Acquisition de ressources pour la simplification de textes médicaux

Natalia Grabar

STL CNRS UMR8163

CENTAL, Louvain La Neuve: 20/10/2017

Présentation

Spécialité — Traitement Automatique des Langues

Informatique Médicale, TAL médical

Health on the Net (HON) à Genève, Suisse

Assistante Hospitalo-Universtaire (AHU) à l'HEGP
INSERM

Depuis 2010 : CR1 CNRS, Lille, UMR8163 STL

Plan

Contexte

- Domaine biomédical :
  - différents types d'utilisateurs
  - experts, patients, pharmaciens, étudiants ...
  - différents niveaux de spécialisation
- Patients : qualité des informations, compréhension
  - Qualité médicale des informations :
    - HONcode éthique [Boyer *et al.*, 1997] : certification des sites de santé
    - autorité, complémentarité, confidentialité, attribution, justification, transparence de financement ...
    - [Risk & Dzenowagis, 2001] : *Review of Internet information quality initiatives*
  - Comfort visuel
  - *Spécialisation technique et scientifique*
  - ...

$\Rightarrow$ Relation directe avec la vie et le bien-être des personnes

- [AMA, 1999, Berland *et al.*, 2001, McCray, 2005, Tran *et al.*, 2009]

Lisibilité des documents de spécialité

documents de santé

- Facilité à lire, comprendre et utiliser les informations de santé
- Dans différents contextes :
  - suivre les instructions de traitement
  - prendre les decisions (maladies chroniques)
  - communiquer avec les médecins
  - réussir le processus de soins
- La difficulté est réelle :
  - compréhension des différentes étapes pour la bonne administration de médicaments [Patel *et al.*, 2002]
  - cohorte de 2 600 patients américains (2 hôpitaux) :
    - entre 26 % et 60 % ne peuvent pas comprendre les instructions sur les médicaments, les consensus informés, les brochures de santé [Williams *et al.*, 1995]
- Documents, sites web de santé à destination des patients :
  - montrent souvent des niveaux de spécialisation élevés [Berland *et al.*, 2001]

ETP : éducation thérapeutique des patients

- Objectif [Golay *et al.*, 2007, Glasgow *et al.*, 2012] :
  - répondre aux priorités politiques de la santé publique et domaine médical
- Aider les patients avec des pathologies chroniques :
  - acquérir et maintenir le savoir-faire
  - mieux gérer la maladie au quotidien
- Aider les professionels médicaux [d'Ivernois *et al.*, 2011, Gross & Gagnayre, 2013, Brin-Henry, 2014] :
  - mieux communiquer avec les patients
  - mieux guider les patients dans leurs parcours médical
- Établir une confiance mutuelle [Sørensen, 1996]
- Améliorer l'efficacité des soins médicaux

Ce que cela donne du côté des patients... [Guilbert, 2014]

- *Docteur, j'ai une hernie fiscale*
  → *...hernie discale*
- *Docteur, j'ai une fuite mistrale*
  → *...fuite mitrale*
- *J'ai dû subir une enculoscopie*
  → *...coloscopie*
- *J'ai fait un coma idyllique*
  → *...coma éthylique*
- *J'ai consulté un gastro-entéropode*
  → *...gastro-entérologue*
- *On m'a fait 3 points de soudure*
  → *...suture*
- *J'ai entendu à la radio que vous pouviez me donner des gélules souches*
- *J'ai une augmentation des triglicirilіques*
- *Faut m'opérer du corps vitreux*

Objectifs

- Rendre les termes plus facilement compréhensibles par les utilisateurs non spécialistes
  - proposer des paraphrases grand public pour les termes techniques
  - exploiter des corpus non spécialisés ou moyennement spécialisés

Travaux existants

Langue générale

- Révision d'articles de Simple wikipedia [Yatskar *et al.*, 2010] :
  - modèle probabiliste + filtres
  - entre 1 079 et 2 970 paires :
    {*stands for*, *is the same as*}, {*indigenous*, *native*}
  - précision : entre 17 % et 86 % ;
- Méthodes issues de la traduction automatique
  [Zhu *et al.*, 2010, Wubben *et al.*, 2012] :
  - corpus parallèles et alignés (Wikipedia/Simple Wikipedia)
- Méthodes distributionnelles
  [Glavas & Stajner, 2015, Kim *et al.*, 2016] :
  - corpus monolingues
  - vecteurs peuvent contenir des équivalents plus simples
  - filtrage

Travaux existants

Langue médicale

- Traducteur automatique de termes médicaux vers des expressions profanes et inversement [McCray *et al.*, 1999] :
  - MEDLINE*plus*
- *Consumer Health Vocabulary* (CHV) [Zeng & Tse, 2006]
  - approche collaborative
- Variations morpho-syntaxiques [Deléger & Zweigenbaum, 2008, Cartoni & Deléger, 2011] :
  - {*consommation régulière*, *consommer de façon régulière*}
    {*gêne à la lecture*, *empêche de lire*}
- Particularités du langage des non experts [Tapi Nzali *et al.*, 2015] :
  - fautes d'orthographe ou d'accent
    {*cirrhose*, *cyrose*}, {*métastase*, *metastase*}
  - formes abrégées
    {*oncologue*, *onco*}, {*chimiothérapie*, *chimio*}

