Traitement des inconnus : une approche systématique de l'incomplétude lexicale

Helena Blancafort[1,2], Gaëlle Recourcé[3], Javier Couto[1],
Benoît Sagot[3], Rosa Stern[3,4], Denis Teyssou[4]
(1) Syllabs, 15, rue Jean-Baptiste Berlier, 75013 Paris, France
(2) Universitat Pompeu Fabra, Roc Boronat, 08013 Barcelona, Espagne
(3) Alpage, INRIA Paris–Rocquencourt & Université Paris 7, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
(4) Agence France-Presse – Medialab, 2 place de la Bourse, 75002 Paris, France

Résumé Cet article aborde le phénomène de l'incomplétude des ressources lexicales, c'est-à-dire la problématique des inconnus, dans un contexte de traitement automatique. Nous proposons tout d'abord une définition opérationnelle de la notion d'inconnu. Nous décrivons ensuite une typologie des différentes classes d'inconnus, motivée par des considérations linguistiques et applicatives ainsi que par l'annotation des inconnus d'un petit corpus selon notre typologie. Cette typologie sera mise en œuvre et validée par l'annotation d'un corpus important de l'Agence France-Presse dans le cadre du projet EDyLex.

Abstract This paper addresses the incompleteness of lexical resources, i.e., the problem of unknown words, in the context of natural language processing. First, we put forward an operational definition of the notion of unknown words. Next, we describe a typology of the various classes of unknown words, motivated by linguistic and applicative considerations as well as the annotation of unknown words in a small-scale corpus w.r.t. our typology. This typology shall be applied and validated through the annotation of a large corpus from the Agence France-Presse as part of the EDyLex project.

Mots-clés : mots inconnus, incomplétude lexicale, acquisition dynamique des ressources lexicales

Keywords: unknown words, lexical incompleteness, dynamic acquisition of lexical information

1 Introduction

L'enrichissement des ressources lexicales nécessaires aux applications de traitement automatique du langage est une tâche souvent longue, fastidieuse et coûteuse. L'exhaustivité de ces ressources est par ailleurs un objectif complexe à atteindre. Le projet EDyLex[1] cherche à proposer des solutions à la problématique de l'incomplétude des ressources lexicales et de l'automatisation de leur enrichissement. Dans cette perspective, de premiers travaux se concentrent sur l'identification des causes de cette incomplétude. Un certain nombre de formes rencontrées dans les contenus à analyser sont en effet qualifiables d'*inconnus* en regard des ressources lexicales en jeu : ces formes ne correspondent à aucune entrée dans le lexique, ou sont inanalysables par les modules d'analyse morphosyntaxique utilisés.

[1] Enrichissement Dynamique de ressources Lexicales multilingues en contexte multimodal. Site internet du projet : http://sites.google.com/site/projetedylex/. Ce projet (2009-2012) a été financé par l'ANR (ANR-09-CORD-008).

Un certain nombre d'études, présentées à la section 2, ont été consacrées au problème des inconnus. La particularité de notre approche étant l'exploitation d'une telle étude à des fins de traitement automatique, nous proposons une définition opérationnelle de la notion d'inconnu (section 3). La typologie décrite à la section 4 établit un cadre permettant de classer les inconnus d'un corpus donné, afin d'en proposer des traitements adaptés. La section 5 expose les modalités de l'annotation, selon cette typologie, des inconnus d'un corpus de dépêches de l'Agence France-Presse, tâche validée lors d'une première expérience à petite échelle.

2 Travaux antérieurs

Différents axes de recherche traitent de la problématique des inconnus : les études linguistiques sur la néologie, d'une part, analysent les pratiques lexicographiques prenant en compte ce problème ; la linguistique de corpus, d'autre part, propose une description de ce phénomène à partir de données (corpus et lexiques) particulières (Cartoni, 2006 ; Dister et Fairon, 2004 ; Maurel, 2004, Ren et Perrault, 1992) afin d'en obtenir une typologie. Enfin, un certain nombre de travaux apportent des solutions de traitement automatique des inconnus (Nakov et al., 2003 ; Schone et Jurafsky, 2001 ; Mikheev, 1997).

