TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Évaluer la pertinence de la morphologie constructionnelle dans les systèmes de Question-Réponse

Delphine Bernhard[1] Bruno Cartoni[2] Delphine Tribout[1]
(1) LIMSI-CNRS, 91403 Orsay, France
(2) Département de linguistique, Université de Genève, Suisse
bernhard@limsi.fr, bruno.cartoni@unige.ch, tribout@limsi.fr

Résumé. Les connaissances morphologiques sont fréquemment utilisées en Question-Réponse afin de faciliter l'appariement entre mots de la question et mots du passage contenant la réponse. Il n'existe toutefois pas d'étude qualitative et quantitative sur les phénomènes morphologiques les plus pertinents pour ce cadre applicatif. Dans cet article, nous présentons une analyse détaillée des phénomènes de morphologie constructionnelle permettant de faire le lien entre question et réponse. Pour ce faire, nous avons constitué et annoté un corpus de paires de questions-réponses, qui nous a permis de construire une ressource de référence, utile pour l'évaluation de la couverture de ressources et d'outils d'analyse morphologique. Nous détaillons en particulier les phénomènes de dérivation et de composition et montrons qu'il reste un nombre important de relations morphologiques dérivationnelles pour lesquelles il n'existe pas encore de ressource exploitable pour le français.

Abstract. Morphological knowledge is often used in Question Answering systems to facilitate the matching between question words and words in the passage containing the answer. However, there is no qualitative and quantitative study about morphological phenomena which are most relevant to this application. In this paper, we present a detailed analysis of the constructional morphology phenomena found in question and answer pairs. To this aim, we gathered and annotated a corpus of question and answer pairs. We relied on this corpus to build a gold standard for evaluating the coverage of morphological analysis tools and resources. We detail in particular the phenomena of derivation and composition and show that a significant number of derivational morphological relations are still not covered by any existing resource for the French language.

Mots-clés : Évaluation, Morphologie, Ressources, Système de Question-Réponse.

Keywords: Evaluation, Morphology, Resources, Question-answering system.

1 Introduction

Les systèmes de Question-Réponse (QR) ont pour objectif de fournir une réponse précise à une question. Pour ce faire, ils reposent généralement sur un composant de recherche d'information (RI) qui vise à apparier les mots de la question avec les mots des documents contenant la réponse potentielle. La principale difficulté pour les systèmes de RI réside dans le fait qu'une réponse peut se trouver dans un document qui ne reprend pas forcément les mots de la question. Les systèmes de RI et de QR doivent donc pouvoir récupérer les documents pertinents sans se baser uniquement sur l'identité formelle entre les mots de la question et les mots du document. À cette fin, la morphologie a souvent été préférée à une analyse sémantique plus complexe dans la mesure où deux mots reliés morphologiquement montrent généralement une similitude formelle qui permet de prendre en compte facilement leur relation sémantique. Les systèmes de RI et de QR intègrent donc généralement des connaissances morphologiques, que ce soit lors de l'indexation des documents ou lors de la recherche, en étendant les requêtes ou en les reformulant au moyen de mots morphologiquement reliés. Cette intégration est généralement effectuée de manière très générique, c'est-à-dire que toutes les relations morphologiques possibles, ou pour lesquelles on dispose d'une ressource, sont incluses. Par ailleurs, les évaluations sont effectuées de manière globale, en évaluant l'amélioration de la performance globale du système, et non l'impact de cet ajout.

La plupart des recherches menées dans ce domaine utilisent des techniques de désuffixation (*stemming*) basées sur des heuristiques simples qui suppriment la fin des mots (Lennon *et al.*, 1988; Harman, 1991; Fuller & Zobel,

1998). Ces méthodes s'avèrent efficaces pour les langues à morphologie moins riche comme l'anglais, mais ne sont pas disponibles pour toutes les langues (McNamee *et al.*, 2009). La plupart du temps l'utilisation de ces méthodes permet d'augmenter légèrement le rappel, mais ces techniques génèrent également du bruit. Bilotti *et al.* (2004) ont par exemple montré que des mots relativement éloignés comme *organisation* et *organ* sont réduits à la même racine par l'algorithme de désuffixation de Porter. Moreau & Claveau (2006) ont quant à eux utilisé une méthode d'acquisition automatique de connaissances morphologiques par apprentissage, et leur étude a montré que l'utilisation des connaissances morphologiques pour étendre les requêtes améliore les résultats pour la plupart des langues européennes qu'ils ont testées.

Dans chacune des études précédentes, les mots de la question sont étendus à l'ensemble des mots appartenant à la même famille morphologique, et les différents types de procédés (tels que la formation d'un nom déverbal ou la formation d'un nom désadjectival) ne sont pas distingués. Ainsi, tous les mots appartenant à la même famille morphologique sont considérés comme sémantiquement proches. Or, nous pensons que tous les mots morphologiquement reliés n'ont pas la même proximité sémantique et que certaines relations morphologiques sont plus pertinentes que d'autres dans le cadre de QR. Par exemple dans ce type de tâche, il nous semble plus pertinent d'étendre une requête contenant le verbe *diviser* au nom d'événement dérivé *division*, plutôt qu'à l'adjectif dérivé *divisible*. Cependant, à notre connaissance, aucune étude qualitative ni quantitative n'a été menée en ce sens, afin de déterminer quels types de relations morphologiques sont pertinents pour la recherche en QR.

