Détection et correction automatique d'entités nommées dans des corpus OCRisés

Benoît Sagot, Kata Gábor

Détection et correction automatique d'entités nommées dans des corpus OCRisés

Benoît Sagot Kata Gábor
Alpage, INRIA & Université Paris-Diderot, 75013 Paris
{prenom.nom}@inria.fr

Résumé. La correction de données textuelles obtenues par reconnaissance optique de caractères (OCR) pour atteindre une qualité éditoriale reste aujourd'hui une tâche coûteuse, car elle implique toujours une intervention humaine. La détection et la correction automatiques d'erreurs à l'aide de modèles statistiques ne permettent de traiter de façon utile que les erreurs relevant de la langue générale. C'est pourtant dans certaines entités nommées que résident les erreurs les plus nombreuses, surtout dans des données telles que des corpus de brevets ou des textes juridiques. Dans cet article, nous proposons une architecture d'identification et de correction par règles d'un large éventail d'entités nommées (non compris les noms propres). Nous montrons que notre architecture permet d'atteindre un bon rappel et une excellente précision en correction, ce qui permet de traiter des fautes difficiles à traiter par les approches statistiques usuelles.

Abstract. Correction of textual data obtained by optical character recognition (OCR) for reaching editorial quality is an expensive task, as it still involves human intervention. The coverage of statistical models for automated error detection and correction is inherently limited to errors that resort to general language. However, a large amount of errors reside in domain-specific named entities, especially when dealing with data such as patent corpora or legal texts. In this paper, we propose a rule-based architecture for the identification and correction of a wide range of named entities (proper names not included). We show that our architecture achieves a good recall and an excellent correction accuracy on error types that are difficult to adress with statistical approaches.

Mots-clés : OCR, Entités nommées, Détection d'erreurs, Correction d'erreurs.

Keywords: OCR, Named entities, Error Detection, Error Correction.

1 Introduction et état de l'art

Le projet dans lequel s'inscrit le travail présenté ici a pour objectif d'optimiser des performances des projets de numérisation patrimoniale par le biais de l'automatisation de la détection et de la correction d'erreurs dans des documents en sortie de systèmes de reconnaissance optique de caractères (Optical Character Recognition, ou OCR). En effet, la correction de données textuelles obtenues par reconnaissance optique de caractères (OCR) pour obtenir une qualité éditoriale est aujourd'hui une tâche coûteuse, même lorsque les documents d'origine sont d'excellente qualité, car elle implique toujours une intervention humaine, seule à même de garantir des taux d'erreurs acceptables par exemple pour publication sous forme de livre électronique. Pour diminuer le coût de cette intervention humaine, des approches automatiques pour la détection d'erreurs, pour la suggestion de corrections voire pour la correction automatique sont progressivement utilisées. Ces approches reposent presque toujours sur l'utilisation de modèles de langage et de modèles d'erreur, mais rarement sur des informations linguistiques plus riches. Pourtant, ces techniques ne permettent pas de traiter de façon utile les erreurs qui ne relèvent pas de la langue générale, et notamment les erreurs présentes dans les entités nommées. C'est pourtant dans certains types d'entités nommées que résident les erreurs les plus nombreuses, surtout dans des données telles que des corpus de brevets (formules chimiques) ou des textes juridiques (identifiants de jurisprudence). En effet, de telles entités ne peuvent être présentes dans les lexiques sur lesquels reposent les systèmes d'OCR. Elles ne sont pas adaptées à des traitements statistiques, et sont au contraire adaptées à une modélisation par règles qui en permette la détection robuste, malgré le bruit issu de l'OCR, et la correction. Dans cet article, nous proposons ainsi d'étudier l'impact sur la correction automatique de sorties d'OCR d'une prise en charge par règles des erreurs présentes dans un large éventail d'entités nommées.

