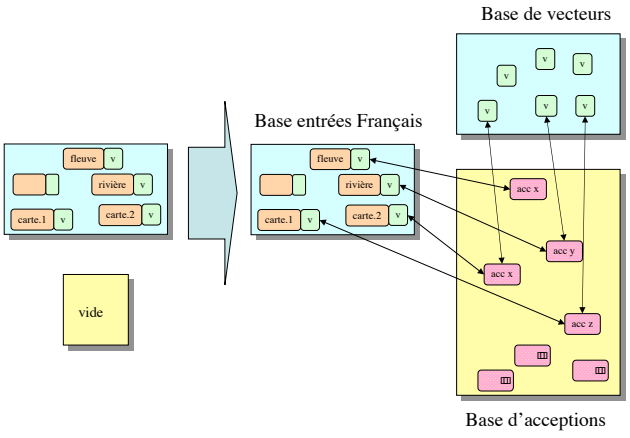

Figure 12. *Chargement initial du dictionnaire interlingue d'acceptions.*

Le travail qu'il reste à faire consiste à réduire cet ensemble d'acceptions à l'aide d'associations bilingues extraites de dictionnaire D_{a-b} (dictionnaires bilingues d'une langue source A vers une langue cible B) que nous définissons comme suit :

$$A(D_{a-b}, w) = \{Sa_i\} \quad \text{et} \quad Sa_i = (w, \text{cat}, \text{glose}^*, \text{equiv}^+)$$

Dans le dictionnaire D_{a-b}, le terme w est associée à n sous-entrée. Chaque sous-entrée contient : une information morphologique (au moins la catégorie morphosyntaxique, Nom, Verbe, Adjectif, Adverbe), zéro ou plus gloses, et au moins un équivalent dans la langue cible. Les gloses sont des termes optionnels qui permettent à l'utilisateur de sélectionner le sens dont il est question si le terme est polysémique. Ce sont ces même gloses qui permettent d'associer via les vecteurs conceptuels une sous-entrée du dictionnaire bilingue à une lexie du dictionnaire monolingue (en cas d'ambiguïté). Un exemple typique d'association bilingue (anglais-français) est :

demand ==
demand.1 VT, g{money, explanation, help}, e{exiger, réclamer})
demand.2 VT, g{higher pay}, e{revendiquer, réclamer})
demand.3 N, g{person}, e{demande})
demand.4 N, g{duty, problem, situation}, e{revendication, réclamation})
demand.5 N, g{for help, for money}, e{demande})

Dans le cas d'associations bilingues entre deux équivalents monosémiques (par exemple *babouin* $\rightarrow$ *baboon*) nous fusionnons les acceptions correspondantes. Un avertissement est émis pour le lexicographe si la distance entre les vecteurs des 2 acceptions est trop importante. Dans ce cas, il est vraisemblable, qu'au moins un des vecteurs conceptuels ne dispose pas d'activation pertinente.

Pour les autres associations bilingues, il nous faut identifier la lexie correspondant à chaque sous-entrée et la lexie correspondant à chaque équivalent.

Appariement lexie-association Pour chaque sous-association Sa_i, nous calculons un vecteur contexte V_C qui est la somme des vecteurs de ses gloses :

Le vecteur de la glose est le vecteur global, toutefois on sélectionne les sens dont les catégories morphosyntaxiques sont compatibles. Le vecteur associé à Sa_i est le calcul de la contextualisation faible (fonction Γ) entre le vecteur issu du dictionnaire monolingue pour w et le vecteur contexte. Le vecteur estimé $V_{\approx}$ de Sa_i est :

$$V_{\approx}(Sa_i) = \Gamma(V(w), V_C(Sa_i)) \tag{2}$$

À chaque sous-association Sa_i il est maintenant possible d'apparier un vecteur issu du dictionnaire monolingue. Il s'agit du vecteur (et donc du sens) qui est le plus proche du vecteur estimé (figure 13).

$$V(Sa_i) = Min(D_A(V(Se_j), V_{\approx}(Sa_i)) \tag{3}$$

Ainsi, nous avons apparié une partie des lexies et des associations bilingues. C'est-à-dire qu'en pratique nous avons pu fournir un vecteur aux associations bilingues ce