Au delà de l'extension de requête, la morphologie trouve également sa place dans les méthodes de reformulation automatique des questions, qui visent à traiter des phénomènes de paraphrase entre question et réponse. Ainsi Ravichandran & Hovy (2002) proposent une méthode d'acquisition de patrons de reformulation de surface pour des types précis de questions. Ces patrons incluent entre autres des mots morphologiquement reliés tels que *discover*, *discovery* et *discoverer*, dans le cas d'une question portant sur la personne ayant fait une découverte donnée. Ces patrons sont ensuite utilisés pour extraire la réponse. Une approche similaire est proposée par Lin & Pantel (2001) et Hermjakob *et al.* (2002). Ces travaux ne se focalisent toutefois pas sur la morphologie et ne proposent donc pas d'évaluation spécifique. Seul Jacquemin (2010) évalue l'apport de la mophologie dans ce contexte en utilisant le lexique des verbes français de Dubois & Dubois-Charlier (1997) et les relations de dérivation qu'il contient pour automatiquement reformuler des énoncés, sur la base des relations de dépendance syntaxique.

Dans cet article, nous présentons les résultats d'une évaluation portant sur la *pertinence* des connaissances morphologiques dans un système de QR. Ces résultats permettent d'une part de déterminer quels types de ressources morphologiques sont nécessaires à l'amélioration des systèmes de QR, et d'autre part d'évaluer la couverture des ressources existantes pour une telle tâche.

Pour évaluer la pertinence des connaissances morphologiques dans un système de QR, nous avons tout d'abord constitué un corpus de paires question-passage contenant la réponse à partir de trois collections de données issues de campagnes d'évaluation en QR. Nous avons ensuite annoté ce corpus afin de déterminer quelles sont les relations morphologiques les plus fréquentes qui relient les mots de la question et les mots du passage. Enfin, nous avons analysé les résultats de cette annotation et évalué la couverture des ressources existantes en français pour les procédés morphologiques observés [1].

2 Constitution et annotation d'un corpus de paires question-passage réponse

2.1 Corpus de questions-passages

Nous avons constitué notre corpus de paires question-passage réponse à partir de trois collections de données utilisées pour l'évaluation de systèmes QR : Quæro, EQueR et Conique. Ces trois collections et le corpus qu'elles nous ont permis de constituer sont décrits ci-dessous. La table 1 présente les statistiques concernant ces trois collections.

1. Cet article présente les résultats de l'évaluation uniquement pour des *ressources morphologiques*. Les mêmes données de référence ont été utilisées pour évaluer des *outils d'analyse morphologique* (Bernhard *et al.*, à paraître)

	Quæro	EQueR-Medical	Conique
#Questions	350	200	201
#paires de question-passage	566	394	664
Longueur moyenne de la question	8,8	9,9	11,4
Longueur moyenne du passage réponse	38,5	29,0	92,4

TABLE 1 – Statistiques sur les sous-corpus de questions-réponses utilisés

2.1.1 Quæro

Le corpus français Quæro a été constitué dans le cadre du projet Quæro avec pour objectif d'évaluer des systèmes de QR (Quintard *et al.*, 2010). Le corpus de documents contient 2,5 millions de documents en français extraits de l'Internet et 757 questions, dont 250 pour la campagne de 2008 et 507 pour celle de 2009. La collection de documents a été constituée en prenant les 100 premières pages retournées par le moteur de recherche Exalead pour une série de requêtes trouvées dans les logs du moteur. Quant aux questions, elles ont été rédigées par des francophones, sur la base du contenu des documents pour la campagne 2008, et sur la base du log uniquement pour la campagne de 2009. Trois types de questions ont été formulées : des questions factuelles, des questions booléennes attendant une réponse de type oui/non, et des questions de type liste. Pour notre tâche d'annotation, nous avons constitué des paires question-passage, formées de l'ensemble des questions factuelles et des passages contenant la réponse qui ont été fournis par les systèmes et validés manuellement lors des deux campagnes d'évaluation 2008 et 2009. Nous avons ainsi obtenu 566 paires de questions-passages contenant la réponse, 338 pour la campagne 2008 et 228 pour la campagne 2009.

2.1.2 EQueR-Medical

Les données du corpus EQueR ont été constituées dans le cadre de la campagne d'évaluation EQueR-EVALDA pour les systèmes de question-réponse du français (Ayache *et al.*, 2006). La campagne comprenait deux tâches principales : (i) question-réponse générale sur une collection d'articles de journaux et de rapports sénatoriaux et (ii) question-réponse spécialisée sur une collection de textes médicaux. Pour ces deux tâches, les passages contenant les réponses retournées par les systèmes participants ont été validés manuellement par des spécialistes du domaine. Pour notre étude, seule la partie médicale a été retenue, constituant ainsi un ensemble de 394 paires de questions-passages, pour un total de 200 questions distinctes.

2.1.3 Conique

Le corpus Conique a été constitué dans le but d'étudier les justifications pertinentes pour les réponses des systèmes de QR (Grappy *et al.*, 2010). Les justifications des réponses fournissent un matériel supplémentaire pour l'utilisateur, afin qu'il ou elle puisse faire confiance à la réponse fournie par le système. Le corpus est basé sur un sous-ensemble de 291 questions de la campagne EQueR pour le français (Ayache *et al.*, 2006) et de plusieurs campagne CLEF. Les passages-réponses candidats ont été extraits de la version française de Wikipedia à l'aide d'un système de RI et ont ensuite été annotés par 7 annotateurs. Contrairement aux deux corpus décrits précédemment, les passages-réponses de Conique ne correspondent pas à une sortie d'un système de QR. Le corpus possède donc un taux de rappel extrêmement haut, et est exempt de tout biais inhérent aux systèmes de QR, comme les taux importants de mots identiques entre les questions et le passage. Nous avons pré-traité ce corpus, pour ne conserver que les justifications complètes ou partielles. De plus, nous avons réduit le passage à une longueur de trois phrases. Au total, le corpus constitué à partir de la collection Conique contient 664 paires de question-passage, pour 201 questions distinctes.