Les premières approches à la correction automatique des erreurs (orthographiques, lexicales ou post-OCR) s'appuyaient fortement sur le lexique : pour chaque mot inconnu, elles cherchaient des candidats proches (par exemple en termes de

distance d'édition (Levenshtein, 1965)) qui figurent dans le lexique et choisissent en prenant en compte la fréquence des candidats, le contexte, ou éventuellement un poids associé au type d'erreur présumé. Cependant, l'utilisation des lexiques pré-existants semble laisser la place à d'autres méthodes plus flexibles qui ne nécessitent pas la consultation d'une telle ressource, ou qui la créent à la volée à partir du corpus à corriger (Reynaert, 2004) ou en exploitant les données du Web (Cucerzan & Brill, 2004; Strohmaier *et al.*, 2003). Deux méthodes fondamentales peuvent être distinguées dans les travaux plus récents visant la correction des documents bruités : les systèmes s'appuyant sur des automates à états finis (pondérés) (Ringlstetter *et al.*, 2006), et des méthodes adoptant le modèle du canal bruité (Kernighan *et al.*, 1990; Brill & Moore, 2000; Kolak & Resnik, 2002). Contrairement aux systèmes symboliques, ces dernières ne font aucune hypothèse a priori sur le type d'erreurs que l'on prévoit de rencontrer dans le document et se prêtent ainsi plus facilement à l'adaptation à de nouveaux domaines. Cependant, comme la matrice de confusion qui sert de modèle d'erreur est construite à partir d'un corpus, il est souhaitable de disposer d'un corpus d'apprentissage de taille conséquente. Les méthodes symboliques évitent le problème du besoin de ressources annotées et permettent une meilleure intégration d'informations résultant d'une analyse linguistique, mais elles présentent l'inconvénient d'être spécifiques à une langue et, typiquement, à un domaine. D'autres méthodes utilisent une combinaison des sorties de plusieurs OCR : Klein & Kope (2002) utilisent un système de vote, alors que Lund & Ringger (2009) produisent un treillis de mots à partir d'un alignement textuel effectué sur les sorties et choisissent l'une des hypothèses par recherche dans un dictionnaire. Mais ces techniques sont mal adaptées à la correction d'erreurs au sein d'entités nommées souvent complexes, dispersées (*sparse*) et spécifiques au domaine.

Pour palier cette limitation, nous avons adapté et étendu la chaîne de traitement de surface SXPipe (Sagot & Boullier, 2008) pour effectuer la tokenisation, la segmentation en énoncés et le traitement de divers types d'entités nommées dans des sorties d'OCR. En particulier, nous avons développé de nouvelles grammaires locales pour traiter des classes d'entités nommées rencontrées dans les corpus traités dans le projet (formules chimiques, identifiants juridiques), nous avons adapté l'ensemble des grammaires locales pour les rendre robustes aux erreurs d'OCR, et nous avons intégré un nouveau mécanisme permettant de compléter la reconnaissance d'une entité par une proposition de correction selon des règles qui dépendent en partie du type d'entité. L'objectif, à terme, est de coupler cette chaîne avec une architecture plus classique pour la correction des tokens hors-entités (modèle de langage, modèle d'erreur).

Dans ce qui suit, nous allons présenter les corpus traités dans le cadre de nos travaux (section 2). Une description des grammaires locales construites pour la reconnaissance et la correction des entités nommées est fournie à la section 3, et les résultats sont décrits à la section 4.

2 Corpus à traiter

Les trois corpus sur lesquels nous avons travaillé sont présentés à la table 1. Contrairement aux deux autres, le corpus BNF/Gallica est un *corpus d'écarts*, c'est-à-dire un corpus constitué d'une version brute (sortie directe d'OCR) et d'une version corrigée par des opérateurs humains (qualité éditoriale), les deux versions étant alignées. Les corpus d'écarts contribuent à créer des modèles d'erreur.

