

L'analyseur syntaxique Fips

Eric Wehrli, Luka Nerima
LATL-Département de linguistique
Université de Genève
Eric.Wehrli@unige.ch, Luka.Nerima@unige.ch

Résumé

L'analyseur Fips a connu beaucoup de développements au cours de ses quelques 15 années d'existence. Ce papier décrit le fonctionnement et quelques unes des propriétés de cet analyseur multilingue et multifonction, basé sur un modèle linguistique inspiré de la grammaire générative et sur une modélisation par objets pour son implémentation. A l'heure actuelle, des versions bien développées existent pour l'anglais, l'allemand, l'espagnol, le français et l'italien. Un accent tout particulier est mis dans cette présentation sur le formalisme utilisé pour l'expression des règles d'attachement, sur le processus de chaînage utilisé, entre autres, pour l'interprétation des syntagmes extraposés, sur le traitement des collocations ainsi que sur la structure et le contenu de la base de données lexicales.

1 Introduction

Depuis près d'une vingtaine d'années, le LATL travaille au développement d'un modèle d'analyse syntaxique. Connu sous le nom de Fips, ce modèle d'analyseur a subi de nombreux développements et modifications. Entre la version de cet analyseur, présentée en 1991 à l'ATALA (cf. Laenzlinger et Wehrli, 1991) et la version actuelle, pratiquement rien en dehors du nom, de l'objectif général et de certains présupposés linguistiques n'a été conservé. Les concepts sur lesquels repose l'analyseur Fips actuel, la stratégie d'analyse, la plateforme et le langage de développement ont changé, parfois à plusieurs reprises. Certains de ces changements ont été effectués rapidement et en une fois; c'est le cas bien sûr de la réécriture complète du système lorsque le LATL a abandonné

l'environnement OpenVMS pour passer à MS-Windows, il y a une douzaine d'années, ou encore lorsque le langage de programmation Component Pascal (une version d'Oberon-2, développée par Oberon Microsystems¹) a été choisi en remplacement de Modula-2. D'autres modifications sont intervenues de manière progressive, souvent à la suite de problèmes rencontrés. Un exemple de ce dernier cas de figure est le traitement des arguments, comme les compléments verbaux, lorsque nous avons tenté d'appliquer le modèle développé sur la base d'un ordre des mots relativement rigide (français, anglais) à des langues à ordre de mots très libre (allemand, grec moderne).

C'est bien sûr la version actuelle de Fips que nous allons décrire dans les pages ci-dessous. Fips, version 2009, est un analyseur multilingue. A ce jour, le modèle a été appliqué de manière extensive au français, à l'anglais, à l'espagnol, à l'italien et à l'allemand. Plusieurs autres langues sont en chantier, à des stades de développement divers. Dans l'ordre décroissant de développement, nous avons le grec moderne, le roumain, le russe, le japonais et le romanche. Rappelons également que l'analyseur Fips n'a pas été défini pour une application particulière, mais au contraire se veut "universel", c'est-à-dire utilisable pour toute application qui nécessite une analyse linguistique. A ce jour, l'analyseur Fips a été utilisé pour la traduction automatique (Wehrli et al. 2009), l'analyse de la parole (Gaudinat et al. 1999), la synthèse de la parole (Goldman et al. 2000), le tagging (Scherrer, 2008), l'extraction terminologique (Seretan, 2008, Seretan et Wehrli, 2008), le résumé automatique (Genest et al. 2008), la correction orthographique (L'haire, 2004), l'extraction d'infor-

¹Voir www.oberon.ch.

mations (Ruch, 2002) et l'assistance terminologique (Wehrli, 2004). Il faut souligner que pour toutes ces applications, l'analyseur, ainsi que les ressources lexicales qu'il utilise sont identiques. Seuls les modules d'entrée et de sortie sont spécifiques.

2 Fondements linguistiques

Au plan linguistique, l'approche sous-jacente repose sur une adaptation du modèle chomskyen "minimaliste" (Chomsky, 1995), avec de nombreux emprunts à d'autres modèles génératifs, tels que la grammaire lexicale fonctionnelle (LFG) (Bresnan, 2001) ou le modèle de *Simple Syntax* de Culicover & Jackendoff (2005). Notre interprétation, parfois très personnelle, de ces modèles va dans le sens d'une simplification maximale des représentations syntaxiques, aussi bien par souci d'efficacité de l'implémentation du modèle que par conviction linguistique.