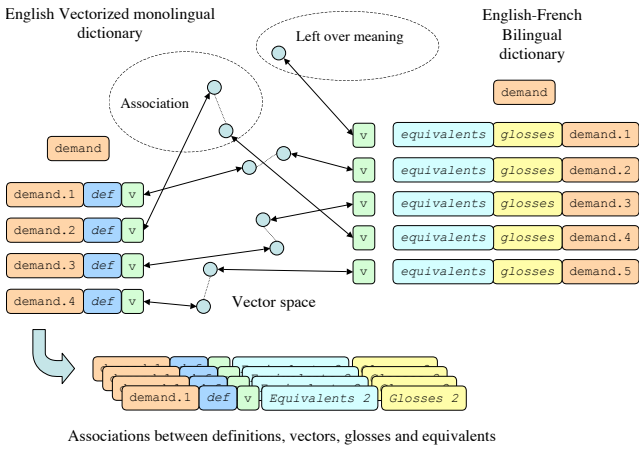

Figure 13. *Les associations sont appariées aux lexies dont le vecteur sémantique est le plus proche.*

qui est la condition pour les relier à des acceptions (dans les cas de polysémie). À la fin de ce processus, certaines des lexies du dictionnaire monolingue vectorisé disposent d'un lien unique vers une entrée du dictionnaire bilingue.

Liaison lexie-acceptions Il s'agit ici d'associer une lexie S_b du langage cible à une acception interlingue Ax_a issue d'une lexie du langage source.

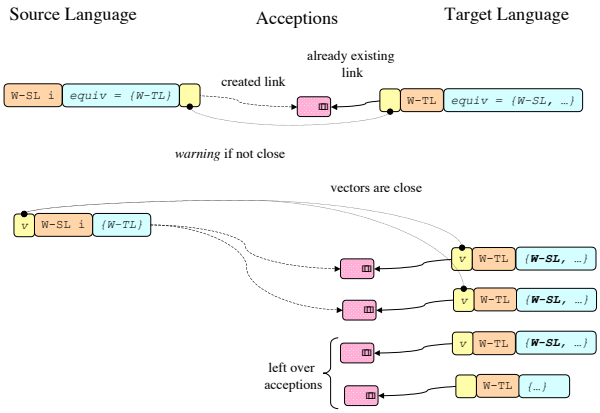

Figure 14. *Liaison Lexie-Acception dans le cas d'un équivalent monosémique (en haut) et d'un équivalent polysémique (en bas).*

On considèrera deux vecteurs conceptuels comme *suffisamment proche* si leur distance thématique est inférieur à un seuil t. Plus ce seuil est faible, plus le niveau de

confiance du lien vers l'acception est fort. En retour, il risque d'être difficile d'automatiquement réaliser l'association. Une valeur de seuil acceptable s'avère être $\pi/4$. Les différentes situations sont les suivantes :

1) **Un sens S vers un seul équivalent monosémique.** Ce cas consiste à sélectionner directement les termes. Les vecteurs conceptuels ne sont pas utilisés ici, si ce n'est pour effectuer une vérification. Si les deux vecteurs conceptuels ne sont pas raisonnablement proches, un message d'alerte est envoyé au lexicographe. Le problème peut aussi bien venir d'une erreur des dictionnaires bilingues, ou qu'un des vecteurs (ou les deux) à des activations inadéquates.

2) **Un sens S vers un équivalent polysémique.** Il faut alors sélectionner le sens équivalent S_b qui pourrait être acceptable (figure 14) Un filtre consiste à sélectionner les équivalents inverses, puis parmi les sens restant (s'il y en a plusieurs) choisir celui dont le vecteur est le plus proche.

3) **Un sens S vers plusieurs équivalent polysémique.** Ce cas est une généralisation des cas précédents.

4) **Cas d'erreur.** L'erreur principale provient de la constitution d'un ensemble vide. Cela peut arriver si les informations dans le dictionnaire bilingue sont inconsistantes. On remarquera que cela arrive relativement souvent en pratique.

4.2.3. *Nettoyage des liens*

L'architecture lexicale choisie impose des contraintes de bonne formation qu'il nous faut prendre en compte dans le peuplement du dictionnaire interlingue d'acceptions.