Corpus	**Genre**	**Nb. de tokens**	**Type de corpus**
BNF/Gallica	littérature	50 000 000	corpus d'écarts
OPOCE	jurisprudence	37 000 000	sortie d'OCR
EPO (European Patent Office)	brevets	15 000 000	sortie d'OCR

TABLE 1 – Corpus utilisés dans cette étude

L'étude de ces corpus nous a permis d'identifier certains types d'entités comportant de nombreuses erreurs d'OCR. Ainsi, dans le corpus de brevets EPO, une proportion importante des erreurs d'OCR sont situées au sein de formules chimiques, indications de mesures et numéros de brevets. Le corpus de jurisprudence européenne contient de nombreux identifiants juridiques tels que des références à des articles de lois, des arrêtés ou des décisions.

De manière générale, pour ces entités comme pour les entités plus classiques (dates, adresses…), qui contiennent également des erreurs d'OCR que nous souhaitons corriger, une approche par règles semble plus adaptée qu'une approche par modèle de langue, en raison de la grande variété des occurrences de ces entités, mais également en raison de leur forte structuration et des nombreuses contraintes non-linguistiques qui s'y appliquent.

3 Grammaires locales pour la correction de sorties d'OCR

3.1 Définition des entités nommées

La tâche d'étiquetage des entités nommées a reçu une attention considérable au sein de la communauté TAL depuis les années 90. Une des tâches partagées de la série de conférences MUC (*Message Understanding Conference* (Chinchor, 1998)) avait pour objectif de reconnaître les entités nommées, définies en tant qu'unités faisant référence à une entité unique et concrète et réalisées par des noms propres (noms de personnes, d'organisations, d'artefacts ou de lieux). Les expressions temporelles et les expressions de quantité sont généralement ajoutées à cette liste, moins en raison de leurs propriétés sémantiques que pour des considérations d'ordre pratique. Toutefois, la plupart des campagnes d'évaluation (p.ex. CoNLL 2003, (Sang & Meulder, 2003)) se contentent d'illustrer les différentes catégories d'entités nommées de façon extensionnelle, par des exemples.

Dans cet article nous définissions une (mention d')entité nommée comme une séquence de caractères ou de tokens qui n'est analysable ni morphologiquement ni syntaxiquement, mais procède d'un motif productif, qui peut contenir des marques de ponctuation. Nous utilisons ainsi le terme d'entité nommée dans un sens légèrement plus large que le sens habituel (Maynard *et al.*, 2001), en y incluant également les nombres, les formules chimiques, les unités monétaires, etc. En revanche, nous ne traitons pas ici des entités nommées classiques que sont les noms de personnes, de lieux ou d'organisations (sauf en tant que composants d'entités plus larges comme les identifiants juridiques).

3.2 La chaîne de traitement SXPipe-postOCR

SXPipe (Sagot & Boullier, 2008) est une chaîne de traitement dont le rôle est d'appliquer à des corpus une cascade de traitements linguistiques de surface. Selon leur fonction, ces traitements peuvent être classés en trois catégories : détection de frontières (entre phrases ou mots), correction d'erreurs et reconnaissance des entités nommées. Tous les modules de la chaîne sont paramétrables, pour s'adapter à différentes applications. De plus, chaque utilisateur peut construire sa propre chaîne de traitement avec ses propres paramètres et jeux de modules, mais également créer et intégrer ses propres modules. Nous avons ainsi construit un exécutable spécifique SXPipe-postOCR, qui repose sur des versions améliorées de modules existants et sur de nouveaux modules, auxquels ont été ajoutés des règles permettant de proposer une correction pour certaines entités reconnues.