Travaux existants

- Remplacement mot à mot [Biran *et al.*, 2011] :
  - In 1900, Omaha was the center of a national uproar over the kidnapping of Edward Cudahy, Jr., the son of a local meatpacking **magnate**.
  - In 1900, Omaha was the center of a national uproar over the kidnapping of Edward Cudahy, Jr., the son of a local meatpacking **businessman**.
  - {*magnate*, *king*}, {*magnate*, *businessman*}

Travaux existants

- Compétition *SemEval 2012* [Specia *et al.*, 2012]
- Pour un texte court et un mot cible, et plusieurs substitutions possibles pour ce mot et satisfaisant le contexte, l'objectif était de trier ces substitutions selon leur degré de simplicité
- *Hitler committed terrible atrocities during the second World War.*
- candidats/synonyms : abomination, cruelty, enormity, violation
- bon choix : cruelty
- *Hitler committed terrible cruelties during the second World War.*

Rationale

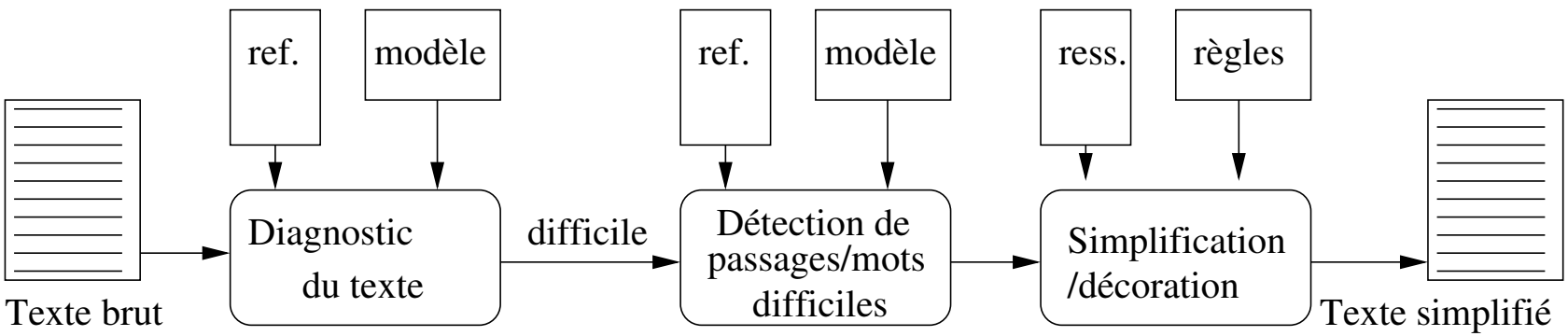
Constatations :

- Peu de travaux sur la langue médicale
- Ressources non disponibles
- Ressources spécifiques :
  - {*otite*, *inflammation de l'oreille*}
  - {*desmorrhexie*, *rupture des ligaments*}
  - {*lombalgie*, *douleurs lombaires*}

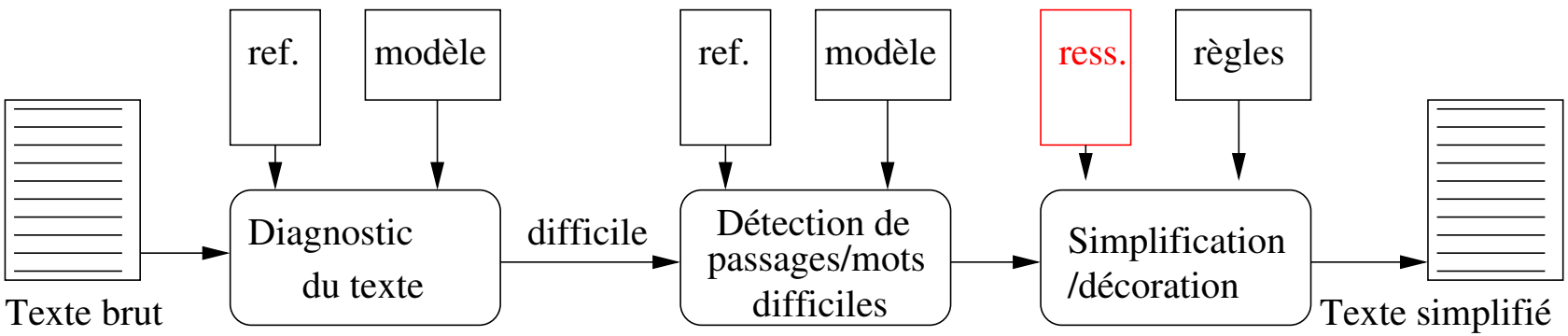
Hypothèses :

- Différentes méthodes et approches sont complémentaires
- Corpus non spécialisés : contiennent les informations recherchées

Rationale


ref.
modèle
ref.
modèle
ress.
règles
Texte brut
Diagnostic du texte
difficile
Détection de passages/mots difficiles
Simplification /décoration
Texte simplifié

