Méthodes et outils pour la lexicographie bilingue en ligne : le cas du Grand Dictionnaire Estonien-Français

Antoine Chalvin, Mathieu Mangeot

MÉTHODES ET OUTILS POUR LA LEXICOGRAPHIE BILINGUE EN LIGNE : LE CAS DU GRAND DICTIONNAIRE ESTONIEN-FRANÇAIS

Antoine CHALVIN

INALCO
2, rue de Lille
F-75005 PARIS

Mathieu MANGEOT

Condillac - LISTIC - Université de Savoie
Campus Scientifique
F-73376 LE BOURGET DU LAC
CEDEX

RÉSUMÉ

Le projet de construction du Grand dictionnaire estonien-français (GDEF), du fait de sa spécificité — une équipe rédactionnelle dispersée —, a immédiatement ressenti la nécessité d'utiliser des méthodes informatiques innovantes permettant le travail à distance et en réseau. Les initiateurs de ce projet ont donc tout naturellement décidé d'utiliser une plate-forme générique de construction de dictionnaires en ligne : la plate-forme Jibiki, fruit de recherches en lexicographie computationnelle.
Après avoir exposé les conditions générales dans lesquelles s'inscrit ce projet de lexicographie bilingue en ligne (nécessité d'un tel dictionnaire, travail à distance, structure complexe, bases de données lexicales utilisées), l'article explique les méthodes de travail mises en œuvre dans ce cadre (protocole de rédaction en trois étapes) et les solutions informatiques qui les rendent possibles (interface de rédaction en ligne, gestion des contributions, import-export de données, outils annexes).

Méthodes et outils pour la lexicographie bilingue en ligne : le cas du Grand dictionnaire estonien-français

L'objet de la présente communication est de décrire les méthodes de travail mises en œuvre dans le cadre d'un projet de lexicographie bilingue en ligne, le Grand dictionnaire estonien-français (GDEF)[1], ainsi que les solutions informatiques qui les rendent possibles.

1. Exposé du problème

1.1. Nécessité d'un nouveau dictionnaire estonien-français

L'adhésion de l'Estonie à l'Union européenne et le développement de l'enseignement du français dans ce pays font apparaître comme une urgente nécessité la réalisation d'un nouveau dictionnaire estonien-français. Il n'existe en effet aucun grand dictionnaire estonien-français répondant aux exigence de la lexicographie moderne et reflétant l'état actuel de la langue estonienne, dont le lexique a considérablement évolué au cours des quinze dernières années. Le dernier dictionnaire d'une certaine ampleur (Kann et Kaplinski 1979), réalisé à l'époque soviétique, est aujourd'hui en grande partie périmé. Il est en outre entaché de graves défauts de conception, auxquels il ne semble pas possible de remédier par une simple mise à jour.

1.2. Une équipe rédactionnelle bilingue et dispersée

Une équipe bilingue de linguistes et de traducteurs a donc décidé en 2002 de se lancer dans la rédaction d'un Grand dictionnaire estonien-français d'environ 80 000 entrées. L'un des problèmes principaux auquel le projet s'est trouvé confronté, du fait de son caractère binational, a été celui de la dispersion géographique des rédacteurs entre trois villes (Paris, Tallinn, Tartu) et deux pays (la France et l'Estonie).

1.3. Travail à distance

Cette dispersion imposait de définir des méthodes et d'élaborer des outils permettant de travailler entièrement à distance. Il était donc nécessaire, avant le début de la rédaction proprement dite, de mettre en place un environnement de travail virtuel permettant : 1) de saisir les articles au format

[1] Dictionnaire réalisé par l'Association franco-estonienne de lexicographie, sous la responsabilité d'Antoine Chalvin (directeur scientifique), Madis Jürviste, Indrek Koff et Jean Pascal Ollivry, avec le soutien de l'Agence intergouvernementale de la Francophonie, de la Fondation Robert Schuman et du Centre culturel français de Tallinn.

XML, 2) de consulter la base de données au fur et à mesure de sa constitution, 3) d'organiser les différentes étapes et les différents aspects du travail rédactionnel (attribution des articles, suivi de l'avancement du travail, révision, validation, discussions entre les rédacteurs, etc.)