SXPipe contenait déjà des grammaires locales effectuant la reconnaissance de nombreux types d'entités nommées, comme les expressions temporelles ou les adresses. Elles ont été améliorées à la fois en couverture et en robustesse, afin de reconnaitre ces entités même dans des données légèrement bruitées issues de systèmes d'OCR. De plus, de nouvelles grammaires locales ont été développées spécifiquement pour ce projet, notamment pour la reconnaissance des dimensions, des unités monétaires, des formules chimiques et des identifiants de jurisprudence, toujours avec un souci de large couverture et de robustesse. Enfin, ces grammaires ont été complétées par des règles de *correction*, dont l'application (facultative) permet d'obtenir pour certaines entités nommées reconnues une version normalisée/corrigée. À cet égard, par prudence, l'accent a été mis sur la précision des propositions de correction, une correction incorrecte étant bien plus dommageable qu'une absence de correction pour ce type de tâche. Ainsi, la quasi-totalité des grammaires se basent sur la présence de marqueurs non-ambigüs, et une grande partie d'entre elles exploitent de l'information contextuelle.

3.2.1 Détection robuste et correction automatique dans les grammaires locales de SXPipe

La reconnaissance et la correction/normalisation des entités sont assurées par ces grammaires locales (cascades d'expressions régulières), chaque type d'entité correspondant à un module distinct. Les propositions de correction/normalisation bénéficient ainsi de connaissances spécifiques sur leur nature et leur structure interne. Par exemple, dans le cas de dimensions (p.ex. : 10^3 l/m^2), la structure en valeur et/ou puissance de 10 puis unité de mesure permet d'appliquer des corrections spécifiques : la puissance de 10 peut être mise en exposant, et l'unité de mesure peut voir tous ses chiffres 1 transformés en lettres l, tous les autres chiffres mis en exposant et tous les espaces supprimés (sous condition). C'est ainsi que l'on peut corriger « 10 3 1/m2 » en « 10^3 l/m^2 ».

L'identification de ces règles de correction a été faite par exploration manuelle du corpus d'écarts BNF/Gallica, à l'exception des formules chimiques. Ce dernier cas appelle par ailleurs quelques précisions supplémentaires (cf. section suivante).

Les tables 2 et 3 illustrent les différents cas possibles (détection ou non-détection, normalisation proposée ou non, correcte

ou non, ayant introduit ou non des erreurs). La table 2 rassemble des exemples pour lesquels cette normalisation est totalement correcte, la table 3 illustre diverses difficultés rencontrées par SXPipe-postOCR.

Sortie d'OCR	**Étiquetage**	**Normalisation proposée = Référence**
23 avril i908 1" janvier 1912 1G février 1859	{23 avril i908} _DATE {1" janvier 1912} _DATE {1G février 1859} _DATE	23 avril 1908 1er janvier 1912 16 février 1859
Al(OH)3 Ba2Na(NbO3)5 C02	{Al(OH)3} _CHEM {Ba2Na(NbO3)5} _CHEM {C02}_CHEM	$Al(OH)_3$ $Ba_2Na(NbO_3)_5$ CO_2
MgA1204	{MgA1204}_CHEM	$MgAl_2O_4$
484,9 g.l-1 2,10 10 6 1/m2	{484,9} _NUM {g.l-1} _DIMENSION {2,10 10 6 1/m2} _DIMENSION	484,9 $g.l^{-1}$ 2,10 10^6 l/m^2
52, rue de Chabrol, Paris, Xèriie rue de l'Université, i3 52, rue Taibout, Paris (9') 1, bd Beau-Séjour, Paris (16°)	{52, rue de Chabrol, Paris, Xèriie} _ADRESSE {rue de l'Université, i3} _ADRESSE {52, rue Taibout, Paris (9')} _ADRESSE {1, bd Beau-Séjour, Paris (16°)} _ADRESSE	52, rue de Chabrol, Paris, Xème rue de l'Université, i3 52, rue Taibout, Paris (9e) 1, bd Beau-Séjour, Paris (16e)

TABLE 2 – Reconnaissance et correction/normalisation des entités bruitées : exemples corrects