On peut considérer que la grammaire est de type lexicaliste, basée sur un lexique riche en informations syntaxiques, spécifiant en particulier les propriétés sélectionnelles d'éléments fonctionnels tels que les prépositions, les auxiliaires, les déterminants, etc. Par exemple, en français, les auxiliaires *avoir* et *être* sélectionnent un élément verbal [+participePassé], alors qu'en allemand, l'auxiliaire *werden* sélectionne un verbe [+infinitif]. Les arguments sélectionnés par des têtes prédicatives (noms, adjectifs, mais surtout verbes) sont également spécifiés dans l'entrée lexicale des items concernés. D'autres renseignements, y compris de nature sémantique, figurent également dans la base de données lexicales (cf. section 4).

Les structures syntaxiques construites par l'analyseur Fips obéissent toutes au schéma (1), où **L** correspond au sous-arbre gauche et **R** au sous-arbre droit de la tête lexicale **X**.

$$(1) [\underset{XP}{\quad} L X R]$$

Dans ce schéma, **X** est une variable dont les valeurs possibles sont les catégories lexicales usuelles, **Adv** (adverbe), **A** (adjectif), **N** (nom), **D** (déterminant), **V** (verbe), **P** (préposition), **C** (conjonction), **Inter** (interjection). A cette liste, nous ajoutons les catégories fonctionnelles **Temps** (tête de TP) et **Fonction** (tête des structures fonctionnelles FP utilisées pour représenter les structures prédicatives de type adjectival, nominal, etc.).

Le schéma (1) correspond à une variante minimale de la théorie X-barre, limitée à deux niveaux : **X**, la tête de la projection, et **XP**, la projection maximale.

3 Fonctionnement de l'analyseur

Dans les grandes lignes, l'analyseur fonctionne comme suit : balayant la chaîne d'entrée de gauche à droite, sans retour en arrière, l'analyseur lit un mot de catégorie **X**, projette un syntagme de catégorie **XP**, dont le mot lu constitue la tête lexicale, puis combine le constituant **XP** avec l'analyse en cours en fonction des règles de la grammaire. Le non-déterminisme inhérent aux langues naturelles (un mot peut appartenir à plus d'une partie du discours, un syntagme peut s'attacher à plus d'un nœud de la structure syntaxique, etc.) est traité de manière (pseudo)-parallèle. L'analyseur calcule toutes les solutions possibles, qui sont ordonnées en fonction de critères en partie psycholinguistiques (préférences d'attachement) et en partie statistiques (p. ex. fréquence des lexèmes).

L'analyse aboutit si, après avoir parcouru toute la chaîne d'entrée, on a au moins une structure couvrant toute la séquence. On peut alors afficher une ou plusieurs des analyses obtenues. En cas d'échec d'analyse globale, l'analyseur retourne un message d'échec et une structure constituée de la concaténation d'analyses partielles.

Les trois mécanismes fondamentaux utilisés par l'analyseur sont (i) le mécanisme de projection, (ii) la combinaison des constituants et (iii) le "déplacement" de constituants (en anglais, *project, merge et move*).

3.1 Projection

Le mécanisme de projection crée ("projette") une structure syntaxique sur la base soit d'une structure lexicale, soit d'une autre structure syntaxique. Le cas "normal", non-marqué, est celui de la projection d'une structure syntaxique à partir d'un élément lexical. Cette opération prend place à l'intersection des composantes lexicales et syntaxiques. Tout item lexical issu de l'analyse lexicale donne lieu à une projection maximale de même catégorie, avec l'item lexical pour tête. Le mécanisme de projection est également invoqué dans d'autres cas, où il permet de créer des projections syntaxiques à partir d'autres projections, ou alors de créer des projections complexes à partir d'éléments lexicaux. Ce mécanisme étendu de

- a. AgreeWith(b, {number, gender})
- b. **VP + DP**
 - a. HasFeature(mainVerb)
 - b. IsArgumentOf(a, directObject)
- c. **NP + AP**
 - b. HasFeature(postNominalAdj)
 - a. AgreeWith(b, {number, gender})

Considérons, à titre d'exemple, la première règle de (3), qui concerne l'attachement d'un adjectif prénominal à un nom. Cette règle stipule qu'une projection de catégorie AP peut s'attacher comme sous-constituant gauche d'une projection de type NP. Deux conditions sont associées à cette règle. La première `a.HasFeature(prenominal)` exige que le premier terme, l'adjectif, porte le trait [+prénominal]. Quant à la deuxième condition, `a.AgreeWith(b, {number, gender})`, elle impose une règle d'accord en nombre et en genre entre les deux constituants. Les variables **a** et **b** utilisées dans l'expression des conditions réfèrent respectivement au premier et au deuxième constituant de la règle.