1.4. Une structure d'article complexe

La réalisation d'une interface d'édition était rendue plus difficile par la relative complexité de la structure d'article (en raison du degré de détail visé par le dictionnaire) : celle-ci comporte en effet jusqu'à cinq niveaux de blocs emboîtés (blocs grammaticaux, sémantiques, sous-blocs sémantiques, blocs contextuels, équivalents) et plusieurs éléments susceptibles d'avoir un nombre d'occurrences infini, ce qui rendait impossible l'utilisation de formulaires HTML classiques.

1.5. Plate-forme de construction

Une solution informatique satisfaisante à l'ensemble de ces problèmes a pu être trouvée grâce à une instance de la plate-forme Jibiki, complétée et adaptée progressivement aux besoins spécifiques du projet entre décembre 2003 et juillet 2005. Cette plate-forme, conçue au départ pour le projet Papillon (Mangeot et al. 2003), est un environnement générique en ligne permettant la rédaction et la consultation de tous types de dictionnaires : glossaires terminologiques, dictionnaires bilingues, bases lexicales multilingues, etc. Elle a été développée principalement à partir de 2001 par Mathieu Mangeot (Université de Savoie) et Gilles Sérasset (Université de Grenoble 1), notamment grâce à des recherches (Mangeot 2001) menées au sein du laboratoire GETA-CLIPS de Grenoble. La plate-forme est implantée en Java à l'aide d'outils en source libre. Elle est basée sur Enhydra, un serveur Web d'objets dynamiques en Java, et Postgres, une base de données relationnelle.

2. Base lexicale

2.1. Structure de la base lexicale

Les données lexicales du GDEF sont stockées dans deux bases distinctes : une base estonienne et une base française. La première contient l'essentiel des informations figurant dans les articles du dictionnaire, à l'exception des équivalents français et de leurs informations grammaticales « permanentes », non dépendantes du contexte d'emploi (genre des substantifs, pluriels irréguliers des substantifs et des adjectifs, féminins irréguliers des adjectifs, *h* aspiré, etc.). L'insertion dans un article du dictionnaire d'un équivalent français et l'affichage des informations le concernant se font

au moyen d'un lien reliant un élément XML de l'article de la base estonienne à un article de la base française. Cette technique permet une gestion centralisée des diverses informations concernant les mots français, ce qui permet d'éviter des erreurs, de corriger immédiatement dans l'ensemble du dictionnaire les données fautives, et d'ajouter ultérieurement des informations non prévues ou non disponibles au départ (classe de conjugaison des verbes, auxiliaire utilisé à la conjugaison active, prononciation).

2.2. Données lexicographiques de départ

Ces deux bases de données ont été constituées en extrayant automatiquement les informations pertinentes de plusieurs bases lexicales existantes.

Pour l'estonien, nous avons utilisé le grand dictionnaire estonien-russe (EVS) en cours de rédaction à l'Institut de la langue estonienne de Tallinn (3 volumes publiés, de A à P, achèvement prévu en 2007). À partir de la version XML de ce dictionnaire, nous avons extrait les mots vedettes et leurs informations morphologiques, les subdivisions sémantiques de chaque article avec les indications sémantiques correspondantes, les indications de domaine de spécialité et de registre, les exemples et les locutions. Nous disposons ainsi d'une nomenclature de 40 000 mots estoniens (de A à P) et d'une première version de chaque article, que les rédacteurs n'ont plus qu'à adapter et à compléter avec les informations françaises.

Environ 6 330 articles estoniens ont en outre été pourvus d'une indication de fréquence tirée du Dictionnaire des fréquences de l'estonien écrit (Kaalep et Muischnek 2002). Cette information nous permet de rédiger en priorité les articles sur les mots les plus fréquents, afin de maximiser l'utilité du dictionnaire en cours d'élaboration qui est immédiatement consultable sur Internet.

La base de données française a été extraite de Morphalou, dictionnaire des formes fléchies du français élaboré par l'ATILF et comprenant environ 66 000 lemmes à catégorie grammaticale unique (caractéristique qui rendait cette base parfaitement adaptée à l'usage que nous voulions en faire). Nous en avons extrait, outre les lemmes, la catégorie grammaticale, le genre des substantifs, ainsi que les pluriels et féminins « irréguliers » (selon nos critères). L'ampleur de la base française ainsi constituée permet d'ores et déjà un fonctionnement satisfaisant du système d'établissement des liens, même s'il nous reste encore à ajouter les données absentes de Morphalou (*h* aspiré, classe de conjugaison, mots composés, etc.).