Sortie d'OCR	**Étiquetage**	**Normalisation proposée**	**Référence**
15 rue Spufflot 3ï, rue Taitboùt, Paris	{15 rue Spufflot} _ADRESSE {3ï, rue Taitboùt, Paris} _ADRESSE	15 rue Spufflot 31, rue Taitboùt, Paris	15 rue Soufflot 31, rue Taitbout, Paris
C8-C20 (CH2)nNR4R5 R1COOH	{C8-C20} _CHEM ({CH2} _CHEM)nNR4R5 R1COOH	C_8-C_2O (CH_2)nNR4R5 R1COOH	C_8-C_{20} $(CH_2)_n NR_4R_5$ R_1COOH

TABLE 3 – Reconnaissance et correction/normalisation des entités avec bruit : autres exemples

3.2.2 Le cas des formules chimiques

Les formules chimiques, malgré les nombreuses régularités et contraintes qui les régissent, représentent un défi important pour la reconnaissance automatique. Compte tenu de leur fréquence dans le corpus de brevets EPO et des nombreuses erreurs d'OCR que l'on y rencontre, nous leur avons consacré une attention particulière. Au lieu d'utiliser un des outils existants (Hawizy *et al.*, 2011; Rocktäschel *et al.*, 2012) pour la reconnaissance des formules, nous avons décidé de créer notre propre grammaire, ce qui permet une meilleure prise en charge des erreurs d'OCR. La grammaire SXPipe dédiée que nous avons développée est donc robuste à certaines erreurs, telles que la substitution de la lettre l par le chiffre 1, de la lettre O par le chiffre 0, et par le fait que les nombres qui devraient être en indices ne le sont pas. Certaines règles de correction exploitent des connaissances chimiques (par exemple, un élément chimique ou une sous-formule ne peut pas être indicé par 0 ou 1). Ainsi, si l'on donne MgA1204 en entrée à ce module, il produit en sortie {MgA1204}_CHEM_$MgAl_2O_4$_. Les tables 2 et 3 donnent quelques autres exemples.

Outre la sutructure interne des formules chimiques, la grammaire s'appuie sur leur contexte d'apparition. Dans la description de la structure des formules potentielles, nous sommes partis d'une liste complète d'éléments chimiques. Certains d'entre eux, tout comme certaines formules simples, sont toutefois ambigus avec des mots de la langue générale (p. ex. Ce, Ne, La). Nous avons donc construit deux ensembles de règles de reconnaissance. Un premier ensemble, fiable, ne reconnaît que des formules qui ne sont pas ambiguës avec des mots de la langue ou des acronymes. Un second ensemble de règles reconnaît de telles séquences ambiguës. Ce n'est que si les règles les plus sûres se sont appliquées avec succès sur une phrase que nous lui appliquons les règles plus ambigües : en effet, nous supposons alors que nous sommes en présence d'un contexte thématique favorisant l'apparition des formules chimiques.

4 Évaluation

Nous avons effectué une évaluation manuelle des grammaires d'entités nommées et des corrections/normalisations, en nous limitant aux dates, adresses et formules chimiques. Pour les deux premiers types d'entités, nous avons effectué cette évaluation à partir du corpus BNF/Gallica ; les formules chimiques ont été évaluées sur le corpus EPO.

Pour l'évaluation de la reconnaissance des dates, nous avons sélectionné aléatoirement 200 phrases parmi un ensemble de phrases susceptibles de contenir au moins une entité nommée de ce type. Pour ce faire, nous avons eu recours à des marqueurs couvrants (mais moins sûrs que nos grammaires) qui permettent de détecter les dates avec un rappel (presque) parfait mais une précision moindre. L'idée sous-jacente est que le nombre de dates qui ne sont pas capturées par nos marqueurs est très faible, de sorte qu'il est possible de calculer sur un tel échantillon non seulement la précision mais également le rappel de notre grammaire de reconnaissance. Nous avons procédé de même pour les adresses.