La deuxième règle de (3) concerne l'attachement d'un sujet à une structure de prédicat, dans notre terminologie, l'attachement d'un DP comme sous-constituant gauche d'une projection TP. Trois conditions sont imposées à cet attachement. La première condition dit que le verbe doit être un verbe principal (et non un auxiliaire)³. La deuxième condition impose un accord en nombre, genre et personne entre ces deux constituants (l'accord sujet-verbe) alors que la troisième condition stipule que **a**, le premier constituant, doit pouvoir être interprété comme argument sujet de **b**.

3.3 Move

Bien que l'architecture générale des structures de surface assignées aux phrases analysées résulte de l'interaction des opérations de projection et de combinaison, un mécanisme additionnel est requis pour rendre compte de certaines conditions de bonne formation syntaxique, comme celles imposées par la théorie thématique. Cette dernière impose, en gros, que tout syntagme nominal (autre

³L'attachement d'un sujet à un groupe verbal comprenant un auxiliaire nécessite une règle un peu plus complexe.

qu'adverbial), soit associé à un rôle thématique distribué par un prédicat sous condition de gouvernement. Dans la théorie chomskyenne, les éléments extraposés sont des éléments déplacés par une transformation de mouvement à partir d'une position dite canonique, gouvernée par un prédicat. C'est par l'intermédiaire de cette position canonique, à laquelle il reste lié, qu'un syntagme nominal extraposé reçoit son rôle thématique.

Une analyse très semblable a été adoptée pour l'analyseur Fips, qui associe à un élément extraposé une catégorie vide en position canonique d'argument (p. ex. position sujet ou position complément d'un prédicat). Par analogie avec la transformation dont il s'inspire, ce mécanisme est appelé **move**.

Pour illustrer le fonctionnement de ce mécanisme, considérons une phrase simple comme (5a). L'analyse de cette phrase est donnée en (5b) et appelle quelques commentaires. Tout d'abord, on observe que la locution interrogative *est-ce que* est considéré comme une sorte de conjonction et apparaît comme tête du constituant phrase CP. Le sous-constituant gauche de cette structure contient le syntagme interrogatif *qui*. N'étant gouverné par aucun prédicat, le syntagme interrogatif doit être associé à une position canonique d'argument du verbe *inviter*. Le mécanisme d'interprétation des éléments extraposés crée un syntagme abstrait (vide), noté $[_{DP} e]$, en position d'objet direct du verbe et le lie à l'élément extraposé. Ce lien, aussi appelé chaînage, est représenté dans notre structure par l'indice **i**.

- (5)a. qui est-ce que Paul a invité ?
- b. $[_{CP} [_{DP} \text{qui}]_i \text{ est-ce que } [_{TP} [_{DP} \text{Paul}] \text{ a } [_{VP} \text{ invité } [_{DP} e_i]]]]]$

3.4 Identification des collocations

L'importance d'un traitement adéquat des collocations en TAL est maintenant largement reconnue. Dans notre optique, c'est à l'analyseur syntaxique – et non à un pré-traitement spécifique – qu'il appartient d'identifier les unités lexicales complexes, dont les collocations constituent un exemple, à côté des mots composés et des expressions idiomatiques (Wehrli, 2000). Cette optique est cohérente avec le fait que les collocations associent deux termes qui entretiennent une relation syntaxique étroite, soit de modification (p. ex. adjectif modifieur de nom), soit de complément (p. ex. objet direct ou objet prépositional