Les travaux de linguistique de corpus (Cartoni, 2006 ; Dister et Fayron, 2004 ; Maurel, 2004, Ren et Perrault, 1992) présentent pour le français une classification des mots non reconnus par un lexique de référence déterminé et un corpus précis, souvent journalistique. Ces auteurs traitent souvent des *mots inconnus*, définis comme les unités lexicales qui ne sont pas répertoriées dans le lexique utilisé (Dister et Fairon 2004). L'objet d'analyse est donc ici le mot : toute autre unité typographique présente dans un corpus est ignorée. Cartoni (2008) donne une définition plus fine en précisant qu'un inconnu est une « suite de caractères qui correspond à la notion graphique d'unité lexicale » et qui est absente du lexique de référence. Il souligne également que l'on trouve parmi eux un nombre important d'hapax. Mais la question de savoir si un inconnu est nécessairement un mot simple ou si l'on peut parler de composés inconnus reste ouverte dans la littérature. Seuls Jonasson et Maurel (2004) définissent les mots inconnus comme des mots simples par opposition aux noms propres qui sont souvent des mots composés. Par ailleurs, tous ces auteurs soulignent les limites de la notion de lexique de référence, un lexique n'étant jamais une ressource exhaustive, et considèrent que la clé pour aborder les mots inconnus est de traiter la créativité lexicale en tant que telle. Ainsi, pour le français, Maurel (2004) explique que 40% des inconnus étudiés correspondent à des créations lexicales, tandis que Ren et Perrault (1992) précisent que 30% des inconnus sont des dérivés. Les résultats sont différents sur d'autres langues. Ainsi, en allemand, la plupart des mots inconnus sont des composés *ad hoc* sémantiquement transparents, formant une unité typographique unique qui n'a pas à été ajoutée au lexique (Geyken, 2009). En chinois, l'enjeu principal est celui de la segmentation. Ainsi, une mauvaise segmentation des inconnus est responsable de 60% des erreurs de segmentation en mots (Wong et al., 2009). Cela dit, la composition est aussi très productive en chinois (Tseng et al., 2005).

Concernant les travaux linguistiques sur la néologie, on peut citer Sablayrolles (1997 ; 2002), qui illustre bien la difficulté du domaine : bien qu'il y ait un consensus au niveau de la définition de la notion de néologisme, on observe un désaccord dès lors qu'il s'agit de les catégoriser. Sablayrolles a comparé plus de 100 typologies des néologismes, montrant ainsi la difficulté de la classification de ceux-ci, et même de définir ce qui est un néologisme et ce qui n'en est pas un.

3 Qu'est-ce qu'un inconnu ?

Comme le montre la problématique des inconnus composés évoquée à la section précédente, la question de la nature même de ce que l'on considère comme connu ou inconnu fait débat. Avant de prendre position sur ce point, il est important de poser deux définitions. En cohérence avec la norme ISO MAF (de la Clergerie et Clément, 2005) et de nombreux travaux (Vilnat et al., 2008), nous définissons comme suit les notions de token et de forme. Un *token* est une unité typographique constituée d'un caractère de ponctuation ou d'une séquence de caractères (sauf l'espace) délimitée par des espaces et/ou des caractères de ponctuation[2].. Ainsi, « gratin aux pommes» contient trois tokens. Par contraste, une *forme* (en anglais *wordform*) est une unité lexicale susceptible de constituer une entrée dans le lexique de référence. La correspondance entre tokens et formes n'est pas bijective : le token « aux » correspond aux deux formes *à* et *les,* alors que les trois tokens « pomme de terre » correspondent (ici) à une seule forme (composée). Notre définition de *token inconnu* repose sur un lexique de référence et une chaîne de traitement particulière : il s'agit d'un token qui n'appartient pas au lexique et qui ne fait pas partie d'une séquence considérée comme productive dans les ressources utilisées dans la chaîne de traitement (par exemple, une entité nommée). De ce fait, l'absence d'un module de reconnaissance des noms de personnes ou une erreur dans un tel module peut conduire à rendre inconnu un token faisant partie d'une entité nommée.

4 Proposition de typologie des inconnus

Afin de construire notre typologie des inconnus, nous avons étudié dans un corpus de dépêches de l'Agence France Presse, en français, anglais et espagnol[3], les différents types de tokens inconnus. On observe les grandes distinctions suivantes (du plus fréquent au plus marginal) :

- Certains tokens facilement analysables, qui relèvent de la **création lexicale productive**.
- Certains tokens difficilement décomposables comme une **forme lexicalisée** (*cupboard* en anglais).
- Certains tokens inconnus font partie de **séquences productives** et auraient pu ou dû être analysés comme tels par un analyseur, comme une date ou un « **cybertoken** » (posts Twitter : *#FT*, ou *m@il*), ou d'autres types de séquences productives.
- Certains tokens sont inconnus parce qu'erronés : des **erreurs** du scripteur (fautes, typo…) ou des erreurs de la chaîne de traitement (mauvais…).

