Conception d'algorithmes de rétroconversion

Pascale Renders (1) (2), Cyril Briquet (1)
(1) ATILF (CNRS & Nancy-Université), (2) Université de Liège
pascale.renders@ulg.ac.be, cyril.briquet@acm.org

Résumé

Rétroconvertir un ouvrage imprimé consiste à le transformer, de la manière la plus automatisée possible, en un ouvrage informatisé. Dans le cas d'un dictionnaire tel que le *Französisches Etymologisches Wörterbuch* (FEW), il s'agit en outre de l'enrichir d'un balisage XML qui permette des consultations ciblées. Le noyau de l'opération est constitué de 37 algorithmes dits "de rétroconversion" qui balisent chacun un type d'information "cible" du dictionnaire. Cet article présente, en guise d'exemple, la mise au point de l'algorithme qui reconnaît et balise le(s) signataire(s) d'un article du FEW.

1 Introduction

L'opération de rétroconversion d'un dictionnaire papier, consistant à transformer un ouvrage papier en un ouvrage informatisé, comprend l'identification et le balisage sémantique des types d'information "cibles". Le balisage, sémantique puisqu'il donne un sens aux données en les typant[1], est en effet le prérequis de toute exploitation complexe. Le balisage dans le *Französisches Etymologisches Wörterbuch* (FEW) des unités lexicales, des langues d'étymons, des états de langue et des datations permet par exemple de répondre à des questions telles que "quels lexèmes sont attestés pour la première fois au 17e siècle?" ou "quels mots attestés en moyen français proviennent d'un étymon grec?". Etant donné la complexité et le nombre d'articles du FEW (~20.000), il est nécessaire d'automatiser le balisage sémantique. Mettre au point un algorithme (dit algorithme de *rétroconversion*) identifiant un type d'information donné et insérant le balisage sémantique correspondant dans une représentation XML d'un article donné du dictionnaire constitue le problème de base de la rétroconversion.

La mise au point d'un algorithme de rétroconversion est loin d'être triviale. La nature implicite des informations du FEW, le manque de consistance syntaxique des articles, les incohérences inévitables dans un ouvrage écrit sur une longue période apportent chacun leur part de complexité. La mise au point de l'algorithme est, de plus, effectuée sur la base d'un échantillon à taille humaine, nécessairement petit en comparaison du nombre total d'articles du dictionnaire. Par conséquent, il est nécessaire d'imaginer et de prévoir des irrégularités qui seront légitimement présentes hors de l'échantillon.

Le reste de l'article présente de manière exemplative la mise au point de l'algorithme de détection automatique de la signature dans les articles du FEW. La définition du résultat attendu (2), l'étude des critères de détection (3) et la prise en compte des cas particuliers (4) sont les étapes préalables à la construction de l'algorithme final (5).

1 Nous opposons balisage sémantique et balisage typographique, ce dernier concernant la mise en forme et ne donnant aucune information sur le contenu du texte.

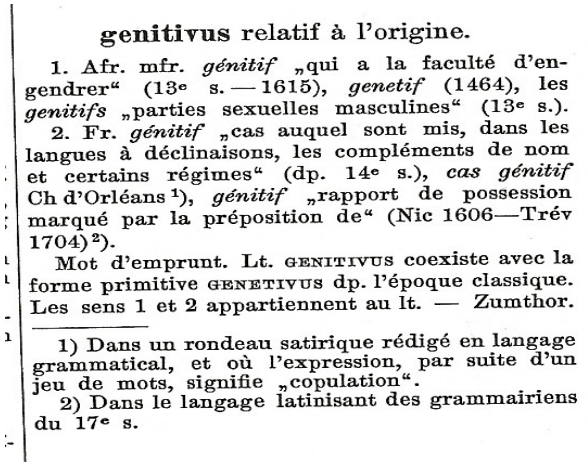
2 Objectifs de l'algorithme

La signature est le nom du rédacteur ou des rédacteurs responsable(s) d'un article (cf. Büchi 1996, 160-161)[2]. Elle apparaît à la fin du commentaire et en est séparée par un tiret (cf. figure 1). Le problème du traitement de la signature lors de la rétroconversion consiste à mettre au point un algorithme qui détecte et balise les signatures d'articles puis identifie le nom du rédacteur et détermine la métalangue utilisée (français ou allemand). Par exemple, la signature de l'article GENITIVUS sera balisée

<signature author = "Zumthor" lang = "french"> Zumthor </signature>

La signature peut être implicite (4.5); dans ce cas, une balise vide sera insérée à la fin de l'article:

<signature author = "Wartburg" lang = "german" />



genitivus relatif à l'origine.