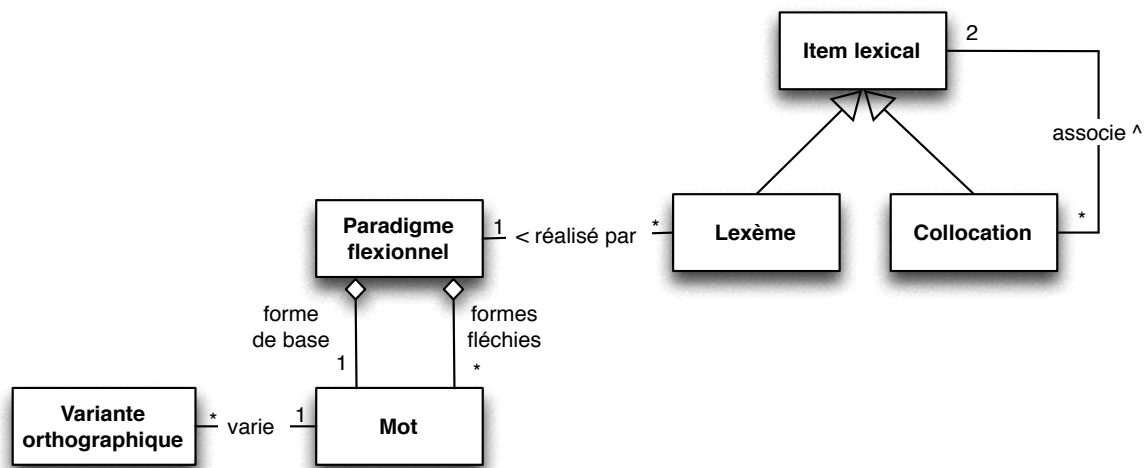


FIG. 1 – Structure de la base de données lexicales (sous forme de classes UML)

le lexique des variantes orthographiques. Pour être complet, il faut encore signaler la possibilité de définir des lexiques privés, destinés à recevoir des mots de domaines spécialisés (médecine, biologie, chimie, etc.) ou des noms propres (lieux géographiques, patronymes, etc.). Pour ne pas "polluer" les lexiques généraux du français, ces lexiques privés constituent une structure lexicale séparée.

Avant de décrire plus en détail les informations contenues dans chacun des lexiques, nous donnons dans la Figure 1 la structure de la base de données lexicales pour le français⁵.

4.1 Le lexique des mots

Une entrée de type *mot* contient l'information décrivant la forme orthographique, comme la catégorie lexicale (valeurs habituelles des parties de discours, *Nom*, *Verbe*, *Adjectif*, etc.), les traits d'accord (nombre, genre, personne, mode et temps verbal...). Comme les lexiques ont été conçus pour d'autres applications que l'analyseur (par exemple la synthèse vocale), l'entrée contient également la représentation phonétique (y compris l'éventuelle consonne latente). Des indices de fréquence d'utilisation des mots dans les corpus sont aussi stockés pour chaque entrée, utilisés par l'analyse Fips pour des traitements heuristiques.

L'insertion des mots dans le lexique se fait manuellement⁶; le lexicographe entre la forme de

base (le mot au singulier pour les noms, l'infinif pour les verbes, etc.), la catégorie lexicale et la classe de flexion. C'est là qu'intervient le générateur morphologique : il génère et insère tous les mots du paradigme flexionnel.

Comme toutes les réalisations possibles des mots se trouvent ainsi dans le dictionnaire, il n'est pas nécessaire lors de l'analyse d'appliquer des règles morphologiques (exception faite pour les mots inconnus) pour calculer la forme de base du mot, l'analyse lexicale se trouve ainsi réduite à la recherche d'un mot dans le dictionnaire. Cette manière de procéder s'avère évidemment efficace en terme de temps de traitement.

4.2 Le lexique des lexèmes

Une entrée de type *lexème* décrit les propriétés syntaxiques (sous-type de la catégorie lexicale, traits sélectionnels, structure argumentale, etc.) et sémantiques (traits sémantiques, télicité des verbes, rôles thématiques des arguments...) du lexème. Tout comme pour les mots, des indices de fréquences d'utilisation du lexème figurent dans l'entrée.

Chaque lexème est associé au paradigme flexionnel qui le réalise et qui est stocké dans le lexique des mots. Si plusieurs lectures syntaxiques (ou sémantiques) existent pour un mot, elles donneront lieu à autant d'entrées de lexème. Par exemple le verbe *manger* sera représenté par deux entrées, l'une transitive, l'autre intransitive.