2.2 Annotation

Pour chaque paire question-passage réponse, nous avons manuellement annoté les mots de la question et les mots du passage afin de déterminer quels mots sont morphologiquement reliés et par quels types de relations.

Les annotations ont été effectuées par trois annotateurs indépendants [2] au moyen de l'outil d'alignement YAWAT (Germann, 2008). Cet outil a été initialement conçu pour aligner les mots de paires de phrases bilingues pour des campagnes d'évaluation de traduction automatique. Dans cette étude, nous l'avons utilisé pour aligner les mots (ou groupes de mots) dans des paires question-passage réponse, et pour assigner à ces paires de mots une étiquette, parmi les trois types de relations morphologiques suivantes : flexion, dérivation, composition. Les figures 1 à 3 présentent des exemples de paires question-passage impliquant respectivement des relations de flexion, dérivation et composition.

Q : Quand **est né** Philippe d'Orléans ?
R : Philippe d'Orléans **naquit** le 2 août 1674.

Q : Comment un *insuffisant* **rénal** doit-il être suivi ?
R : Du fait du risque de transmission nosocomiale du VHC chez les *insuffisants* **rénaux** hémodialysés et chez les transplantés rénaux, une surveillance annuelle de la sérologie doit être réalisée.

Q : À combien de milliards de dollars s'élève le **déficit** budgétaire américain ?
R : Politique budgétaire. Le PIB des États-Unis s'élève à environ 10 000 milliards de dollars et les **déficits** atteindraient au moins 300 ou 400 milliards de dollars en 2003

FIGURE 1 – Exemples de paires question-passage impliquant une relation de flexion

Q : En quelle année Martin Luther King **a**-t-il **été assassiné** ?
R : Il dit avoir été à côté du pasteur King à Memphis lors de son **assassinat**, le 4 avril 1968. Il est ordonné ministre baptiste à la fin de cette même année.

Q : Quels sont les quatre **réalisateurs** du film "Le jour le plus long" ?
R : Le Jour le plus long (The Longest Day) est un film américain **réalisé** par Ken Annakin, Andrew Marton, Bernhard Wicki et Gerd Oswald sorti en salle en 1962...

Q : La pose d'amalgame dentaire peut-elle provoquer des **allergies** ?
R : Il est certain que la pose d'amalgames peut entraîner des réactions **allergiques** plus ou moins graves et prononcées chez les patients.

FIGURE 2 – Exemples de paires question-passage impliquant une relation de dérivation

Q : Où Marcos fut-il **dictateur** ?
R : Imelda Marcos, le 22 février 2006. Imelda Romualdez Marcos (née le 2 juillet 1929) fut la femme de Ferdinand Marcos, **président-dictateur** des Philippines de 1965 à 1986.

Q : Le mercure est-il un métal **toxique** ?
R : En grande concentration ou lorsque l'exposition est prolongée, le mercure a des effets **neuro-toxiques** connus, principalement dans sa forme organique, soit le méthylmercure

Q : Qu'engendre la corticothérapie sur l'**os** ?
R : Les fractures de l'**ostéoporose** cortisonique surviennent au moins en partie en raison d'une perte osseuse induite par la corticothérapie

FIGURE 3 – Exemples de paires question-passage impliquant une relation de composition

Étant donné que deux mots morphologiquement liés peuvent être reliés par plus d'une relation, des instructions spécifiques ont également été définies. Ainsi, nous n'avons pas annoté les variantes flexionnelles des auxiliaires et des déterminants, dans la mesure où ils sont très fréquents et apportent donc peu d'information sémantique. Nous

2. Co-auteurs de cet article.

avons également décidé de donner la priorité aux relations de dérivation et de composition sur les relations flexionnelles. Par exemple, dans la paire de question-passage présentée à la Figure 4, il y a deux étapes morphologiques entre le nom *Australie* dans la question et l'adjectif féminin *australienne* dans la réponse : la première étape est la dérivation de l'adjectif *australien* à partir du nom propre ; la seconde est la flexion de l'adjectif dérivé au féminin. Dans ce cas, la relation morphologique la plus pertinente est la relation dérivationnelle entre le nom propre et l'adjectif, c'est pourquoi dans un tel cas seule la relation dérivationnelle a été annotée. Enfin, une étiquette spécifique "autre" a été utilisée pour annoter des mots qui ne sont par directement reliés morphologiquement, mais qui sont le résultat de deux constructions à partir de la même base. Par exemple dans la paire présentée à la Figure 5 le nom *utilité* et le verbe *utiliser* (sous la forme passive *est utilisée*) ne dérivent pas l'un de l'autre, mais sont tous deux dérivés de l'adjectif *utile*.

Q : Quelle est la capitale de l'**Australie** ?
R : le territoire sur lequel est située la capitale fédérale **australienne**, Canberra .

FIGURE 4 – Exemple de paires question-passage où deux types de relation sont présentes (flexion et dérivation)

Q : Quelle est l'**utilité** de la pierre d'alun ?
R : La pierre d'alun est un excellent déodorant corporel. Elle **est utilisée** pour neutraliser la transpiration, empêcher la fermentation et éliminer les mauvaises odeurs.