1. Afr. mfr. *génitif* „qui a la faculté d'engendrer" (13e s. — 1615), *genetif* (1464), les *genitifs* „parties sexuelles masculines" (13e s.).

2. Fr. *génitif* „cas auquel sont mis, dans les langues à déclinaisons, les compléments de nom et certains régimes" (dp. 14e s.), *cas génitif* Ch d'Orléans [1]), *génitif* „rapport de possession marqué par la préposition de" (Nic 1606—Trév 1704) [2]).

Mot d'emprunt. Lt. GENITIVUS coexiste avec la forme primitive GENETIVUS dp. l'époque classique. Les sens 1 et 2 appartiennent au lt. — Zumthor.

1) Dans un rondeau satirique rédigé en langage grammatical, et où l'expression, par suite d'un jeu de mots, signifie „copulation".

2) Dans le langage latinisant des grammairiens du 17e s.

Figure 1: signature "Zumthor" par référence à Paul Zumthor (FEW 4, 102b, genitivus)

3 Critères de détection

L'existence d'indicateurs spécifiques, ou d'une combinaison spécifique d'indicateurs, est la condition *sine qua non* pour qu'un balisage automatique soit possible.

3.1 Indicateurs textuels

Une première catégorie d'indicateurs, dits textuels, concerne le contenu de l'information. Quelles chaînes de caractères particulières caractérisent les signatures? Peut-on définir ces chaînes de caractères et s'en servir comme critère de détection?

Les divers rédacteurs des volumes 1 à 25 du FEW étant connus (ils sont répertoriés dans les préfaces des différents volumes), il est envisageable d'en constituer la liste. Cette liste peut servir de critère de détection à condition 1/ d'être exhaustive et 2/ de correspondre exactement à ce qu'on trouve dans le FEW. Elle ne peut donc se résumer aux noms de famille des rédacteurs ("Wartburg"), mais doit comprendre toutes les *formes* sous lesquelles le FEW les mentionne en tant que signataires. Outre les abréviations ("Wbg"), on mentionnera les noms précédés de l'initiale du prénom en majuscule ("MHoffert"). Chacune de ces formes sera associée au rédacteur qui lui

2 Plus rarement, elle indique aussi le responsable d'une partie de l'article, lorsqu'il diffère du rédacteur (cf. Büchi 1999, 162; 161, n 218). Ce cas particulier ne sera pas étudié ici.

correspond grâce à la constitution d'une liste en deux colonnes, la première colonne relevant toutes les formes et la seconde, le nom de famille du rédacteur correspondant.

3.2 Indicateurs positionnels

Comme leur nom l'indique, les indicateurs positionnels se basent sur la position de l'information par rapport à d'autres éléments du texte. Les signatures occupent-elles une place particulière dans les articles du FEW?
La signature constitue toujours la dernière partie du champ alloué au commentaire. Elle est délimitée à droite par le point final de ce champ, puis, soit par le champ alloué aux notes de fin d'article, soit par l'interligne qui sépare l'article en question de l'article suivant. À sa gauche, la signature est normalement précédée d'un tiret cadratin ou semi-cadratin et d'une espace éventuelle. Ces règles contextuelles simples deviennent plus complexes en cas de signature multiple (articles signés par plusieurs rédacteurs). Les noms des signataires se suivent alors, séparés par un point-virgule, une virgule ou un point (avec espace éventuelle). Le début et la fin de la séquence respectent toutefois les règles énoncées pour les signatures uniques.
Un élément important est à souligner: la prise en compte des indicateurs positionnels nécessite le balisage préalable, soit du champ alloué au commentaire, soit de la fin d'article et du champ alloué aux notes. L'algorithme de reconnaissance de la signature *dépend* donc d'autres algorithmes.

3.3 Combinaison des critères

Les critères textuels seuls ne sont pas suffisants pour reconnaître une signature. Il arrive par exemple que le nom d'un rédacteur du FEW soit identique à une référence bibliographique. Les critères positionnels seuls ne sont pas suffisants non plus, pour deux raisons. La première raison est qu'ils génèrent des "faux négatifs", c'est-à-dire le non-balisage de vraies signatures. En cas d'absence ou de mauvaise transcription du tiret (semi-)cadratin par exemple, la signature ne sera pas reconnue. La première raison est qu'ils génèrent également des "faux positifs", c'est-à-dire le balisage de fausses signatures: une information qui occupe la même position avec le même contexte serait balisée comme une signature alors que c'est une référence bibiographique (par ex. "ML 3744" s.v. GĔRMEN, -INE), un commentaire (par ex. "Lehnwort" s.v. ISCHURIA) ou encore un renvoi à un autre article (par ex. "S. noch *TRAGŬLARE" s.v. STRAGŬLUM). Éliminer les faux positifs demanderait de complexifier les règles positionnelles ou de baliser préalablement – selon d'autres critères – toutes les informations susceptibles d'occuper la même position.
En revanche, la combinaison des critères textuels et des critères positionnels permet de reconnaître une signature sans aucune ambiguïté : un morceau de texte est une signature s'il correspond à une forme de la liste des rédacteurs et si son contexte vérifie les critères positionnels énoncés ci-dessus.