Rationale


ref.
modèle
ref.
modèle
ress.
règles
Texte brut
Diagnostic du texte
difficile
Détection de passages/mots difficiles
Simplification /décoration
Texte simplifié

Matériel

- Termes :
  - Snomed International [Côté *et al.*, 1993], partie française d'UMLS [Lindberg *et al.*, 1993]
    - mots des termes
    - pas de nombres
- Corpus :
  - Forum *masante*
    - 6139 réponses
    - 315 362 occurrences
  - Wikipédia, Portail de la Médecine
    - version de janvier 2015
    - 18 434 articles
    - 15 235 219 occurrences

Matériel

- Ressources linguistiques :
  - Ressources morphologiques (155 468)
    - {*aorte*, *aortique*}, {*aortique*, *aortiques*}
  - Ressources de synonymes (4 914)
    - {*embolie*, *thrombose*}, {*tumeur*, *fibrome*}
  - Ressources supplétives (1 022)
    - {*base supplétive*, *mot du français*}
    - {*andr*, *mâle*}, {*ectomie*, *ablation*}, {*myo*, *muscle*}

Exploitation de contextes définitoires

Contextes définitoires

Méthode

- Définition : structure avec deux éléments :
  - *definiendum* (terme à définir) et *definiens* (la définition)
  - *Myocarde est le tissu musculaire du coeur*
- Application de quatre patrons [Péry-Woodley & Rebeyrolle, 1998]
  - *désigne*
  - *est un*
  - *est appelé*
  - *peut être défini comme*
- ...avec des variations flexionnelles
- Déclencheur : terme

Contextes définitoires

Résultats

- Extraction :
  - 2 037 contextes définitoires
  - 1 286 termes uniques
- Type de termes définis :
  - composés :
    hypoglycémie, *acidocétose*, *angiographie*, *hypokaliémie*,
  - mots affixés :
    curetage, *capsulite*, *arthrose*, *glaucome*, *durillon*, *pré-diabète*,
  - mots morphologiquement non construits :
    cataracte, *impétigo*, *zona*

Contextes définitoires

Résultats

Définitions correctes :

- L'hypoglycémie est un manque de sucre dans l'organisme
- Une septicémie est un empoisonnement du sang du à un microbe
- Le curetage est un nettoyage en profondeur d'une gencive inflammée
- Pour un être humain adulte, une hypoglycémie est une glycémie inférieure à 0,8 g/L
- Les signes classiques annonciateurs de l'hypoglycémie sont des sueurs, pâleur, palpitations, fringales en particulier
- L'impétigo est une infection cutanée, qui provoque des pustules qui dégénèrent en croûtes jaunâtres, l'impétigo est due à...

Contextes définitoires

Résultats

Définitions possiblement correctes :

- La mélancolie est une douceur qui nous berce
- Une injection est une agression, qui sauve, mais c'est quand même une agression

Contextes définitoires

Résultats

- Compréhension (*péricarde*) :
  - \+ La couche extérieure du cœur est appelée péricarde.
  - ∼ Le péricarde est un sac à double paroi contenant le cœur et les racines des gros vaisseaux sanguins.
  - − Le péricarde est un organe de glissement, formé de deux feuillets limitant une cavité virtuelle, la cavité péricardique, qui permet les mouvements cardiaques.

Contextes définitoires

Résultats

- Évaluation :
  - précision stricte : 52,5 %
    - définitions correctes : 849
  - précision lâche : 68 %
    - définitions correctes et possiblement correctes : 1 028