Pour l'évaluation manuelle de la correction/normalisation des dates et adresses, nous n'avons pas pu nous appuyer sur ces deux ensembles de 200 phrases, dans la mesure ou le pourcentage d'entités pour lesquelles une modification est proposée est trop bas. Nous avons donc évalué manuellement l'intégralité des cas où une modification est proposée par nos grammaires parmi 500 000 phrases du corpus BNF/Gallica, en nous référant si besoin à la version de qualité éditoriale corrigée par des humains.

Pour les formules chimiques, l'absence de marqueurs vraiment couvrants nous a conduit à choisir un échantillon aléatoire de 200 phrases parmi celles où notre grammaire a détecté au moins une formule chimique. Le rappel calculé sur cet échantillon est donc une borne supérieure du rappel réel (c'est pour cette raison qu'il est entre parenthèses dans la table 4). Toutefois, un examen manuel d'un certain nombre de phrases extraites aléatoirement de tout le corpus nous a montré qu'il n'arrivait presque jamais qu'une formule chimique non-détectée soit dans une phrase dans laquelle aucune autre formule chimique ne l'a été. Ceci laisse penser que le rappel réel est très proche du score obtenu.

L'évaluation manuelle de la correction/normalisation des formules chimiques a pu se faire directement sur ces 200 phrases, dans la mesure où une modification est proposée pour plus des trois quarts des formules chimiques trouvées.

Type d'entité	**Reconnaissance**			**Correction/normalisation**		
	Précision	Rappel	#occurrences pour 10^6 tokens	Précision	Rappel	Proportion d'entités corrigées parmi les entités détectées
Dates	0,98	0,97	4713	0,96	–	2%
Adresses	0,83	0,86	269	0,76	–	3%
Formules chimiques	0,91	(0,88)	300	0,95	0,90	72%

TABLE 4 – Evaluation de la reconnaissance des entités nommées et de leur correction/normalisation. Les fréquences sont calculées sur le corpus BNF/Gallica, sauf pour les formules chimiques où le corpus utilisé est EPO

La table 4 donne les résultats de notre évaluation. Elle indique également la fréquence de chaque type d'entité dans son corpus d'évaluation. Pour le calcul des scores de reconnaissance, une entité n'est considérée comme correctement reconnue que si son type et ses deux bornes l'ont été : les reconnaissances partielles sont considérées comme des erreurs.

La précision de la correction/normalisation est définie comme le pourcentage d'entités modifiées par nos grammaires telles que le résultat de cette modification est complètement correct. Ainsi, une correction partielle compte comme une erreur. Ceci explique que la précision obtenue pour les adresses soit plus faible que pour les autres types d'entités, puisqu'elles contiennent presque toujours des noms propres auxquels nos corrections ne s'appliquent pas.

Le rappel est le pourcentage d'entités correctement modifiées parmi celles qui n'étaient pas totalement correctes dans la sortie brute d'OCR. Pour le calcul des scores de correction/normalisation des formules chimiques, toutes les entités ont été prises en compte, qu'elles aient été ou non correctement détectées par SXPipe-postOCR. Ceci nous a permis le calcul du rappel. Cependant, la taille des corpus utilisés pour l'évaluation des dates et des adresses nous a contraint à nous limiter pour ces deux types d'entités aux occurrences reconnues par SXPipe-OCR. Le calcul du rappel est alors impossible.

5 Conclusion et perspectives

Nous avons présenté le composant linguistique d'une architecture pour la correction automatique des erreurs dans des documents numériques obtenus par reconnaissance optique de caractères. Notre système, constitué d'une cascade de grammaires locales implémentées dans l'outil SXPipe, permet la détection robuste ainsi que la correction de certains types d'entités nommées fréquemment rencontrées dans des corpus spécialisés. Contrairement à un modèle statistique d'erreurs, notre méthode permet une meilleure prise en charge des propriétés formelles non-linguistiques de ces entités. De plus, les règles de corrections sont spécifiques à chaque type d'entité, parfois à des types de segments au sein des entités (par exemple, le numéro d'arrondissement dans une adresse). L'évaluation effectuée à la main, en s'appuyant sur

notre corpus d'écart de qualité éditoriale, confirme l'utilité de grammaires locales de reconnaissance et de correction au sein d'une chaîne de correction automatique des sorties de systèmes d'OCR.