3. Méthodes de travail et solutions informatiques

3.1. Procédure de rédaction

Le travail de rédaction se déroule par groupes, chacun d'eux rassemblant au moins un rédacteur estonien et un rédacteur français, qui sont amenés tous les deux à intervenir sur l'article. La réalisation d'un article comporte trois étapes. Dans un premier temps, le rédacteur — Français ou Estonien — contrôle la pertinence de la structure sémantique proposée par la base de données, insère les équivalents français en spécifiant éventuellement leurs contextes d'emploi, leur registre et leurs rections, il adapte au besoin le choix d'exemples et de locutions proposé et en fournit une traduction. Il peut consulter son coéquipier locuteur natif de l'autre langue. Il doit en outre tester les équivalents sur des exemples relevés dans deux corpus estoniens en ligne (corpus de l'estonien écrit de l'Université de Tartu et corpus de l'Institut de la langue estonienne) ainsi que sur Google. Dans une deuxième étape, l'article est révisé par un autre membre du groupe. Tout article rédigé par un Français est révisé par un Estonien et vice-versa. La nature de la révision dépend évidemment de la langue maternelle du réviseur. Un Estonien vérifiera surtout la pertinence des distinctions sémantiques, la correction des indications sémantiques et contextuelles rédigées en estonien par le rédacteur français, la pertinence des exemples et des locutions. Un réviseur français s'attachera surtout à vérifier la pertinence des équivalents et des indications concernant leurs contextes d'emploi. Une fois révisé, l'article est soumis à une ultime vérification, opérée par l'un des deux validateurs français. Le nombre de validateurs est volontairement limité afin de garantir un minimum de cohérence dans la présentation des articles. Une fois validé, l'article devient accessible à la consultation par tous les visiteurs du site.

3.2. Interface d'édition

L'interface d'édition (formulaire de saisie) permet d'éditer des structures relativement complexes grâce à des outils de gestion de listes. L'ajout et la suppression de champs ou de blocs de champs dans le formulaire se font très simplement, au moyen d'un bouton à cliquer. Il est également possible de modifier l'ordre des blocs au sein de certaines listes de blocs. Un module spécifique permet d'établir les liens vers la base française : lorsque le rédacteur saisit un équivalent français dans la fenêtre surgissante de ce module, le système recherche les articles correspondants. S'il n'en trouve qu'un seul, il place automatiquement le lien dans le champ adéquat du formulaire. S'il en

trouve plusieurs, il affiche un résumé des différents articles (avec leur catégorie grammaticale et leur genre) et propose au rédacteur de choisir celui vers lequel il veut établir le lien. S'il n'en trouve aucun, il propose au rédacteur de créer lui-même l'article français. Lors de la sauvegarde, les informations saisies dans le formulaire sont converties automatiquement au format XML.

3.3. Gestion des contributions

Pour gérer le travail de rédaction-révision-validation, la plate-forme permet de définir des groupes et des droits d'accès. Il existe trois groupes prédéfinis (réviseurs, validateurs, administrateurs) et il est possible de créer un nombre illimité de groupes de travail. Si l'utilisateur n'est pas logué, il n'appartient à aucun groupe. Il peut simplement consulter les ressources disponibles sur la plate-forme. Lorsqu'il est logué, il accède à l'interface d'édition des articles. Les réviseurs peuvent modifier et réviser les contributions des autres membres de leur groupe de travail. Les validateurs peuvent modifier toutes les contributions et valider les contributions révisées. Les administrateurs peuvent gérer les utilisateurs et les groupes, ajouter de nouvelles ressources sur la plate-forme, etc.

Les données sont gérées de la façon suivante : les données lexicographiques, constituant l'article proprement dit, sont encapsulées dans une *contribution*, qui contient également un certain nombre de méta-informations sur l'article : auteur, réviseur, validateur, dates de création, révision, validation, ainsi qu'un historique de toutes les modifications.

La contribution comporte un statut, qui évolue pendant le cycle de rédaction en passant successivement par quatre états : « non-finie », « finie », « révisée », « validée ». Une contribution validée est en principe définitive, mais il est possible de la modifier en en faisant une copie, qui repasse alors par toutes les étapes du processus de rédaction. Une fois cette copie validée, ses données remplacent celles de la contribution d'origine.