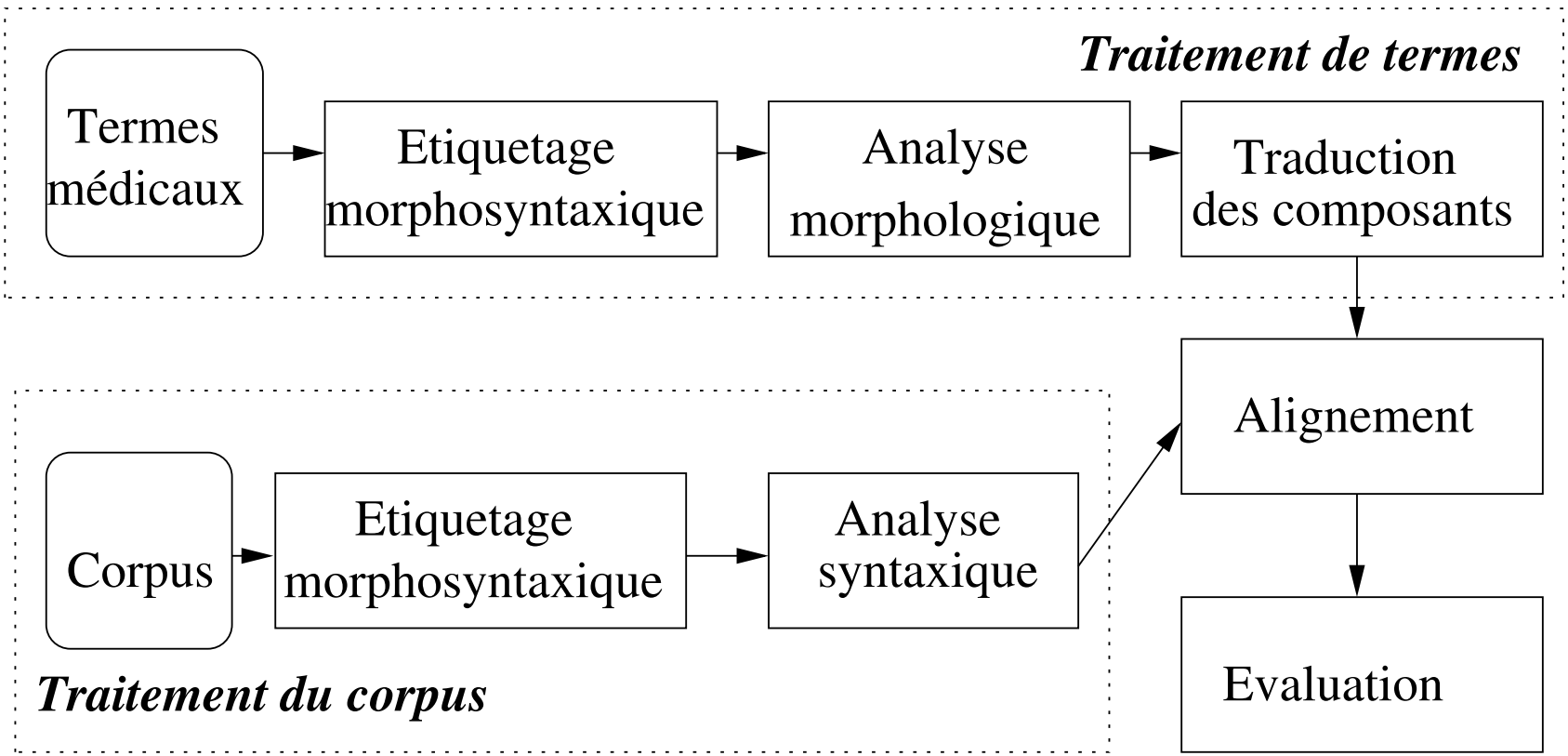
Contextes définitoires

Conclusion

- Acquisition de définitions de termes médicaux
- Différents types de termes
  - non construits, affixés, composés néoclassiques
- Résultats :
  - jusqu'à 1 028 termes
- Précision :
  - stricte : 52,5 %
  - lâche : 68 %

Composition morphologique

Composition morphologique



Composition morphologique

1. Traitement de termes médicaux

- Étiquetage morpho-syntaxique et lemmatisation `Cordial` [Laurent *et al.*, 2009]
  - *myocardique/A*, *cholécystectomie/N*
- Analyse morphologique `DériF` [Namer, 2003]
  - *myocardique/A* : *[[[myo N*] [carde N*] NOM] ique ADJ]*
  - *cholécystectomie/N* : *[[cholécysto N*] [ectomie N*] NOM]*
- Association avec les mots du français (ressource supplétive)
  - *myocardique/A* :
    - *myo*=*muscle*, *carde*=*cœur*
  - *cholécystectomie/N* :
    - *cholécysto*=*vésicule biliaire*, *ectomie*=*ablation*

Composition morphologique

2. Traitement du corpus

- `Cordial` [Laurent *et al.*, 2009]
  - étiquetage morpho-syntaxique et lemmatisation
  - analyse syntaxique
- Définir les frontières des syntagmes

Composition morphologique

3. Extraction de paraphrases

- Mise en parallèle :
  - syntagmes et décompositions morphologiques des termes
- Tout type de contextes :
  - *Les causes de tachycardie ventriculaire sont superposables à celles des extrasystoles ventriculaires : infarctus du myocarde, insuffisance cardiaque, hypertrophie du muscle du cœur et prolapsus de la valve mitrale.*

  $\Rightarrow$ {myocarde, muscle du cœur}
- Quatre paramètres à varier :
  1. taille de la fenêtre : 1, 2, 3 syntagmes
  2. ressources linguistiques :
     - formes brutes
     - ressources morphologiques (flexions, dérivations)
     - ressource de synonymes
  3. taux d'alignement des termes
  4. taux d'alignement des syntagmes

Composition morphologique

4. Évaluation

- Validation :
  1. paraphrase correcte : {myocardique, muscle du cœur},
  2. analyse morphologique incorrecte : {sanglot, lot sang}
  3. traduction vers le français incorrecte : *antisolaire*, {*sol*, *sol*} au lieu de {*sol*, *solaire*}
  4. informations correctes au milieu d'autres informations, informations partielles
     - partiel : {*endophtalmie*, *interne de l'œil*}
     - complet : *inflammation des tissus internes de l'œil*
  5. extraction fausse
- Précision :
  - précision stricte $P_{stricte}$: cas 1
  - précision lâche P_{lache} : cas 1 et 4
  - taux d'erreurs : cas 5
  - cas 2 et 3 : pas pris en compte