FIGURE 5 – Exemple de paires question-passage où la relation morphologique implique deux constructions différentes à partir de la même base

Nous avons mesuré la qualité des annotations à l'aide du coefficient kappa de Fleiss (Fleiss, 1971) [3]. Le coefficient κ varie en fonction du corpus et du type de relation considéré : il est fort (0,674) à presque parfait (0,83) pour la flexion, bon pour la dérivation (0,662 à 0,729) et faible (0,39) à bon (0,665) pour la composition.

Tous les désaccords ont été résolus et les données validées par l'ensemble des annotateurs afin de constituer un corpus de référence. Nous avons ensuite classé et caractérisé les paires de mots que nous avions considérés comme morphologiquement reliés lors de l'annotation. Les résultats de cette analyse sont présentés dans la section suivante et fournissent un panorama des relations morphologiques mises en jeu dans le cadre de Question-Réponse.

3 Analyse des résultats

Au terme de l'annotation, nous avons obtenu un ensemble de mots morphologiquement reliés jouant un rôle dans le lien opéré entre la question et le passage contenant la réponse. Plusieurs observations peuvent être faites suivant différents points de vue. Nous présentons tout d'abord la répartition des différentes relations morphologiques observées (flexion, dérivation et composition). Puis nous décrivons précisément les procédés de dérivation et de composition les plus fréquemment observés. Enfin, nous étudions la position du mot construit dans la paire et en particulier s'il se trouve dans la question ou dans le passage réponse.

3.1 Types de relations morphologiques

Les résultats de l'annotation de chaque sous-corpus en fonction des différents types de relations morphologiques sont présentés dans la table 2 [4]. Ces chiffres montrent que chaque sous-corpus semble favoriser un type particulier de relation morphologique : le sous-corpus Conique contient un majorité de relations de dérivation, le sous-corpus Quæro contient davantage de flexion, alors que pour le sous-corpus EQueR, c'est la composition qui semble la plus fréquente. De plus, si l'on étudie les relations morphologiques en fonction des sous-corpus, on constate que la composition est quasiment absente des sous-corpus Quæro et Conique.

3. Le kappa de Fleiss permet de mesurer l'accord inter-annotateurs lorsqu'il y a plus de deux annotateurs. Il a été calculé en fonction de l'accord des annotateurs sur la présence d'une paire de mots morphologiquement reliés pour une même paire de questions-réponse

4. Les paires de question-passage (paire qp) ne contiennent pas toujours des relations morphologiques, et certaines paires peuvent contenir plus d'une relation morphologique, impliquant parfois les mêmes mots.

	Flexion		Dérivation		Composition	
Corpus (paire qp)	nbr	%	nbr	%	nbr	%
Conique (664)	159	41,8	188	**49,5**	33	8,7
Quæro (566)	136	**61,8**	80	36,4	4	1,8
EQueR (394)	69	26,5	81	31,0	111	**42,5**

TABLE 2 – Flexion, dérivation et composition dans les trois sous-corpus

Il est notable que le sous-corpus Conique contient plus de relations de dérivation que le sous-corpus Quæro. Ceci est lié à la manière dont le corpus Conique a été construit. En effet, il n'a pas été constitué à partir des réponses fournies par un système de QR, mais sur la base de réponses identifiées et annotées manuellement. De plus, Conique contient en moyenne les passages les plus longs (c.f. table 1), ce qui peut expliquer la présence d'un nombre plus important de paires dérivationnelles dans ce sous-corpus. Quant à EQueR, la proportion importante de mots composés est liée au domaine de spécialité du sous-corpus qui contient un grand nombre de termes médicaux, ceux-ci étant souvent composés, comme le montre la figure 6.

Q : Quelle est la conséquence de la **corticothérapie** sur l'*os* ?
R : Le problème essentiel des **corticoïdes** réside dans leurs effets secondaires (... *ostéoporose*, *ostéonécrose* aseptique des têtes fémorales ou parfois humérales ...).

FIGURE 6 – Exemple de paire de question-passage de EQueR

Il est également intéressant de noter le rôle important de la dérivation dans les trois sous-corpus (entre 31 % et 49 %), et l'importance de la composition dans le domaine médical (42,5% dans EQueR). Ceci confirme l'intérêt d'inclure des connaissances morphologiques de ce type dans un système de QR.

Dans la suite, nous décrivons plus précisément ces deux types de construction, en analysant quels procédés morphologiques sont les plus présents. Nous ne nous attardons pas sur la morphologie flexionnelle dans le cadre de cet article car elle est considérée comme pertinente par les systèmes de QR existants. De plus, elle est généralement bien prise en compte par les systèmes de QR existant, notamment *via* l'utilisation de lemmatiseurs.

3.2 Dérivation

Comme nous venons de le montrer (table 2), la dérivation joue un rôle important dans les trois sous-corpus. Dans certain cas, la relation morphologique entre le mot de la question et le mot du passage-réponse implique plus d'une étape dérivationnelle : soit l'un des mots est plus complexe que l'autre, mais n'est pas un dérivé direct (par exemple, *lune* et *alunissage*, ce dernier dérivant du verbe *alunir*, lui-même dérivé de *lune*) ; soit les deux mots sont complexes et sont tous deux dérivés d'un même mot (par exemple *joueur* et *jouable* tous deux dérivés du verbe *jouer*). Le premier cas de figure représente 1,70% des relations observées et le second 8,30%, ce qui représente une proportion assez faible et indique que dans la grande majorité des cas les mots morphologiquement reliés entretiennent une relation de dérivation directe.

Les relations non directes impliquent une prédictibilité moindre, et influencent le choix des méthodes d'implémentation, comme nous l'expliquons dans la section 3.4. Dans un premier temps nous étudions donc uniquement les paires de mots morphologiquement reliés par une dérivation directe.