La prochaine étape de notre travail est de réaliser effectivement l'intégration de SXPipe-postOCR au sein d'une chaîne complète de correction de sorties d'OCR, c'est-à-dire de coupler nos grammaires locales de reconnaissance et de correction avec un modèle d'erreur et un modèle de langage statistiques. Pour cela, nous envisageons d'utiliser SXPipe-postOCR comme un pré-traitement qui viendra en amont du composant reposant sur ces deux types de modèles statistiques.

Les perspectives scientifiques de ce travail sont doubles. Nous prévoyons tout d'abord de réaliser des expériences avec différents modèles de langage exploitant l'information issue de l'étiquetage en entités nommées. Une autre extension possible notre travail serait de faire en sorte que les grammaires locales puissent, en cas de doute, proposer plus d'une correction, charge au modèle de langage (voire au modèle d'erreurs) de choisir la meilleure.

Références

BRILL E. & MOORE R. C. (2000). An Improved Error Model for Noisy Channel Spelling Correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL'00)*, p. 286–293, Hong Kong.

CHINCHOR N. (1998). MUC-7 named entity task definition. In *Seventh Message Understanding Conference (MUC-7)*, Fairfax, États-Unis.

CUCERZAN S. & BRILL E. (2004). Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, p. 293–300, Barcelone, Espagne.

HAWIZY L., JESSOP D. M., ADAMS N. & MURRAY-RUST P. (2011). ChemicalTagger : A tool for semantic text-mining in chemistry. *Journal of Cheminformatics*, **3**(17).

KERNIGHAN M., CHURCH K. & W.A. G. (1990). A spelling correction program based on a noisy channel model. In *Proceedings of the 13th conference on Computational linguistics*, p. 205–210, Helsinki, Finlande.

KLEIN S. T. & KOPE M. (2002). A voting system for automatic OCR correction. In *Proceedings of the Workshop On Information Retrieval and OCR : From Converting Content to Grasping Meaning*, p. 1–21, Tampere, Finlande.

KOLAK O. & RESNIK P. (2002). OCR error correction using a noisy channel model. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT'02)*, p. 257–262, San Diego, États-Unis.

LEVENSHTEIN V. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, **10**, 707–710.

LUND W. B. & RINGGER E. K. (2009). Improving optical character recognition through efficient multiple system alignment. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'09)*, p. 231–240, Austin, États-Unis.

MAYNARD D., TABLAN V., URSU C., CUNNINGHAM H. & WILKS Y. (2001). Named entity recognition from diverse text types. In *In Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP'01)*, p. 257–274, Tzigov Chark, Bulgarie.

REYNAERT M. (2004). Multilingual text induced spelling correction. In *Proceedings of the Workshop on Multilingual Linguistic Ressources (MLR'04)*, p. 117–117.

RINGLSTETTER C., SCHULZ K. U., MIHOV S. & LOUKA K. (2006). The same is not the same—postcorrection of alphabet confusion errors in mixed-alphabet ocr recognition. In *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05)*, p. 406–410, Séoul, Corée du Sud.

ROCKTÄSCHEL T., WEIDLICH M. & LESER U. (2012). Chemspot : a hybrid system for chemical named entity recognition. *Bioinformatics*, **28**.

SAGOT B. & BOULLIER P. (2008). SxPipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues*, **49**(2), 155–188.

SANG E. F. T. K. & MEULDER F. D. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (CONLL'03)*, p. 142–147, Edmonton, Canada.

STROHMAIER C., RINGLSTETTER C., SCHULZ K. & MIHOV S. (2003). Lexical postcorrection of ocr-results : the web as a dynamic secondary dictionary ? In *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03)*, p. 1133–1137, Edinburgh, Royaume-Uni.