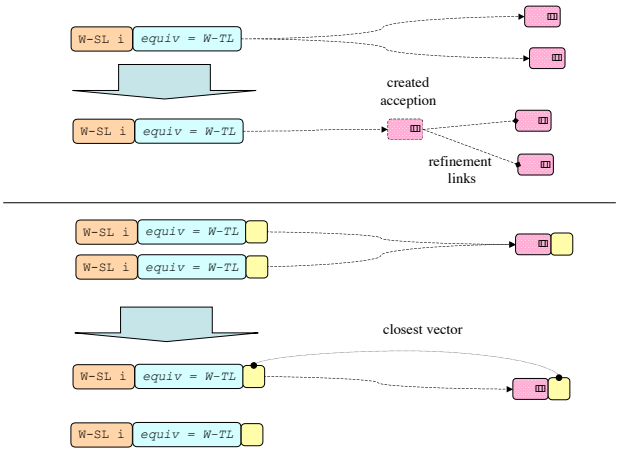

Figure 15. *Nettoyage de liens.* Partie supérieure *: Liens multiples. Une acception intermédiaire est créée ainsi que des liens inter-acceptions.* Partie inférieure *: Sens multiples. On conserve le lien minimisant la distance entre acception interlingue et lexie.*

1) Il y a, au plus, un lien possible d'une lexie vers une acception.

2) Deux lexies ne peuvent pas être liée à la même acception. Si ces sens sont synonymes, cette relation sera explicitée à l'aide d'une fonction lexicale (voir [SCH 01, SCH 02] pour une généralisation de l'approche à plusieurs fonctions lexicales caractéristiques).

Le nettoyage des liens consiste à détruire des liens qui auraient été créés abusivement (au vue des contraintes de bonne formation de la base interlingue) et d'en réévaluer certains. Deux cas peuvent se présenter : une lexie particulière est liée à plusieurs acceptions interlingues, ou une acception interlingue particulière est liée à plus d'une lexie dans une même langue.

1) **Liens multiples.** Pour résoudre ce problème, il est nécessaire de créer une acception intermédiaire. Le sens est alors lié seulement à la nouvelle acception (les liens précédents sont détruits). Des liens de raffinement de sens sont créés entre la nouvelle acception et les précédentes (figure 15).

2) **Sens multiples.** Nous avons à choisir parmi les sens lequel doit être retenu comme liable à l'acception, les autres liens étant détruits. La sélection se fait sur la base du plus proche vecteur.

En toute généralité, les deux situations peuvent survenir en même temps. Le processus de nettoyage est appliqué itérativement avec une priorité en faveur de la création d'acceptions intermédiaires.

4.3. *Premiers résultats et discussion*

Un certain nombre d'aspects doivent être mentionnés leur développement dépassant l'objet de cet article.

1) Vecteurs d'acceptions. Pour chaque acception, nous calculons son propre vecteur conceptuel. Ce vecteur est la moyenne des vecteurs des sens monolingues (dans l'ensemble des langues de la base) associés à l'acception. Ces vecteurs sont stockés comme s'ils constituaient un nouveau dictionnaire monolingue (cf. figure 14). Ils servent essentiellement à confirmer ou infirmer une proposition de liens lors de l'ajout de nouvelles entrées monolingues (à des acceptions déjà existantes).

2) Pondération des liens. Chaque lien créé automatiquement se voit attribuer une valeur de confiance (entre 0 et 1). Cette valeur correspond à la similarité entre le vecteur de l'acception et celui de l'entrée monolingue. Si un lien est confirmé par le lexicographe, le niveau de confiance vaut 1. L'exploitation des fonctions lexicales permet de moduler la valeur de confiance attribuée par le système ([SCH 01]).

3) Le processus de peuplement est effectué itérativement par des agents autonomes. Un agent explore la base d'acceptions et essaye d'évaluer les liens ou d'en créer. Par exemple, une acception pendante (avec un seul lien) doit être reliée à une entrée monolingue pour chacune des autres langues. Dans le cas d'entrée polysémique ou d'équivalent multiple, seule la source monolingue vectorisée nous concerne dans

la mesure où seuls les vecteurs conceptuels sont à la base du processus de décision. Les entrées orphelines doivent également être traité par la recherche d'une acception adéquate. Le processus est globalement convergent surtout dans la mesure où des liens sont fortement confirmés par les contributeurs humains.