4. Gestion et suivi du travail

4.1. Attribution des articles

L'attribution des articles aux rédacteurs se fait au moyen d'une interface spécifique accessible aux validateurs. Cette interface permet d'attribuer des paquets d'articles constitués selon différents critères combinables entre eux. Il est par exemple possible d'attribuer des tranches alphabétiques avec ou sans prise en compte de la fréquence, des groupes de termes relevant d'un même domaine

de spécialité, des séries de mots composés comportant le même élément final, de réattribuer tous les articles non finis d'un auteur donné à un autre auteur, etc. La liste complète des articles concernés est affichée pour vérification avant l'attribution.

4.2. Tableau résumé

Pour permettre le suivi de l'avancement du travail et le calcul de la rémunération des rédacteurs, il est possible d'obtenir un tableau résumé de toutes les contributions finies, révisées et validées sur une période donnée. Cette fonctionnalité est accessible à tous les membres permanents de l'équipe, ce qui leur permet notamment de comparer leur productivité à celle des autres rédacteurs.

4.3. Exportation et importation des données

La plate-forme dispose d'une fonction d'exportation, qui peut être utilisée par exemple pour la relecture à l'écran ou sur papier de groupes d'articles. Il est possible de combiner plusieurs critères de recherche pour constituer la liste des articles à exporter. L'utilisateur peut aussi définir le format d'exportation : XML, HTML, format texte et format PDF pour l'impression.

Il est également possible d'importer des articles, à condition que ceux-ci soient au format XML. Cette fonctionnalité pourrait permettre aux lexicographes qui le désirent de travailler en local puis de réimporter leurs contributions une fois le travail terminé et lorsqu'une connexion à l'Internet est accessible. En pratique, nous l'utilisons surtout pour ramener des contributions à un statut antérieur (par exemple dans le cas d'une contribution validée par erreur) ou lorsque le format des données doit être modifié, soit pour ajouter de nouvelles informations, soit pour raffiner certaines parties de la structure. Dans ce cas, les contributions concernées sont exportées au format XML, modifiées, puis réimportées sur la plate-forme.

5. Outils annexes

5.1 Forum

Pour faciliter la communication entre les rédacteurs et avec le public, un forum a été installé sur le serveur du projet. Il est constitué de deux parties : l'une réservée aux membres du projet et l'autre d'accès public. La partie réservée aux membres comprend plusieurs sous-forums : sur l'un figure la dernière version du protocole de rédaction, un autre est réservé à la discussion sur les problèmes

rédactionnels généraux non encore couverts par le protocole, un troisième permet de discuter des futurs développements de la plate-forme, un quatrième est consacré au partage de liens thématiques vers des glossaires ou autres ressources en lignes, et un cinquième, organisé selon l'ordre alphabétique estonien, permet de discuter de problèmes concernant des mots précis.

Bibliographie

A. Dictionnaires

Grand dictionnaire estonien français (GDEF) : http://www.estfra.ee/

Eesti-vene sõnaraamat [dictionnaire estonien-russe] I-III. Tallinn. Eesti Keele Instituut. 1997-2004.

Kaalep, H-J., Muischnek, K. (2002) *Eesti kirjakeele sagedussõnastik* [dictionnaire des fréquences de l'estonien écrit]. Tartu, Tartu Ülikooli kirjastus.

Morphalou, lexique morphologique ouvert du français : http://actarus.atilf.fr/morphalou/

B. Autres

Corpus de l'Institut de la langue estonienne (Tallinn) : http://www.eki.ee/corpus/

Corpus de l'estonien écrit (Université de Tartu) : http://www.cl.ut.ee/korpused/kasutajaliides/index.html.en

Mangeot, Mathieu (2001) *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse de doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble I, 2001, 280 p.

Mangeot, Mathieu ; Thevenin, David (2004) 'Online Generic Editing of Heterogeneous Dictionary Entries in Papillon Project.' in Proc. COLING 2004, ISSCO, Université de Genève, 23-27 August 2004, vol 2/2, 1029-1035.

Mangeot, Mathieu ; Sérasset, Gilles ; Lafourcade, Mathieu (2003) 'Construction collaborative de données lexicales multilingues, le projet Papillon.' *Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux ?* ed. M. Zock and J. Carroll, *TAL Traitement Automatique des Langues*, Vol. 44 : 2/2003, 151-176.