La table 3, qui présente la proportion des différents types de procédés morphologiques observés, montre que les sous-corpus diffèrent selon les procédés de dérivation majoritairement utilisés. Si Conique contient une majorité d'adjectifs dénominaux (environ 47% des procédés de dérivation), Quæro et EQueR montrent une préférence pour des procédés de nominalisation (avec respectivement 61% et 54% des procédés de dérivation).

	Exemple	Conique (174)		Quæro (70)		EQueR (70)	
		nbr	%	nbr	%	nbr	%
Nom > Adj	commerce > commercial	37	21	16	23	**28**	**40**
Nom propre> Adj	Afrique > africain	**45**	**26**	8	11,5	1	1
Nom > Nom	président > présidente	29	17	5	7	2	3
Nom propre > N	Arménie > Arméniens	6	3	8	11,5	2	3
Adj > Nom	national > nationalité	3	2	0	0	9	13
Verbe > Nom	traiter > traitement	41	24	**30**	**43**	25	36
Autres	complet > complètement	13	7	3	4	3	4

TABLE 3 – Procédés dérivationnels dans les paires de question-passage

3.2.1 Adjectifs dénominaux

Dans les trois sous-corpus, les adjectifs dérivés d'un nom propre sont toujours des adjectifs relationnels, qui peuvent être remplacés par un complément du nom équivalent, comme *chilien* dérivé de *Chili* ou *africain* dérivé de *Afrique*. Les adjectifs dérivés d'un nom commun sont la plupart du temps relationnels, comme le montrent les chiffres de la table 4. Par exemple, *commercial* dérivé de *commerce* ou *solaire* dérivé de *soleil*. Cependant, on trouve également quelques adjectifs qualificatifs comme *âgé* dérivé de *âge* ou *montagneux* dérivé de *montagne*. La table 4 présente les proportions d'adjectifs relationnels et qualificatifs dans notre corpus, et montre que ce sont les adjectifs relationnels qui sont les plus fréquents dans les trois sous-corpus.

	Adj. relationel		Adj. qualificatif	
	nbr	%	nbr	%
Conique (37)	23	62	14	38
Quæro (16)	10	62	6	38
EQueR (28)	24	86	4	14
Total (81)	57	70	24	30

TABLE 4 – Types d'adjectifs dénominaux

3.2.2 Procédés de formation des noms

En ce qui concerne les procédés de formation de noms, on trouve dans les trois sous-corpus un grand nombre de nominalisations déverbales, ainsi que quelques cas de nominalisations désadjectivales ou dénominales. La formation de noms sur des bases nominales est relativement rare, sauf dans Conique qui contient plusieurs noms de profession féminisés, comme *infirmier* et *infirmière*, *directeur* et *directrice*, *président* et *présidente*, que nous avons considérés comme dérivés l'un de l'autre et non comme deux formes fléchies du même mot. Nous avons également trouvé quelques noms diminutifs, comme *rame > ramette* et quelques préfixés comme *président > vice-président*. Nous avons aussi considéré les formations de noms à partir de noms propres comme des nominalisations. Ces noms dérivés sont principalement des gentilés comme *Colombien* dérivé de *Colombie*.

Les noms désadjectivaux sont rares dans les trois sous-corpus, voire inexistants dans le cas de Quæro. Ces noms désadjectivaux sont principalement des noms de propriété, comme *toxicité* construit sur *toxique*. La plupart des noms désadjectivaux se trouvent dans le corpus EQueR. Cela s'explique par le fait que le corpus médical contient beaucoup de noms de maladie ou de pathologie (comme *toxicité* ou *insuffisance*), or ces noms réfèrent la plupart du temps à la propriété de se trouver dans un état particulier (*toxicité* $\approx$ 'propriété d'être toxique', *insuffisance* $\approx$ 'propriété d'être insuffisant').

Quant aux noms déverbaux, qui sont les plus fréquents, ce sont essentiellement des noms d'événements, comme *débarquement* dérivé du verbe *débarquer*. Les noms d'événements représentent presque 85% des noms déverbaux, comme le montre la table 5. Cependant, on trouve également un petit nombre de noms d'agents dans les sous-corpus Conique et Quæro, comme *réalisateur* construit sur *réaliser*, et quelques cas de noms résultatifs comme *produit* dérivé de *produire*.

	Exemple	Conique (41)		Quæro (30)		EQueR (25)	
		nbr	%	nbr	%	nbr	%
Verbe > N événement	inaugurer > inauguration	**34**	**83**	**25**	**83**	**22**	**88**
Verbe > N agent	réaliser > réalisateur	4	10	4	13	0	0
Verbe > N autre	produire > produit	3	7	1	4	3	12

TABLE 5 – Types sémantiques des noms déverbaux dans les paires de questions-passage