Nos expériences de croisement de dictionnaires et de peuplement et liage automatique de la base d'acceptions nous ont permis dans le cas du français-anglais de générer environ 20000 acceptions dont environ 15000 était correctement liés. Le reste consistait en acceptions pendantes (soit du côté anglais soit du côté français). La plus grande difficulté concerne les entrées qui ne sont pas directement lexicalisées dans une langue. Dans ce cas, l'équivalent se réduit à une phrase explicative ou à une paraphrase. Ces traductions ne se retrouvent pas dans les dictionnaires monolingues de la langue cible. Par exemple le terme abêtir se traduit par *to make stupid, to turn into a moron* qui ne constitue pas des entrées du dictionnaire monolingue anglais. Afin de régler ce problème nous avons décidé de générer de telles entrées monolingues qui seront complétées par la suite (en particulier au niveau de leurs fonctions lexicales). Une petite frange d'acceptions (moins de 4%) et de sens (monolingue français ou anglais) sont incorrectement liés ou disposent de liens dont le seuil de confiance est inférieur à $1/2$. Il s'agit en général de termes très polysémiques (verbes support par exemple) qui génèrent beaucoup de forme lexicale voire de locutions dont la forme exacte peut être sujette à des variations. Cela engendre en particulier des amas d'acceptions lié par des raffinements de sens. En toute objectivité, l'approche fournit des résultats dans le rappel est important mais dont la précision est parfois médiocre. En l'occurrence, c'est très exactement la situation souhaitée où le lexicographe humain peut intervenir. Les termes polysémiques dont les champs sémantiques sont relativement distincts (par exemple un terme comme botte) sont correctement traités par les vecteurs conceptuels. Notre approche permet également d'améliorer la qualité des vecteurs conceptuels. Il s'agit ici d'une exploitation du graphe que représente les liens et les acceptions (indépendamment de sa construction). En particulier l'apport des lexicographes sur des informations lexicales permet d'augmenter la pertinence de certains vecteurs ce qui en retour améliore les performances du processus de peuplement.

5. Conclusion

Cet article présente un projet de construction d'une base lexicale multilingue linguistiquement riche. Par l'utilisation d'une stratégie basées sur le modèle « open source », nous souhaitons réduire les coût d'une telle construction en utilisant les compétences naturelles d'internautes volontaires. L'adoption d'une architecture lexicale interlingue qui sépare clairement les informations monolingues des informations interlingues présente de nombreux avantages dans ce cadre. D'une part, elle permet de distinguer les contributions interlingues des contributions monolingues qui requièrent des compétences différentes. D'autre part, elle permet de s'appuyer sur des langues bien dotées pour construire des données interlingues brutes impliquant des langues plus pauvres.

Le choix d'une architecture linguistique monolingue basée sur la composante lexicale de la théorie sens-texte d'Igor Mel'čuk, favorisera une réutilisation future de ces données dans de nombreuses et diverses applications de traitement des langues. De plus, la richesse des données construites rend plus intéressant le travail de contribution, les utilisateurs apportant de nombreuses informations originale, que l'on ne trouve actuellement dans aucun dictionnaire existant.

Le serveur de communauté sur lequel s'appuie le dictionnaire Papillon est en cours de construction, néanmoins, il est déjà fonctionnel et permet notamment un accès unifié à des dictionnaire existants. Ce service permet de rendre le site attractifs à des utilisateurs qui seront peut-être nos futurs contributeurs. Il nous a permis de plus de valider notre choix d'utiliser des outils standards (serveurs d'applications Java, XML pour la représentation des données, XSL pour leur manipulation, etc.). Ce serveur facilite aussi le travail des différents partenaires en proposant des services de partage de documents et d'archivage de liste de diffusion. Notre prochain objectif est d'ouvrir un service qui permettra des contributions en ligne par l'intermédiaire d'interfaces adaptables à l'utilisateur et à son environnement. À terme, nous souhaitons permettre des contributions en ligne (à partir d'un navigateur internet standard) et hors ligne (à partir d'application autonomes spécialisées). Le défi qui suivra consistera à animer une communauté de contributeurs et à trouver différentes motivations à même d'encourager la participation d'utilisateurs aux profils divers.