Composition morphologique

Résultats

- 274 131 termes UMLS et Snomed International
- 76 536 mots sans nombres
- 15 121 mots analysés par `Dérif`
  - décomposés en deux bases au moins
- Alignement syntagme/terme (pourcentage d'alignement) :

  *E*1 : terme et syntagme complets dans l'alignement :
  - {*myo pathie*, *maladie du muscle*}

  *E*2 : terme complet, syntagme partiel :
  - {*myo pathie*, *maladie du muscle cardiaque*}

  *E*3 : terme partiel, syntagme complet :
  - {*myopathie*, *la maladie*}

  *E*4 : terme et syntagme partiels :
  - {*myopathie*, *l' origine de la maladie*}
- Travail avec E1 (le plus optimisé)

Composition morphologique

Extraction de paraphrases

Nb de	*unigrammes*			*bigrammes*			*trigrammes*		
	b	*l*	*s*	*b*	*l*	*s*	*b*	*l*	*s*
syntagme	9854	16093	22110	11875	18504	27670	7936	12284	19984
terme unique	1513	1947	2090	1780	2260	2463	1523	1966	2231
$syntagme_{E1}$	2681	4163	5370	1109	1611	2521	403	634	988
$terme\ unique_{E1}$	668	1023	1051	492	670	962	239	358	472

- total et E1
- ressources linguistiques : augmentent le volume
  - *b* : sans les ressources
  - *l* : ressources morphologiques
  - *s* : ressources de synonymie
- n-grammes de syntagmes : diminuent le volume
  - seuil d'alignement acceptable

Composition morphologique

Évaluation

Nombre de	*unigrammes*			*bigrammes*			*trigrammes*		
	b	*l*	*s*	*b*	*l*	*s*	*b*	*l*	*s*
paraphrases correctes	549	785	644	378	517	461	195	290	257
possibl. correctes	39	32	67	22	45	75	10	19	41
traitement de termes	47	60	44	28	28	46	9	10	26
paraphrase incorrectes	33	146	296	64	80	380	25	39	148
$P_{stricte}$	82	77	61	77	77	48	82	81	55
P_{lache}	88	80	68	81	84	40	86	86	63
$\%_{incorrect}$	5	14	28	13	12	39	11	11	31

- Évaluation :
  - précision stricte 82 à 55 %
  - précision lâche 86 à 40 %
  - taux d'erreurs 5 à 39 %
- Ressources
  - sans ressources : précision la plus élevée
  - ressources morphologiques : bonne précision
  - ressources de synonymie : la plus faible précision

Composition morphologique

Analyse morphologique

- Analyse ambiguë
  - *[post [[uro N*] [graphie N*] NOM] NOM]*
  - *[[posturo N*] [graphie N*] NOM]*
- Analyse incorrecte
  - *sanglot* : *lot* et *sang*
  - *exotique* : *externe* et *oreille*

Composition morphologique

Extraction de paraphrases et leur évaluation

Extraction de paraphrases correctes

- Brut
  - {podalgie, douleur du pied}
  - {mastite, inflammation du sein}
  - {cystoprostatectomie, ablation de la vessie et de la prostate}
- Morphologie
  - {desmorrhexie, rupture des ligaments} (ligament→ligaments)
  - {bronchite, inflammation des bronches/inflammation bronchique} (bronche→bronches, bronche→bronchique)
  - {dentalgie, douleurs dentaires} (dents→dentaires)
- Synonymie
  - {aclasie, absence de fracture} (cassure→fracture)
  - {enterectomie, résection des intestins} (ablation→résection)

Composition morphologique

Extraction de paraphrases et leur évaluation

- Relations sémantiques entre composants :
  - bien gérées sur la base du corpus
  - erreurs : coordination/subordination
    - *hematospermie* : *le sang ou le sperme*, au lieu de

    → *le sang dans le sperme*
- Termes non compositionnels :
  - *ostéodermie* : *peau* et *os*, au lieu de

    → *une structure d'écailles, de plaques osseuses ou d'autres compositions dans les couches dermiques de la peau, comme chez les lézards ou dinosaures*
- Couverture des 15 121 termes analysés morphologiquement :
  - 6,8 % (1 031) paraphrases correctes
  - 7,5 % (1 128) paraphrases correctes et possiblement correctes correctes

Composition morphologique

Ressources linguistiques

Synonymie : valeurs sémantiques contextuelles

Peut extraire des paraphrases incorrectes :

- *cardialgie* :
  - correct : *douleur de cœur*
  - extrait : *plaie du cœur* (douleur→plaie)
- *cheiropathie* :
  - correct : *maladie des mains*
  - extrait : *Le syndrome main* (maladie→syndrome)
- *cinépathie*
  - correct : *mal des transports*
  - décomposé en *mouvement* et *maladie*
  - extrait : *évolution du syndrome* (mouvement→évolution, maladie→syndrome)