La figure 1 illustre la typologie, les grands nœuds représentent les catégories principales décrites ci-après.

1. **Entité nommée** : Les tokens faisant partie d'**entités nommées** forment une catégorie de séquences productives très représentée, probablement majoritaire dans les corpus journalistiques. De nombreux inconnus résultent de l'absence ou de l'imperfection du module de reconnaissance des EN dans la chaîne de traitement. Pour annoter ces inconnus, nous différencions les EN simples et les EN complexes, et donc les tokens faisant partie d'une EN complexe. Par exemple, si la chaîne de traitement n'a pas reconnu *La Rochelle*, une double annotation est appliquée. « Rochelle » est d'abord annotée comme élément d'une EN complexe, puis « La Rochelle » est annoté comme EN. Par ailleurs, une sous-catégorie « abréviation » est dédiée aux sigles et acronymes et aux troncations (*Sarko*).

[2] Sauf lorsque l'on considère qu'ils ne servent pas à la ponctuation, comme les points dans les URL.
[3] Il s'agit d'un corpus constitué dans le cadre du projet EDyLex mentionné auparavant.

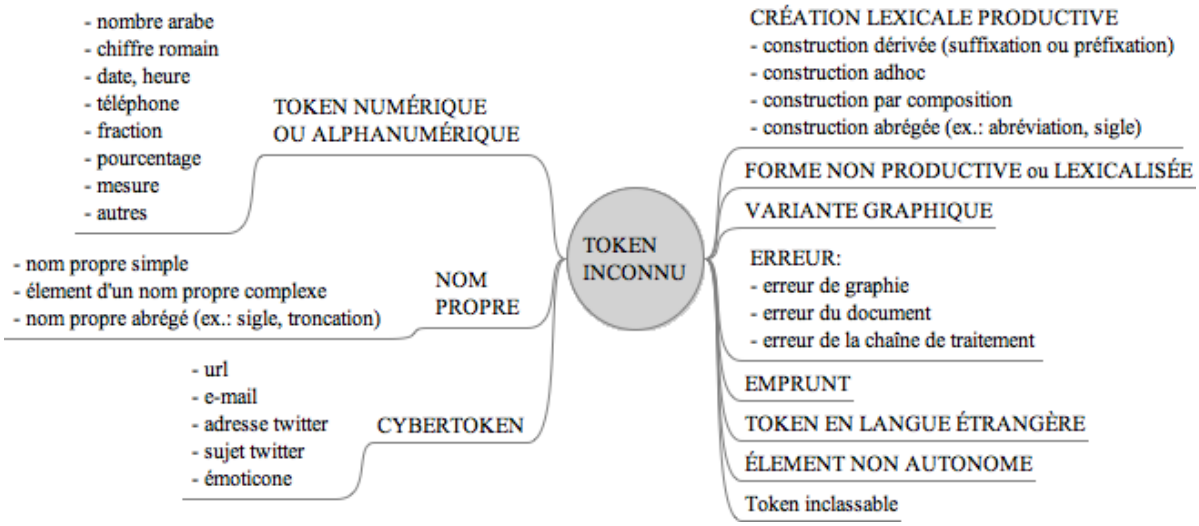

FIG. 1 – Vue générale de la typologie des inconnus.