3.2.3 Autres procédés de formation

Parmi les autres procédés de formation observés dans le corpus, on trouve des formations d'adverbes, comme *complètement* dérivé de *complet* ou *directement* construit sur *direct*, ainsi qu'un certain nombre de verbes préfixés, comme *déboucher*, ou d'adjectifs préfixés, comme *international*. Il est également intéressant de noter que nous n'avons observé aucun verbe désadjectival (comme *national* > *nationaliser*) et très peu de verbes dénominaux (quatre cas seulement dans le sous-corpus Conique, dont trois sont des verbes convertis : *border*, *fusionner* et *suicider*). La quasi-absence de verbes dénominaux peut s'expliquer par la faible prédictibilité du sens d'un verbe dénominal. Comme l'ont décrit Hopper & Thompson (1984), il existe une asymétrie entre les catégories lexicales, dans la mesure où un nom déverbal continue de référer à l'événement dénoté par le verbe base, alors qu'un verbe dénominal ne réfère pas à l'entité dénotée par le nom base, mais dénote un événement associé à cette entité. Or, les événements associés à une entité peuvent être nombreux et variés. Par exemple, le nom *destruction* dénote le même événement que sa base verbale *détruire*, alors qu'un verbe dénominal comme *hospitaliser* ne réfère pas à l'objet dénoté par la base nominale (*hôpital*), mais à l'un des événements liés à cet objet. Ainsi, dans le cadre d'un système de QR, la relation sémantique entre un nom et son verbe dérivé est moins informative que la relation entre un verbe et son nom dérivé.

3.3 Composition

En ce qui concerne la composition, comme nous l'avons vu dans la section 3.1 et la table 2, elle est surtout présente dans les sous-corpus EQueR et Conique, mais quasiment absente du sous-corpus Quæro. Dans notre analyse de la composition nous avons distingué les paires de question-passage contenant un composé et au moins l'un de ses constituants (comme dans *filmographie* composé de *film*), des paires contenant deux composés qui partagent un même constituant (comme *aéronautique* et *aéroport* partageant le constituant *aéro*). La table 6 présente les résultats de cette classification, et montre que le second cas (deux composés partageant un constituant) est avant tout présent dans le corpus spécialisé.

	composé-constituant(s)		2 composés	
	nbr	%	nbr	%
Conique (33)	**26**	**79**	7	21
Quæro (4)	**4**	**100**	0	0
EQueR (111)	**70**	**63**	41	37

TABLE 6 – Types de relations de composition

3.4 Conclusion sur les relations morphologiques observées dans les trois sous-corpus

Comme le montre l'analyse des résultats de notre annotation, la morphologie joue un rôle important pour établir le lien de similarité entre question et passage-réponse. D'après notre étude qualitative des relations morphologiques observées, la flexion est loin d'être la seule connaissance morphologique présente dans notre corpus, et la dérivation, tout comme la composition, jouent un rôle important. Notons également que les types de procédés morphologiques employés sont très similaires dans les corpus de langue générale, alors que la langue de spécialité – EQueR, corpus médical – montre des tendances nettement différentes.

Nous avons également étudié la position privilégiée du mot le plus complexe morphologiquement, afin se savoir s'il se trouve plutôt dans la question ou dans le passage réponse. La prédominance de l'une ou l'autre position

joue un rôle essentiel dans la manière de gérer la morphologie dans les systèmes de QR. Pour les paires impliquant une relation de dérivation, le mot complexe se trouve majoritairement dans le passage réponse (52% des cas dans EQueR, 59% dans Quæro et 65% dans Conique). Ce résultat confirme l'intérêt de l'expansion de requête aux mots dérivés des mots de la question.

De plus, le nombre important de relations de composition où les deux membres de la paire sont des composés partageant un même constituant (37% dans EQueR et 21% dans Conique) pointe les limites de l'apport de la morphologie dans de tels systèmes, étant donné qu'il est difficilement envisageable de vouloir générer tous les mots composés à partir d'un élément de la question.

Dans la suite de cet article, nous utilisons les résultats de l'annotation comme gold-standard (ensemble de référence) pour évaluer les ressources morphologiques existantes en français. Cette évaluation permet donc implicitement de connaître l'impact de telles ressources si elles étaient intégrées dans un système de QR. À noter également qu'en plus des ressources statiques rendant compte des relations morphologiques, il existe également des outils d'analyse, à base de règles ou d'heuristiques comme Dérif (Namer, 2009) [5].

4 Evaluation des ressources morphologiques existantes

En français, il n'existe pas de ressources contenant des relations morphologiques dérivationnelles à large échelle similaire à la base CELEX qui contient un nombre important d'informations morphologiques pour le néerlandais, l'anglais et l'allemand (Baayen *et al.*, 1995). Il existe pour le moment uniquement des ressources conçues pour traiter d'un phénomène morphologique particulier. Dans le cadre de notre étude, nous nous sommes intéressés à trois ressources qui couvrent des phénomènes morphologiques particulièrement fréquents dans notre corpus : les noms déverbaux et les adjectifs dénominaux. Ces ressources sont VerbAction, Dubois et Prolexbase. Nous laissons de côté dans le cadre de cette étude la couverture des procédés de composition, pour lesquels il n'existe pas de ressource.

4.1 Présentation des ressources existantes

Verbaction [6] est une ressource lexicale regroupant tous les noms d'actions dérivés d'un verbe (Hathout *et al.*, 2002; Hathout & Tanguy, 2002). Elle contient un total de 9 393 paires de nom-verbe.

Dubois [7] Cette ressource XML, constituée à partir du travail de (Dubois & Dubois-Charlier, 1997) est une description des verbes français regroupés en classes syntaxico-sémantiques, qui fournit également des informations sur les dérivés de ces verbes. Elle contient au total 25 609 entrées pour lesquels elle mentionne 33 955 dérivés.

Prolexbase [8] est un dictionnaire multilingue de noms propres (Tran & Maurel, 2006; Bouchou & Maurel, 2008). Bien qu'elle ne contienne pas explicitement de connaissances morphologiques, cette ressource fournit des informations sur les noms relationnels et les adjectifs associés aux noms propres. Par exemple, *Français* et *français* sont explicitement associés à l'entrée *France*. Au total, Prolexbase contient 76 118 lemmes et 20 614 relations dérivationnelles.