Composition morphologique

Termes non paraphrasés

- Plus de 2 composants :
  - *hémi-desmo-some*, *hémo-histio-blaste*
- Composants et leurs combinaisons rares :
  - *hémi-desmo-some* : *demi*, *ligament*, *corpuscule*
- Ressource supplétive :
  - trop stricte
  - d'autres méthodes [Claveau & Kijak, 2014]

Composition morphologique

Conclusion

- Paraphrases grand public pour les termes médicaux
- Composés néoclassiques
- Résultats :
  - jusqu'à 1 128 termes
- Précision moyenne :
  - toutes les expériences : 76 %
  - sans synonymes : 86 %

Exploitation de reformulations

Reformulations

Motivation

- Reformulation : redire différemment ce qui a déjà été dit [Le Bot *et al.*, 2008]
- Présence de reformulations :
  - indique les mots/termes difficiles
  - offre les indices pour l'extraction
- Exploiter des données fiables

Reformulations

Matériel

- Forum de discussion *masante.net* (questions/réponses)
  - corpus de développement
  - 6 139 réponses, 315 362 occurrences
  - réponses des médecins
  - reformuler les propos pour qu'ils soient mieux compréhensibles
- Wikipédia francophone, Portail de la Médecine
  - corpus de test
  - 18 434 articles, 15 235 219 occurrences
- Trois méthodes :
  - abréviations
  - marqueurs de reformulation :
    - *c'est-à-dire, autrement dit, encore appelé(e)(s)*
  - parenthèses

Reformulations

Matériel

- Ressources linguistiques :
  - liste de mots de vides
  - morphologique : 163 823 paires de mots (dérivations, flexions) ({*aorte*, *aortique*}) et ({*aortique*, *aortiques*})
  - synonymique :
- Terminologie médicale en français :
  - UMLS : Unified Medical Language System [Lindberg *et al.*, 1993]
  - SNOMED Int : Systematized Nomenclature of Medicine [Côté *et al.*, 1993]