3.4 Attribution de la métalangue

La métalangue d'un article peut être déterminée en fonction de son rédacteur. La liste de mots-clés contenant les noms des rédacteurs est dès lors augmentée d'une troisième colonne indiquant la métalangue correspondant à chaque rédacteur.

4 Problèmes particuliers

Les critères présentés ci-dessus (3) permettent de construire la base de l'algorithme de rétroconversion. Tel quel, il reconnaîtra correctement un grand nombre de signatures. Un certain nombre de problèmes particuliers contribuent toutefois à la complexité du problème et doivent être intégrés à l'algorithme.

4.1 Variantes graphiques

La méthode choisie – reconnaissance par mots-clés – impose une stricte correspondance entre les mots-clés de la liste et les formes effectivement présentes dans le FEW. La question se pose donc de savoir si toutes les possibilités de variantes peuvent être ajoutées à la liste. Dans le cas des signatures, trois problèmes sont à envisager. Le premier concerne l'apparition ou non d'une petite espace dans les signatures du type "M Hoffert", espace susceptible de disparition ("MHoffert") ou de transcription comme une espace normale ("M Hoffert") lors de la récupération du texte. Une solution facilement réalisable en pratique est d'intégrer ces variantes dans la liste de mots-clés. Un deuxième problème concerne le codage des caractères spéciaux, présents par exemple dans le nom du rédacteur Lubomir Smiřický. Garantir un même codage UTF-8 des articles du FEW et de la liste de mots-clés constitue une première réponse à ce problème. Les variantes "Smiricky", "Smirický" et "Smiřicky" peuvent de plus être intégrées à la liste, afin que ce mot-clé soit reconnu en cas de mauvaise saisie ou d'incohérence du FEW. Enfin, un dernier problème concerne la présence de traits d'union de fin de ligne, qui mènent à des coupures du type "Cham-bon". Doit-on également mettre dans la liste toutes les possibilités de coupure des noms de rédacteurs ? Ce problème général concerne d'autres algorithmes et requiert une solution globale, grâce à laquelle les variantes comportant des traits d'union internes ne doivent pas être ajoutées à la liste.

4.2 Interférences

D'autres informations peuvent interférer avec les règles décisionnelles que nous avons établies jusqu'à présent. Des appels de notes apparaissent parfois après la signature ou après le point final du commentaire (ex. FEW 24, 495b ; 525b ; 636a). Ces interférences doivent être prises en compte lors de l'application des critères de détection. Deux solutions sont envisageables : soit une définition plus large du contexte permis entre le mot-clé et la fin du paragraphe, soit le balisage préalable des appels de note assorti de la possibilité de les rendre invisibles lors de la détection des signatures. Cette deuxième solution, plus élégante, a été choisie.

4.3 Incohérences et inconsistances

La solution proposée pour l'attribution de la métalangue (3.4) ne répond pas à toutes les situations permises par le FEW. Hubschmid, pourtant germanophone, a rédigé les articles du volume 25 en français, après en avoir rédigé en allemand dans les volumes précédents. Smiřický a quant à lui rédigé ses articles tantôt en français, tantôt en allemand, dans le même volume. Une liste annexe de mots-clés a été établie, contenant les 15 articles qu'il a écrits en français, désignés par l'étymon-vedette. Si l'étymon-

vedette de l'article traité appartient à cette liste, l'algorithme désignera le français comme métalangue de l'article.

4.4 Erreurs du FEW

Par rapport aux règles établies ci-dessus, le FEW comporte des erreurs, c'est-à-dire des situations non permises. L'absence d'une signature dans des articles non rédigés par von Wartburg dans la majeure partie du FEW (ou par Jänicke, dans la section slave), en est une. Il est impossible de repérer ce type d'erreur : les articles en question seront attribués à von Wartburg ou à Jänicke, respectivement. Cependant, nous pouvons signaler une absence de signature lorsqu'elle a lieu dans les parties du FEW où tous les articles sont censés être signés, par exemple dans les volumes 24 (à partir du fascicule 142) et 25. C'est le cas de l'article ANIMANS (FEW 25, 594a)[3]. La résolution de telles erreurs n'est pas décidable de façon automatisée. L'algorithme se contentera de signaler ces erreurs, afin que l'information puisse être ajoutée manuellement.