4.2 Résultats

L'évaluation de ces trois ressources dérivationnelles ne peut se faire sur l'ensemble du gold-standard dans la mesure où chacune d'elles a été conçue pour couvrir un phénomène morphologique spécifique. De ce fait nous avons évalué les ressources uniquement sur la partie du gold-standard concernée par le phénomène pour lequel

5. À ce sujet voir Bernhard *et al.* (à paraître).
6. http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=hathout&subURL=verbaction/main.html
7. http://rali.iro.umontreal.ca/Dubois/
8. http://www.cnrtl.fr/lexiques/prolex/

elles ont été conçues. La couverture de VerbAction et de Prolexbase a été évaluée en comptant le nombre de paires de mots morphologiquement reliés qui s'y trouvent. Dubois, en revanche, ne contient pas les dérivés mais mentionne uniquement leur existence et fournit des informations permettant de déduire la forme du dérivé. Pour évaluer Dubois nous avons donc pris en compte les cas où le dérivé pouvait être automatiquement calculé à partir des informations fournies.

La table 7 résume la couverture de VerbAction et de Dubois pour les noms d'événement observés dans notre corpus. La couverture de VerbAction est meilleure que celle de Dubois, en particulier pour les sous-corpus de langue générale Conique et Quæro. Quant aux noms déverbaux agentifs, Dubois couvre 100% des noms de Conique et 75% de ceux de Quæro (aucun nom agentif n'a été trouvé dans le sous-corpus EQueR). VerbAction est limité aux noms d'action et ne contient donc aucun nom d'agent.

	VerbAction		**Dubois**	
Corpus (nbr.)	nbr.	%	nbr.	%
Conique (34)	33	97	19	56
Quæro (25)	25	100	9	36
EQueR (22)	22	100	19	86
Total (81)	**80**	**99**	**47**	**58**

TABLE 7 – Couverture des ressources pour les noms d'événements déverbaux

Pour ce qui est des gentilés et des adjectifs relationnels dérivés de noms géographiques, les résultats de l'évaluation de Prolexbase sont présentés dans la table 8. Nous distinguons les gentilés (habitants d'un lieu), les adjectifs relationnels, et les noms de lieu ou d'entité institutionnelle que nous avons appelés "LocOrg". Les chiffres montrent que Prolexbase a une très bonne couverture pour les gentilés dérivés d'un nom de lieu, et pour les adjectifs relationnels dérivés d'un nom de lieu ou d'un gentilé [9].

Les ressources existantes VerbAction, Dubois et Prolexbase offrent donc une bonne couverture des noms déverbaux et des gentilés et adjectifs dérivés de noms propres. Cependant, si l'on évalue ces trois ressources sur l'ensemble des relations dérivationnelles observées dans le corpus, le taux de couverture global est relativement faible (environ 52%). Ce faible taux de couverture n'est pas étonnant dans la mesure où les ressources évaluées sont conçues pour des phénomènes particuliers. Cela démontre également qu'il reste un nombre important de relations morphologiques dérivationnelles pour lesquelles il n'existe pas encore de ressource exploitable. En premier lieu, il manque une ressource associant les adjectifs dénominaux aux noms dont ils dérivent, lorsqu'il ne s'agit pas de noms géographiques. Or, ce type de relation dérivationnelle est une des plus fréquentes dans notre corpus puisqu'elle concerne environ 21% des paires de mots reliés par un procédé dérivationnel dans le sous-corpus Conique, 23% dans le sous-corpus Quæro, et 40% dans le sous-corpus EQueR (cf. table 3). Une telle ressource spécifiant la relation entre un adjectif dénominal et son nom base permettrait donc d'augmenter de façon significative la couverture globale des ressources morphologiques du français pour les relations morphologiques observées dans notre corpus de question-passage.

5 Conclusion et perspectives

Nous avons présenté une étude détaillée des phénomènes morphologiques permettant de faire le lien entre question et réponse dans le cadre des systèmes de QR. Pour réaliser cette étude, nous avons constitué un corpus de paires de question-passage réponse à partir de divers corpus utilisés pour l'évaluation en QR. Nous avons réalisé une annotation détaillée du corpus, portant sur les liens morphologiques entre question et réponse. Cette annotation nous a permis d'obtenir des données de référence, que nous avons analysées de manière détaillée selon plusieurs axes : types de relations morphologiques, procédés dérivationnels utilisés, relations de composition. Cette analyse nous a permis de tirer les conclusions suivantes : (i) la morphologie dérivationnelle constitue une forte proportion des phénomènes morphologiques à l'œuvre dans le corpus, (ii) les phénomènes de dérivation observés concernent essentiellement les adjectifs dénominaux et les nominalisations verbales, (iii) le procédé de composition s'observe essentiellement dans le sous-corpus spécialisé de la langue médicale EQueR.

9. Dans le corpus Quæro, aucune paire gentilé>adjectif relationnel n'a été trouvée, et dans le corpus EQueR, seule une paire LocOrg>adjectif relationnel a été trouvée et est analysée correctement.