Une première réponse à ce défi réside dans la stratégie de construction employée, qui impose une phase d'amorçage assez complexe, mais néanmoins nécessaire pour disposer d'un ensemble de données suffisant qui sert à la fois de base de travail (les contributions sont vues comme des modifications de ces données) et de motivation (les données ainsi construites sont accessibles en ligne). Nos travaux préliminaires nous ont permis de produire une base d'acception sur le français et l'anglais par une méthode adaptable à d'autres langues avec un coût raisonnable. Il nous ont permis aussi d'associer un vecteur conceptuel à chaque acception créée. Ainsi, lors de la prise en compte de contributions, nous disposons de critères nous permettant de construire un agent automatique servant à la validation.

6. Bibliographie

[BAR 01] BARRIÈRE C., COPECK T., « Building Domain Knowledge from Specialized Texts », *TIA 2001*, Nancy, 2001.

[BLA 95] BLANC E., « Une maquette de base lexicale multilingue à pivot lexical : PARAX », *Lexicomatique et Dictionnairique, Actes du colloque LTT*, Universités Francophones, Actualités scientifiques, AUPELF-UREF, 1995, p. 43-58.

[BOU 93] BOURRIGAULT D., « Analyse locale pour le repèrage des termes complexes dans les textes », *TAL*, vol. 34, n° 2, 1993, p. 105-118.

[CHA 90] CHAUCHÉ J., « Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance », *TA Informations*, vol. 31, n° 1, 1990, p. 17-24.

[CHA 96] CHAUCHÉ J., SANDFORD E., « Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance », *Actes de MIDDIM-96*, Le Col de Porte, France, Août 1996, p. 56-66.

[CRU 95] CRUSE D. A., TOGIA P., « Towards a cognitive model of antonymy », *Lexicology*, vol. 1, 1995, p. 113-141.

[DEE 90] DEERWESTER S., DUMAIS S., LANDAUER T., FURNAS G., HARSHMAN R., « Indexing by latent semantic analysis », *Journal of the American Society of Information Science*, vol. 416, n° 6, 1990, p. 391-407.

[DES 02] DESPERRIER J.-M., « Analyzis of the results of a collaborative project for the creation of a Japanese-French dictionary », *Papillon'2002 Seminar*, NII, Tokyo, Japan, July 2002, http ://www.papillon-dictionary.org/ConsultInformations.po.

[FEL 95] FELLBAUM C., « Co-occurrence and antonymy », *International Journal of Lexicography*, vol. 8, 1995, p. 281-303.

[FIS 73] FISCHER W. L., *Äquivalenz und Toleranz Strukturen in der Linguistik zur Theory der Synonyma*, Max Hüber Verlag, München, 1973.

[GWE 87] GWEI G., FOXLEY E., « A Flexible Synonym Interface with application examples in CAL and Help Environments », *The Computer Journal*, vol. 30, n° 6, 1987, p. 551-557.

[HAM 01] HAMON T., NAZARENKO A., « La structuration de terminologie : une nécessaire coopération », *TIA 2001*, Nancy, 2001.

[HAT 01] HATHOUT N., « Analogies morpho-synonymiques. Une méthode d'acquisition automatique de liens morphologiques á partir d'un dictionnaire de synonymes », *TALN 2001*, Tours, July 2001, p. 223-232.

[HEA 98] HEARST M., « Automated discovery of Wordnet relations », FELLBAUM C., Ed., *Wordnet An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998, p. 131-151.

[JUS 91] JUSTESON J., KATZ S., « Co-occurrences of antonymous adjectives and their contexts », *Computational Linguistics*, vol. 17, 1991, p. 1-19.

[LAF 99] LAFOURCADE M., SANDFORD E., « Analyse et désambiguïsation lexicale par vecteurs sémantiques », *TALN'99*, Cargèse, juillet 1999, p. 351-356.

[LAF 01a] LAFOURCADE M., « Lexical sorting and lexical transfer by conceptual vectors », *First International Workshop on MultiMedia Annotation (MMA'2001)*, Tokyo, January 2001, page 6.

[LAF 01b] LAFOURCADE M., PRINCE V., « Synonymies et vecteurs conceptuels », *TALN 2001*, Tours, Juillet 2001, p. 233-242.

[LAF 02] LAFOURCADE M., PRINCE V., SCHWAB D., « Vecteurs conceptuels et structuration émergente de terminologies », *TAL*, vol. 43, n° 1, 2002, p. 43-72.

[MAN 01] MANGEOT-LEREBOURS M., « Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue », Thèse de nouveau doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble I, Septembre 2001.