2. **Création lexicale productive** : Nous proposons la sous-classification suivante qui comprend trois classes principales citées dans la littérature (Ren et Perrault, 1992 ; Morel, 2004 ; Dister et Fairon, 2004 ; Cartoni, 2008) et plusieurs autres classes ajoutées suite à l'observation de divers corpus, comme les néographies. Cette classe regroupe des tokens issus de la créativité lexicale, des processus morphologiques de création des mots et peut être subdivisée en sous-catégories en fonction du type de construction. Nous avons intégré la notion de « transparent » pour limiter cette catégorie aux formes régulières et facilement analysables par un des procédés cités ci-dessous.
   - Constructions dérivées par suffixation ou préfixation (*relire, tartinabilité*), formes flexionnelles non conventionnelles comme la forme « *pleuvent* », féminisations des noms comme « *professeure* » mentionnées par Dister et Fairon (2004).
   - Constructions *ad hoc* (*52ème*, *bleu-vert, internet-based* en anglais) : ces tokens sont fréquemment construits en contexte. Leur caractère « inconnu » dépend fortement de la chaîne de traitement utilisée. Ces éléments sont *a priori* exclus de l'enrichissement du lexique et relèvent de l'enrichissement du module de segmentation.
   - Constructions par composition (*porte-bébé, microclimat*) dont une partie significative correspond à des termes scientifiques.
   - Constructions abrégées : abréviations, sigles et acronymes, troncations, néographies (*à+, 4x4*).
3. **Forme non productive ou lexicalisée** : Il s'agit des tokens correspondant à des formes lexicalisées non transparentes et qu'il est difficile d'analyser par un des procédés répertoriés ci-dessus. Cette catégorie peut inclure des régionalismes, des termes, ainsi que des mots argotiques, mais également des lacunes du lexique. Nous n'avons pas créé ces sous-catégories explicites, car leurs frontières nous semblent floues et liées à des considérations subjectives, comme l'origine sociale ou géographique.

4. **Variante graphique** : Cette catégorie contient les tokens qui sont une variante de graphie d'une forme connue du lexique. À la différence de la catégorie 5, il s'agit ici de graphies correctes. La variante peut être justifiée par une réforme de l'orthographe (*surement*), par l'usage (*clé* vs. *clef*) ou par une variante géographique (*color* en anglais américain).

5. **Erreur** : Fautes d'orthographe et coquilles, erreurs du document (format ou encodage) et erreurs de la chaîne de traitement, en particulier les erreurs d'analyse de certains tokens qui en font des inconnus.

6. **Emprunt** : Tokens correspondant à des formes étrangères importées et utilisées dans la langue du texte de façon similaire aux autres formes (cf. le mot anglais *tagger*). Lorsque la forme a subi une modification pour être intégrée dans la langue du document, les tokens correspondants appartiennent à la sous-catégorie « emprunt adapté » (cf. *taggué*).

7. **Token en langue étrangère** : C'est le cas des tokens en langue étrangère mais non intégrés à la langue (ce ne sont donc pas des emprunts).

8. **Composant non autonome** : Tokens pouvant exister comme partie d'une forme (simple ou complexe) mais qui ne constituent pas une forme à eux seuls. C'est le cas d'un token comme « priori » s'il n'est pas reconnu comme faisant partie de la forme *a priori*.

9. **Token numérique ou alphanumérique** : Unités composées de chiffres ou de chiffres et de lettres, normalement non répertoriées dans un lexique, et lorsqu'elles ne sont pas reconnues en tant que séquences productives ; il peut s'agir par exemple de nombres (25 000), de mesures (7cm), etc.

10. **Cybertoken** : Chaîne liée aux nouvelles technologies de la communication, comme une URL ou une adresse électronique.

11. **Token inclassable** : Token ne rentrant pas dans une des catégories ci-dessus ; cette catégorie pourra être étendue au fil de l'annotation.

5 Annotation des inconnus en corpus et travaux futurs

L'adéquation de la typologie proposée à la section 4 au problème de l'incomplétude des ressources lexicales doit être confrontée à des données constituées en corpus, afin de mieux identifier les contours qualitatifs et quantitatifs des inconnus. Une annotation à échelle très réduite a été menée dans un premier temps afin de vérifier la cohérence et la pertinence générale de la typologie face aux données. Conduite sur le français et l'anglais (deux documents d'environ 5000 mots chacun) par trois annotateurs, cette expérience a permis d'introduire des modifications qui se sont avérées nécessaires durant l'annotation (notamment la prise en compte des NP simples vs complexes) et de valider la typologie comme guide pour une expérience à plus grande échelle : l'annotation d'un corpus de dépêches de l'Agence France Presse, en français, anglais et espagnol. Une centaine de dépêches pour chaque langue et alignées pour former un corpus parallèle, afin de bénéficier d'une comparaison inter-langues utile pour l'adaptation ultérieure des outils à améliorer. Ce corpus sera ensuite soumis à un traitement automatique à l'aide des ressources lexicales et des outils d'analyse utilisés dans le cadre du projet EDyLex : la chaîne d'analyse de surface SXPipe paramétrée pour chaque langue et basée sur les lexiques morphosyntaxiques L*efff* pour le français, L*effe* pour l'espagnol et EnLex pour l'anglais[4]. Le résultat permettra l'identification des tokens, tels que définis à la section 3, qui demeurent inconnus pour les ressources de référence, grâce à l'attribution de la catégorie adéquate fournie par la typologie. Cette expérience doit également servir à identifier d'éventuels cas d'inconnus non prévus par la typologie afin d'améliorer sa couverture et sa pertinence en termes de description du phénomène. Dans un souci de vérification de la cohérence de la typologie au cours de l'annotation, les accords inter-annotateurs seront calculés à intervalles réguliers.