Corpus	**Relation morphologique (nbr.)**		**Trouvé dans Prolexbase**	
			nbr.	%
Conique	Gentilé - Adj. Rel	(1)	1	100
	LocOrg - gentilé	(6)	6	100
	LocOrg - Adj. rel	(45)	43	96
Quæro	LocOrg - Gentilé	(8)	5	62
	LocOrg - Adj. rel.	(8)	8	100
EQueR	LocOrg - Adj. rel.	(1)	1	100
Total		**69**	**64**	**93**

TABLE 8 – Couverture de Prolexbase pour les relations morphologiques de type "géographique"

Si ces résultats soulignent l'importance de la morphologie dans l'appariement en question-réponse, et montrent clairement quelles relations sont les plus pertinentes (car les plus fréquentes), ils ne permettent pas d'évaluer l'impact, notamment en termes de bruit, de la prise en compte de ces relations, fréquentes ou non, dans un système de QR. Une intégration modulaire de chacune des relations, et une évaluation précise de leur impact sur les résultats d'un système de QR permettraient sans doute d'avoir une meilleure idée sur la question.

Nous avons également évalué la couverture de ressources morphologiques existantes pour le français par rapport aux phénomènes observés. Si certains procédés bénéficient d'une très bonne couverture (noms d'événement dans VerbAction, gentilés et adjectifs relationnels dérivés de noms géographiques dans Prolexbase), d'autres souffrent d'un manque de ressource adaptée, comme par exemple les adjectifs dénominaux.

Les perspectives de ces travaux sont multiples. D'un point de vue linguistique, l'observation des relations morphologiques les plus fréquentes, et l'absence constatée de certaines autres, semblent indiquer que certains types de relations morphologiques sont plus informatifs et donc plus pertinents que d'autres. D'un point de vue applicatif cette hypothèse mériterait néanmoins d'être évaluée empiriquement. L'analyse a également permis de distinguer les procédés dérivationnels à intégrer de façon prioritaire dans les systèmes de QR. Nous envisageons d'intégrer ces observations dans un système de QR existant, en définissant notamment des patrons de reformulation de question basés sur la morphologie.

Remerciements Ces travaux ont été partiellement financés par OSEO dans le cadre du programme QUAERO.

Références

AYACHE C., GRAU B. & VILNAT A. (2006). EQueR : the French Evaluation campaign of Question-Answering Systems. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*, p. 1157–1160.

BAAYEN R. H., PIEPENBROCK R. & GULIKERS L. (1995). *The Celex Lexical Database (Release 2) [CD-ROM]*. Philadelphia, PA : Linguistic Data Consortium.

BERNHARD D., CARTONI B. & TRIBOUT D. (À paraître). A Task-Based Evaluation of French Morphological Resources and Tools : A Case Study for Question-Answer pairs. *Linguistic Issues in Language Technology - LiLT*.

BILOTTI M. W., KATZ B. & LIN J. (2004). What Works Better for Question Answering : Stemming or Morphological Query Expansion. In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, Sheffield, England.

BOUCHOU B. & MAUREL D. (2008). Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres. *Traitement Automatique des Langues*, **49**(1), 61–88.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). *Les verbes français*. Larousse-Bordas.

FLEISS J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**(5), 378–382.

FULLER M. & ZOBEL J. (1998). Conflation-based comparison of stemming algorithms. In *Proceedings of the Third Australian Document Computing Symposium*, p. 8–13, Sydney.

GERMANN U. (2008). Yawat : yet another word alignment tool. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (HLT '08)*, p. 20–23.

GRAPPY A., GRAU B., FERRET O., GROUIN C., MORICEAU V., ROBBA I., TANNIER X., VILNAT A. & BARBIER V. (2010). A Corpus for Studying Full Answer Justification. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

HARMAN D. (1991). How effective is suffixing ? *Journal of the American Society of Information Science*, **42**(1), 7–15.

HATHOUT N., NAMER F. & DAL G. (2002). Many Morphologies. chapter An Experimental Constructional Database : The MorTAL Project, p. 178–209. Cascadilla Press.

HATHOUT N. & TANGUY L. (2002). Webaffix : Discovering Morphological Links on the WWW. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, p. 1799–1804, Las Palmas de Gran Canaria, Espagne : ELRA.

HERMJAKOB U., ECHIHABI A. & MARCU D. (2002). Natural Language Based Reformulation Resource and Wide Exploitation for Question Answering. In *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*.

HOPPER P. & THOMPSON S. (1984). The discourse basis for lexical categories in universal grammar. *Language*, **60**, 703–752.

JACQUEMIN B. (2010). A Derivational Rephrasing Experiment for Question Answering. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

LENNON M., PIERCE D. S., TARRY B. D. & WILLETT P. (1988). An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science*, **3**(4), 177–183.

LIN D. & PANTEL P. (2001). Discovery of Inference Rules for Question Answering. *Natural Language Engineering*, **7**(4), 343–360.

MCNAMEE P., NICHOLAS C. & MAYFIELD J. (2009). Addressing morphological variation in alphabetic languages. In *SIGIR '09 : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, p. 75–82, New York, NY, USA : ACM.

MOREAU F. & CLAVEAU V. (2006). Extension de requêtes par relations morphologiques acquises automatiquement. In *Actes de la Troisième Conférence en Recherche d'Informations et Applications CORIA 2006*, p. 181–192.

NAMER F. (2009). *Morphologie, Lexique et TAL : l'analyseur DériF*. TIC et Sciences cognitives. London : Hermes Sciences Publishing.

QUINTARD L., GALIBERT O., ADDA G., GRAU B., LAURENT D., MORICEAU V., ROSSET S., TANNIER X. & VILNAT A. (2010). Question Answering on Web Data : The QA Evaluation in Quæro. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

RAVICHANDRAN D. & HOVY E. (2002). Learning surface text patterns for a Question Answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, p. 41–47.

TRAN M. & MAUREL D. (2006). Prolexbase : un dictionnaire relationnel multilingue de noms propres. *Traitement Automatique des Langues*, **47**(1), 115–139.