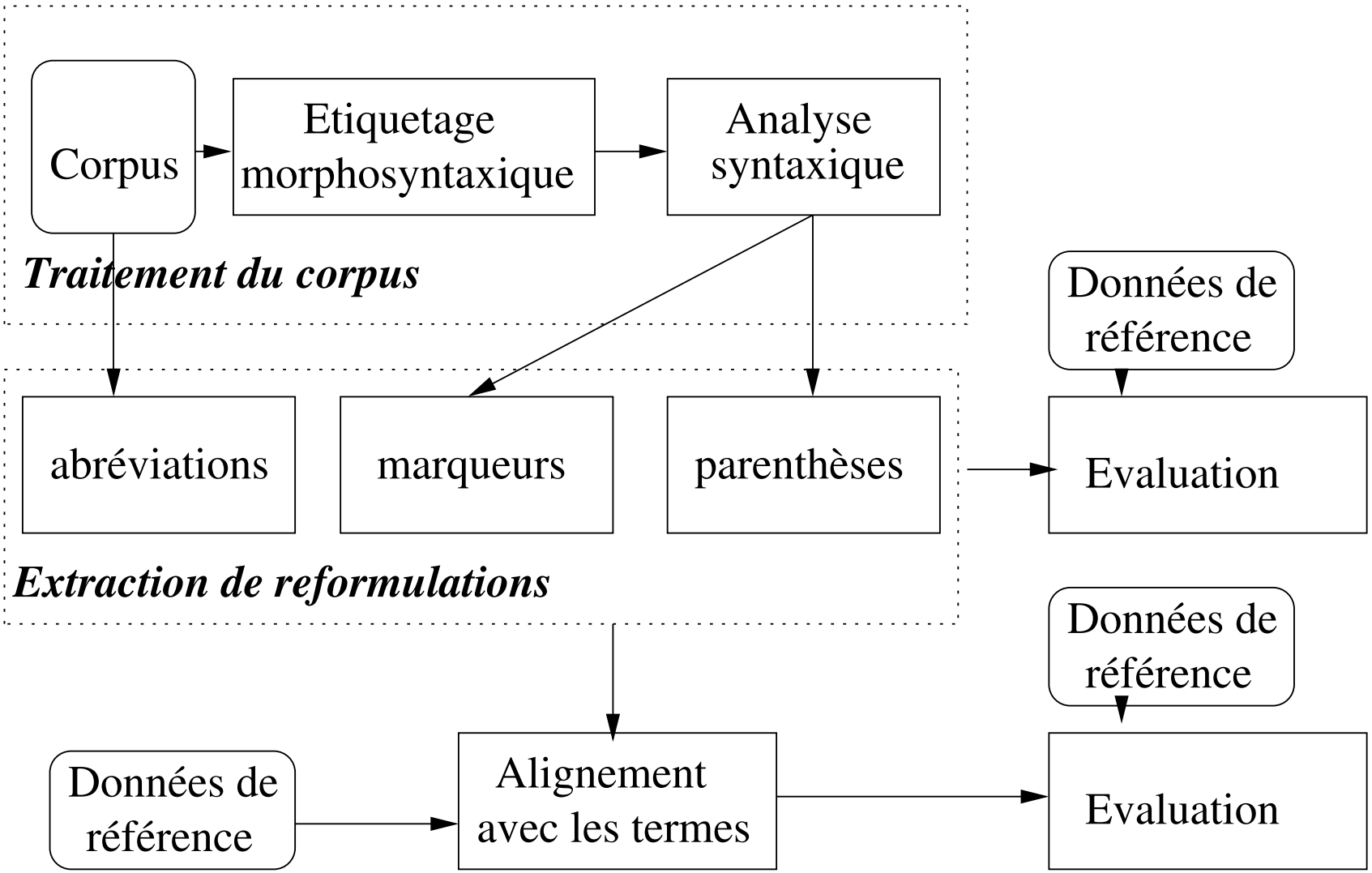
Reformulations

Matériel

- Données de référence pour les extractions
- Toutes les phrases avec les reformulations
- Annotations de reformulations avec un guide d'annotation
  - *<C>d'origine labyrinthique</C>, <M>c'est à dire</M>, <Rgen>venant de l'oreille interne</Rgen>*
- Accord inter-annotateur : kappa de Cohen [Cohen, 1960]
  - 2 niveaux : phrase et token
  - accord binaire : O/N

	Extraction	
	Phrase	*Token*
Abréviations	0,661	0,967
Marqueurs	0,24	0,816
Parenthèses	0,651	0,575

Reformulations



Reformulations

Abréviations : méthode

- Inspiré de [Schwartz & Hearst, 2003]
- 2 types de patrons :
  1. anti-inflammatoires non stéroïdiens (AINS)
  2. AVC (Accident Vasculaire Cérébral)
- Utilisation du texte brut
- Reconnaissance : majuscules, parenthèses
- Association lettre $\rightarrow$ mot
- Gestion des doublons : leucémie aiguë lymphoblastique (LAL)

Reformulations

Abréviations : résultats

	Dév.	*Test*		*P*	*R*	*F*
nb occ.	75	88 762	*exact*	0.74	0.74	0.74
nb types	42	8 106	*inexact*	0.94	0.94	0.94

- Types d'extractions :
  - Complètes : {AINS, anti inflammatoire non stéroïdien}
  - Partielles mais correctes : {CIV, communication interventriculaire}
  - Partielles et exploitables : {CHU, hôpital universitaire}
  - Partielles et inexploitables : *NFS : faire sang*
  - Pas d'extraction : *comment sont les ALAT(ou SGPT) et les ASAT (ou SGOT)*
- Difficultés :
  - *index de masse corporelle (BMI)*
  - *thyroglobuline (TG)*
  - *dispositif intra-utérin qui est imbibé de progestatif (DIU)*
  - *normaliser les transaminases (ALAT)*

Reformulations

Abréviations : résultats

- {AINS, anti inflammatoires non stéroïdiens}
- {IST, infection sexuellement transmissible}
- {ECBU, examen urine}
- {ECBC, examen crachats}
- {SADAM, syndrome algo dystrophique de appareil mandibulaire}
- {SA, semaines aménorrhée}
- {HAS, Haute Autorité Santé}
- {DIU, dispositif intra utérin}
- {HON, health on net}
- {FIV, fécondation in vitro}
- {TG, thyroglobuline}
- {AVC, accident vasculaire cérébral}
- {VS, vitesse sédimentation}

Reformulations

Marqueurs : méthode

concept *marqueur* *reformulation*
vésiculaire, *c'est-à-dire*, *venant de la vésicule biliaire*

- 3 marqueurs :
  - *c'est-à-dire*
  - *autrement dit ; Autrement dit*
  - *encore appelé(e)(s)*
- Pré-traitement
- Étiquetage et analyse morpho-syntaxique de Cordial [Laurent *et al.*, 2009]
- Déclencheur : marqueurs
- Extraction du concept et de la reformulation :
  - informations syntaxiques
  - frontières : syntagmes ou propositions

Reformulations

Marqueurs : méthode

forme	*lemme*	*POS*	*POSMT*	*GS*	*type GS*	*Prop*
Vous	vous	PPER2P	Pp2.pn	1	S	1
ne	ne	ADV	Rpn	3—1	S	1
devez	devoir	VINDP2P	Vmip2p	3	V	1
pas	pas	ADV	Rgn	3	Q	1
employer	employer	VINF Vmn	–	5	D	2
de	de	PREP	Sp	7	D	2
savons	savon	NCMP	Ncmp	7	D	2
ou	ou	COO	Cc	7	F	2
des	de le	DETDPIG	Da-.p-i	10—7	F	2
laits	lait	NCMP	Ncmp	10—7	F	2
sophistiqués	sophistiqué	ADJMP	Afpmp	10—7	F	2
,	,	PCTFAIB	Ypw	-	-	2
c'	ce	PDS	Pd-..-	13	N	2
est	est	ADV	Rgp	-	p	2
-à	à	PREP	Sp	16	F	2
-dire	dire	VINF	Vmn–	16	F	2
contenant	contenant	NCMS	Ncms	17	D	2
plusieurs	plusieurs	ADJIND	Dt-.p-	19	D	2
composants	composant	NCMP	Ncmp	19	D	2