4.5 Insertion de signatures implicites

Un grand nombre d'articles ne contiennent pas de signature en structure de surface. En structure profonde (cf. Büchi 1996, 117), on peut y distinguer deux groupes. Le premier est constitué d'articles dont la signature est véritablement absente. C'est le cas des articles de renvoi, qui doivent donc être exclus du traitement. Dans le second groupe, la signature est seulement implicite. La règle permettant de la rétablir nécessite de connaître le volume et la page du FEW où l'on se trouve. Le signataire par défaut est W. von Wartburg, excepté dans le tome 20, p. 33-52 (section slave, rédigée par O. Jänicke) et le tome 22, p. 193-217 (articles rédigés par M. Hoffert).

5 Description de l'algorithme tag-signature

L'analyse proposée permet de construire un algorithme de détection des signatures basé sur des critères très fiables requérant, d'une part, le balisage préalable des renvois, des étymons, des notes et, d'autre part, l'établissement d'une liste principale et d'une liste annexe de mots-clés, toutes deux codées UTF-8 comme les articles. Un morceau de texte sera balisé comme signature s'il correspond à une forme de la liste des rédacteurs et si son contexte droit et gauche correspondent aux critères positionnels définis. L'algorithme détecte donc toutes les occurrences de mots-clés de la liste principale. Pour chacune, les critères positionnels sont vérifiés (3.2). La détection des mots-clés et la vérification de leur contexte sont effectués avec les appels de note rendus invisibles (4.2). Les mots-clés licites sont balisés avec les attributs "author" et "lang", en considérant les cas particuliers présentés par Hubschmid et Smiřický (4.3). Si aucune signature explicite n'a été trouvée, l'algorithme insère à la fin de l'article une balise vide, contenant les attributs "author" et "lang" correspondant au rédacteur implicite de l'article, en fonction du volume et des pages où se trouve l'article.

3 La paternité de cet article doit probablement être attribuée à J.-P. Chambon, qui a rédigé les articles ANĬMA, ANĬMARE etc.

```
POUR CHAQUE article:
SI mot-clé suivi d'un point trouvé juste avant <notes>
   (ou, si pas de notes, juste avant la fin de l'article) ALORS
     chercher si d'autres mots-clés précèdent
        (séparés par tiret, point ou point-virgule);
     POUR CHAQUE chaque mot-clé trouvé:
       SI (mot-clé = "Hubschmid") ET (vol = "25") ALORS
              métalangue = "french"
       SINON SI mot-clé = "Smiricky" ALORS
              métalangue = langue de l'étymon définie dans liste annexe
       SINON
              auteur et métalangue correspondant au mot-clé dans liste
       FIN SI;
       baliser <signature author="..." lang="..."> mot-clé </signature>
SINON
     SI vol="20" ET page="[33-52]" ALORS
              author = "Jänicke"; lang = "german"
     SINON SI (vol="24") OU (vol="25") ALORS
              author = "?"; lang= "?"; émettre avertissement
     SINON
              author = "Wartburg" ; lang = "german"
     FIN SI ;
     insérer <signature author="..." lang="..."/>
        juste avant <notes> (ou, si pas de notes, avant la fin de l'article)
FIN SI
```

Figure 2: algorithme (simplifié) de détection des signatures

6 Conclusion

Un algorithme de rétroconversion se construit par étapes, du cas général aux cas particuliers, avec affinages successifs. Une connaissance théorique et pratique du FEW ne suffit pas: des tests sur corpus sont une nécessité tant pour mettre au point l'algorithme que pour le valider.
L'algorithme de détection des signatures d'articles présente des caractéristiques communes avec la plupart des algorithmes de rétroconversion du FEW: recherche de mots-clés, traitement de l'implicite et prise en compte d'une série de cas particuliers; dépendance par rapport à d'autres algorithmes de rétroconversion.
En pratique, un substrat complexe d'analyse et de modification de document XML est requis, permettant la détection et l'insertion de balises, le positionnement par rapport à une balise ainsi que la recherche, sensible à la présence de balises, de mots-clés.

Remerciements

Nous tenons à remercier nos collègues, dont Gérard Dethier, pour leurs interventions pertinentes lors du colloque, ainsi qu'Eva Buchi pour sa relecture de l'article.

Références

FEW = Wartburg, Walther von *et al.*, 1922-2002. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes* (25 vol.), Bonn/Heidelberg/Leipzig-Berlin/Bâle : Klopp/Winter/Teubner/Zbinden.

Büchi (1996). *Les Structures du 'Französisches Etymologisches Wörterbuch'. Recherches métalexicographiques et métalexicologiques*, Tübingen : Niemeyer.