[MAN 02] MANGEOT-LEREBOURS M., « An XML Markup Language Framework for Lexical Databases Environments : the Dictionary Markup Language », *LREC Workshop on International Standards of Terminology and Language Resources Management*, Las Palmas, Islas Canarias, Spain, May 2002, p. 37-44.

[MEL 84] MEL'ČUK I., ARBATCHEWSKY-JUMARIE N., ELTNISKY L., IORDANSKAJA L., LESSARD A., *DEC : Dictionnaire explicatif et combinatoire du français contemporain*,

recherches lexico-sémantiques I, Presses de l'université de Montréal, Montréal(Quebec), Canada, 1984.

[MEL 88] MEL'ČUK I., ARBATCHEWSKY-JUMARIE N., DAGENAIS L., ELNITSKY L., IORDANSKAJA L., LEFEBVRE M.-N., MANTHA S., *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques II,* Presses de l'université de Montréal, Montréal(Quebec), Canada, 1988.

[MEL 92] MEL'ČUK I., ARBATCHEWSKY-JUMARIE N., IORDANSKAJA L., MANTHA S., *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques III,* Presses de l'université de Montréal, Montréal(Quebec), Canada, 1992.

[MEL 96] MEL'ČUK I., ARBATCHEWSKY-JUMARIE N., IORDANSKAJA L., MANTHA S., POLGUÈRE A., *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques IV,* Presses de l'université de Montréal, Montréal(Quebec), Canada, 1996.

[MEL 01] MEL'ČUK I., WANNER L., « Towards a Lexicographic Approach to Lexical Transfer in Machine Translation (Illustrated by the GermanRussian Language Pair) », *Machine Translation*, vol. 16, n° 1, 2001, p. 21-87, Kluwer Academic Publishers.

[PLO 98] PLOUX S., VICTORRI B., « Construction d'espaces sémantiques á l'aide de dictionnaires de synonymes », *TAL*, vol. 39, n° 1, 1998, p. 161-182.

[POL 00] POLGUÈRE A., « Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French », *Proceeding of EURALEX'2000, Stuttgart*, 2000, p. 517-527.

[POL 02] POLGUÈRE A., « Notions de base en lexicologie », OLST-Département de linguistique et de traduction, Université de Montréal, 2002.

[QUA 01] QUAH C. K., BOND F., YAMAZAKI T., « Design and Construction of a machine-tractable Malay-English Lexicon », *Proceedings of AsiaLex*, Seoul, 2001, p. 200-205.

[RES 95] RESNIK P., « Using Information contents to evaluate semantic similarity in a taxonomy », *IJCAI-95*, 1995.

[RIL 95] RILOFF E., SHEPHERD J., « A corpus-based bootstrapping algorithm for Semi-Automated semantic lexicon construction », *Natural Language Engineering*, vol. 5, n° 2, 1995, p. 147-156.

[SAL 68] SALTON G., *Automatic Information Organisation and Retrieval*, McGraw-Hill, New York, 1968.

[SCH 01] SCHWAB D., « Vecteurs conceptuels et fonctions lexicales : application à l'antonymie », Mémoire de DEA Informatique, LIRMM, Montpellier, 2001.

[SCH 02] SCHWAB D., LAFOURCADE M., PRINCE V., « Amélioration de la représentation sémantique lexicale par les vecteurs conceptuels : le rôle de l'antonymie », *JADT 2002, vol. 2*, 2002, p. 701-712.

[SÉR 94a] SÉRASSET G., « Approche oecuménique au problème du codage des structures linguistiques », BLACHE P., Ed., *TALN-94 : Le traitement automatique du langage naturel en France aujourd'hui*, vol. 1, 7 to 8 April 1994, p. 109-118.

[SÉR 94b] SÉRASSET G., « Sublim : un système universel de bases lexicales multilingues et Nadia : sa spécialisation aux bases lexicales interlingues par acceptions », Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1, Décembre 1994.

[SPA 86] SPARCK JONES K., *Synonymy and Semantic Classification*, Edinburgh Information Technology Series, Edinburgh University Press, 1986.

[VER 01] VERLINDE S., SELVA T., « DAFA - Dictionnaire d'Apprentissage du Français des Affaires », http ://www.projetdafa.net, 2001.

[YAR 92] YAROWSKY D., « Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora », *COLING'92*, Nantes, 1992, p. 454-460.

Table des matières