⁵Bien que nous décrivions ici la base de données lexicales du français, précisons que les lexiques des autres langues que nous traitons sont tous structurés de la même manière.

⁶Si des ressources sont disponibles, l'entrée se fait de manière semi-automatique : pour l'espagnol nous avons exploité

un corpus étiqueté et entré ainsi environ 15'000 mots. Le lexicographe a cependant dû valider ces entrées.

4.3 Le lexique des collocations

Une entrée de type *collocation* contient :

- la configuration syntaxique de l'expression, appelée type de la collocation (*nom + adjectif*, *nom + nom*, *nom + préposition + nom*, *sujet + verbe*, *verbe + objet*, etc.) ;
- la référence aux deux items lexicaux composant l'expression (base + collocatif). Il faut noter ici que l'item lexical peut aussi bien être de type *lexème* que *collocation* : s'il est de type *collocation*, on est en présence d'une définition récursive de la collocation. C'est de cette manière que nous avons choisi de représenter les collocations de plus de 2 termes, comme par exemple *tomber en panne sèche* ou *travail à temps partiel* ;
- la préposition s'il y a lieu (p. ex. *salle de conférence* ;
- les traits de figement (collocation plurielle, complément sans déterminant, complément figé, etc.).

Exemples d'entrées :

(10)a. *prendre rendez-vous*

type : verbe - objet direct

lexème n°1 : *prendre*, verbe transitif

lexème n°2 : *rendez-vous*, nom commun

préposition : Ø

traits de figement : { }

(11)a. *miroir aux alouettes*

type : nom - préposition - nom

lexème n°1 : *miroir*, nom commun

lexème n°2 : *alouette*, nom commun

préposition : à

traits de figement : {compl. avec déterminant, compl. pluriel}

L'interface du système de saisie des collocations effectuée par l'intermédiaire de Fips une analyse syntaxique complète de l'expression entrée par l'utilisateur et détermine quelles sont les unités lexicales qui la composent, ainsi que le type de la collocation et les traits de figement. C'est ensuite au lexicographe de valider ou de modifier ces paramètres.

Un outil d'extraction automatique de collocation a été également développé au LATL (Seretan et al., 2004 ; Seretan, 2008). Il génère des listes de collocations à partir de corpus choisis et fournit les contextes d'utilisation. Le lexicographe peut s'aider de cet outil en parcourant la liste des colloca-

tions et en validant celles qui lui paraissent pertinentes.

4.4 Le lexique des variantes orthographiques

Les entrées du lexique des variantes orthographiques contiennent la graphie de la variante, le type de la variante (graphie rectifiée par exemple) et le principe de la variation s'il y a lieu. Elles permettent lors de l'analyse de traiter les variantes de la même manière que les formes préférées stockées dans le lexique des mots (et éviter ainsi les mots inconnus). Nous avons aussi utilisé la notion de variante pour traiter les abréviations. A noter qu'un mot peut avoir plusieurs variantes alors qu'une variante est associée à exactement un mot.

Dans une application future, on envisage de laisser le soin à l'utilisateur de choisir le type d'orthographe auquel se conformer en génération, par exemple l'orthographe réformée.

5 Quelques chiffres

En guise de conclusion, donnons quelques estimations de la taille des lexiques et des performances obtenues par Fips. Pour les langues relativement bien développées, les lexiques comptent entre 30 000 (espagnol) et plus de 70 000 (anglais et français) lexèmes (le nombre de mots dépend évidemment de la richesse flexionnelle de la langue – il s'élève à plus de 400 000 pour l'allemand et 230 000 pour le français). Quant aux collocations, elles sont plus de 14 000 pour le français, 6 500 pour l'anglais, moins pour les autres langues.