[4] Toutes ces ressources sont librement disponibles sur les sites web des projets Lingwb pour SXPipe et Alexina pour les lexiques : http://lingwb.gforge.inria.fr/ et http://alexina.gforge.inria.fr/

La typologie présentée dans cet article et le corpus annoté qui résultera de cette expérience permettront le développement d'un outil automatique de détection et de typage des inconnus. Ceci constitue le point de départ de travaux menés au sein du projet EDyLex permettant de rendre *dynamique* les ressources lexicales et les chaînes de traitement. En effet, la classification automatique des inconnus permet de les traiter de façon adaptée en fonction de leur type. On peut alors proposer l'ajout temporaire ou définitif dans le lexique d'entrées lexicales manquantes ou nouvelles, éventuellement validées manuellement, puis immédiatement prises en compte dans une chaîne de traitement linguistique. Cet enrichissement dynamique est particulièrement pertinent face à flux riche en nouveautés lexicales, comme c'est le cas avec des dépêches d'agence de presse.

Références

CARTONI B. (2006). Constance et variabilité de l'incomplétude lexicale. *Noûs* (3), pp. 10–13.

DE LA CLERGERIE É., CLEMENT L. (2005). MAF: a Morphosyntactic Annotation Framework. Actes de *LTC'05* pp. 90–94, Poznań, Pologne.

DISTER A. ET FAIRON, C. (2004). Extension des ressources lexicales grâce à un corpus dynamique. *Lexicometrica.*

JANSSEN, M. (2009). Detección de Neologismos: una perspectiva computacional. *Debate Terminológico*, 5, p. 68–75.

GEYKEN, A. (2009). Automatische Wortschatzerschließung großer Textkorpora am Beispiel des DWDS. *Linguistik online, nº 39(3), pp. 97-108.*.ISSN 1615-3014.

JONASSON, K. (1994). *Le nom propre. Constructions et interprétations.* Duculot.

MAUREL, D. (2004). Les mots inconnus sont-ils des noms propres ? Actes des JADT 2004.

NAKOV P., BONEV Y., ANGELOVA G., GIUS E., VON HAHN, W. (2003). Guessing Morphological Classes of Unknown German Nouns. In *Proceedings of RANLP'03*. pp. 319-326. Borovets, Bulgarie.

MIKHEEV A., (1997). Automatic Rule Induction for Unknown-Word Guessing. In Computational Linguistics 23(3), pp. 405–423.

REN X, PERRAULT F. (1992). The Typology of Unknown Words: An Experimental Study of Two Corpora. In: Proceedings of *COLING'92*. Nantes; 1992.

SABLAYROLLES, J.F. (1997). Néologismes : Une typologie des typologies. Cahier du CIEL, pp. 11–48.

SABLAYROLLES, J.F. (2002). « Fondements théoriques des difficultés pratiques du traitement des néologismes », *Revue française de linguistique appliquée*, 7(1). « Lexique : recherches actuelles », pp. 97–111.

SCHONE, P., JURAFSKY, D. (2001). Knowledge-Free Induction of Inflectional Morphologies. *Proceedings of the 2nd Meeting of the North American Chapter of the ACL (NAACL'01)*.

TSENG, H., JURAFSKY, D., MANNING, C. (2005). Morphological features help POS tagging of unknown words across language varieties. Actes du 4th SIGHAN Workshop on Chinese Language Processing.

VILNAT A., FRANCOPOULO G., HAMON O., LOISEAU S., PAROUBEK P., VILLEMONTE DE LA CLERGERIE É. (2008). Large Scale Production of Syntactic Annotations to Move Forward. In : Proceedings of the COLING 2008 Workshop on Cross-Framework and Cross-Domain Parser Evaluation.

WONG, K., LI, W., XU, R., ZHANG, Z. (2009). *Introduction to Chinese natural language processing*. Synthesis lectures on human language technologies, Morgan & Claypool Publishers.