Le temps de traitement dépend beaucoup de la complexité des corpus, mais varie généralement de 100 à 250 symboles (mot orthographique, ponctuation) à la seconde, ce qui signifie que l'analyse d'un corpus d'un million de mots prend approximativement entre 2 et 4 heures. Nous ne disposons pas à l'heure actuelle d'évaluation précise des résultats de l'analyseur Fips et espérons que notre participation à la campagne d'évaluation Passage comblera cette lacune. Les seules données statistiques à grande échelle sont le nombre d'analyses complètes et le nombre de mots "inconnus" (c-à-d. absents dans la base de données). Sur la base de gros corpus journalistiques (plus de 5 000 articles de l'hebdomadaire britannique *The Economist* pour l'anglais et le corpus du journal *Le Monde* pour le français), nous obtenons un peu plus de 80% d'analyses complètes. De nombreux

pointages et examens forcément partiels suggèrent que si les analyses "complètes" comportent fréquemment des erreurs au niveau des attachements de constituants (c'est en particulier le cas des attachements de syntagmes prépositionnels !), elles sont généralement très fiables quant à l'identification des unités lexicales. Pour ce qui est du nombre de mots "inconnus", il s'élève à moins de 0.2% pour l'anglais, et à moins de 0.1% pour le corpus français, sans compter les noms propres.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse de la recherche scientifique, subside no 101412-103999. Nous sommes reconnaissants à Lorenza Russo et aux autres membres du LATL pour une lecture attentive de ce texte et bien sûr pour leur contribution aux recherches décrites dans cet article.

6 Bibliographie

- Chomsky, N. 1995. *The Minimalist Program*, Cambridge, Mass., MIT Press.
- Culicover, P. et R. Jackendoff, 2005. *Simpler Syntax*, Oxford, Oxford University Press.
- Emonds, J. 1976. *A Transformational Approach to English Syntax*, New York, Academic Press.
- Gaudinat, A., J.-Ph. Goldman et E. Wehrli, 1999. "Syntax-Based Speech Recognition : How a Syntactic Parser Can Help a Recognition System", *EuroSpeech 1999*, 1587-1591.
- Genest, P.-E., G. Lapalme, L. Nerima et E. Wehrli, 2008. "Ness : a Symbolic Summarizer for the Update Task of TAC 2008", in *Actes de la conférence TAC 2008*, Gaithersburg, MD.
- Goldman J-P., A. Gaudinat et E. Wehrli, 2000. "Utilisation de l'analyse syntaxique pour la synthèse de la parole, l'enseignement des langues et l'étiquetage grammatical" *TALN 2000*, Lausanne.
- Laenzlinger, C. 2003. *Initiation à la syntaxe formelle du français*, Berne, Peter Lang.
- Laenzlinger, C. et E. Wehrli, 1991. "FIPS : Un Analyseur interactif pour le français", *TA Informations*, 32 :2, 35-49.
- L'haire, S. 2004. "Vers un feedback plus intelligent : les enseignements du projet Freetext", in *Actes de la journée d'étude de l'ATALA. TAL & Apprentissage des langues*, Grenoble, 1-12.
- Pollock, J.-Y. 1989. "Verb Movement, UG and the Structure of IP", *Linguistic Inquiry* 20, 365-425.
- Ruch, P. 2002. *Applying Natural Language Processing to Information Retrieval in Clinical Records and Biomedical Texts*, thèse de doctorat, Université de Genève.
- Y. Scherrer, 2008. "Part-of-speech tagging with a symbolic full parser : Using the TIGER treebank to evaluate Fips". Workshop on Parsing German, ACL 2008, Columbus, Ohio, USA.
- Seretan, V., Nerima, L. and E. Wehrli, 2004. "A Tool for Multi-Word Collocation Extraction and Visualization in Multilingual Corpora", in *Actes d'EURALEX 2004*, 755-766.
- Seretan, V. 2008. *Collocation Extraction Based on Syntactic Parsing*. Thèse de doctorat, Université de Genève.
- Seretan, V. and E. Wehrli, 2008. "Multilingual collocation extraction with a syntactic parser", in *Language Resources and Evaluation*. [available at www.springerlink.com]
- Wehrli, E. 2000. "Parsing and Collocations", in D. Christodoulakis (éd.) *Natural Language Processing - NLP 2000*, Springer-Verlag, 272-282.
- Wehrli, E. 2004. "Traduction, traduction de mots, traduction de phrases", in B. Bel et I. Marlien (éd.), *TALN XI*, Fes, 483-491.
- Wehrli, E. 2007. "Fips, a Deep Linguistic Multilingual Parser", in *Proceedings of the ACL 2007 Workshop on Deep Linguistic Parsing*, Prague, 120-127.
- Wehrli, E., L. Nerima et Y. Scherrer, 2009. "Deep linguistic multilingual translation and bilingual dictionaries", *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athènes, 90-94.