Reformulations

Marqueurs : méthode

forme	*lemme*	*POS*	*POSMT*	*GS*	*type GS*	*Prop*
Vous	vous	PPER2P	Pp2.pn	1	S	1
ne	ne	ADV	Rpn	3—1	S	1
devez	devoir	VINDP2P	Vmip2p	3	V	1
pas	pas	ADV	Rgn	3	Q	1
employer	employer	VINF Vmn	–	5	D	2
de	de	PREP	Sp	7	D	2
savons	savon	NCMP	Ncmp	7	D	2
ou	ou	COO	Cc	7	F	2
des	de le	DETDPIG	Da-.p-i	10—7	F	2
laits	lait	NCMP	Ncmp	10—7	F	2
sophistiqués	sophistiqué	ADJMP	Afpmp	10—7	F	2
,	,	PCTFAIB	Ypw	-	-	2
c'	ce	PDS	Pd-..-	13	N	2
est	est	ADV	Rgp	-	p	2
-à	à	PREP	Sp	16	F	2
-dire	dire	VINF	Vmn–	16	F	2
contenant	contenant	NCMS	Ncms	17	D	2
plusieurs	plusieurs	ADJIND	Dt-.p-	19	D	2
composants	composant	NCMP	Ncmp	19	D	2

Reformulations

Marqueurs : méthode

forme	*lemme*	*POS*	*POSMT*	*GS*	*type GS*	*Prop*
Vous	vous	PPER2P	Pp2.pn	1	S	1
ne	ne	ADV	Rpn	3—1	S	1
devez	devoir	VINDP2P	Vmip2p	3	V	1
pas	pas	ADV	Rgn	3	Q	1
employer	employer	VINF Vmn	–	5	D	2
de	de	PREP	Sp	7	D	2
savons	savon	NCMP	Ncmp	7	D	2
ou	ou	COO	Cc	7	F	2
des	de le	DETDPIG	Da-.p-i	10—7	F	2
laits	lait	NCMP	Ncmp	10—7	F	2
sophistiqués	sophistiqué	ADJMP	Afpmp	10—7	F	2
,	,	PCTFAIB	Ypw	-	-	2
c'	ce	PDS	Pd-..-	13	N	2
est	est	ADV	Rgp	-	p	2
-à	à	PREP	Sp	16	F	2
-dire	dire	VINF	Vmn–	16	F	2
contenant	contenant	NCMS	Ncms	17	D	2
plusieurs	plusieurs	ADJIND	Dt-.p-	19	D	2
composants	composant	NCMP	Ncmp	19	D	2

Reformulations

Marqueurs : résultats

	Dév.	*Test*		*P*	*R*	*F*
nb occ.	96	2 757	*exact*	0.24	0.24	0.24
nb types	96	2 710	*inexact*	0.98	0.98	0.98

- Difficultés :
  - détection des frontières
    - *en* **c'est-à-dire** *au contact du sang circulant*
    - *une toxi-infection*, **c'est-à-dire**, *qu' elle peut*
  - sémantique
    - *en 10 ans* **autrement dit** *sur 64 millions de personnes*
    - *un objectif* **c'est-à-dire** *une finalité*

Reformulations

Marqueurs : résultats

- des canaux galactophores c'est-à-dire sécrètent le lait
- erratiques c'est-à-dire qu'ils changent de d'aspect et d'endroit
- par une lithiase c'est-à-dire un caillou
- clivage du moi c'est-à-dire comme une opposition entre le moi et la réalité
- au gré de la désintégration radioactive du 18 F c'est-à-dire avec une demi-vie d'environ
- un trouble de l'identité sexuelle c'est-à-dire qu'ils s'identifient à un genre ne correspondant pas à leur sexe biologique
- une enzyme protéolytique c'est-à-dire digère les protéines comme le fait le suc pancréatique
- celle de troubles fonctionnels intestinaux encore appelés colopathie fonctionnelle

Reformulations

Parenthèses : méthode

concept (reformulation)
avec des prélèvements (biopsie)
myopie (difficulté à voir de loin)
boutons (on parle d'éruption cutanée)

- Pré-traitement
- Étiquetage et analyse morpho-syntaxique de Cordial [Laurent *et al.*, 2009]
- Déclencheur : parenthèses
- Extraction :
  - concept : informations syntaxiques
  - reformulation : entre parenthèses

Reformulations

Parenthèses : méthode

- Sémantique des parenthèses :
  - une incise : *de fièvre (en avez-vous)*
  - un exemple : *des infections sexuellement transmissibles (papillomavirus)*
  - une précision : *d'une maladie génétique (pas d'une dégénérescence)*
  - une énumération/instantiation :
    - *un pansement intestinal (type smecta)*
    - *prescrire (primpéran, vogalène, llioresal par exemple)*
- Une série de filtres lexicaux et de structure

Reformulations

Parenthèses : résultats

	Dév.	*Test*		*P*	*R*	*F*
nb occ.	312	100 103	*exact*	0.23	0.23	0.23
nb types	305	92 971	*inexact*	0.68	0.68	0.68

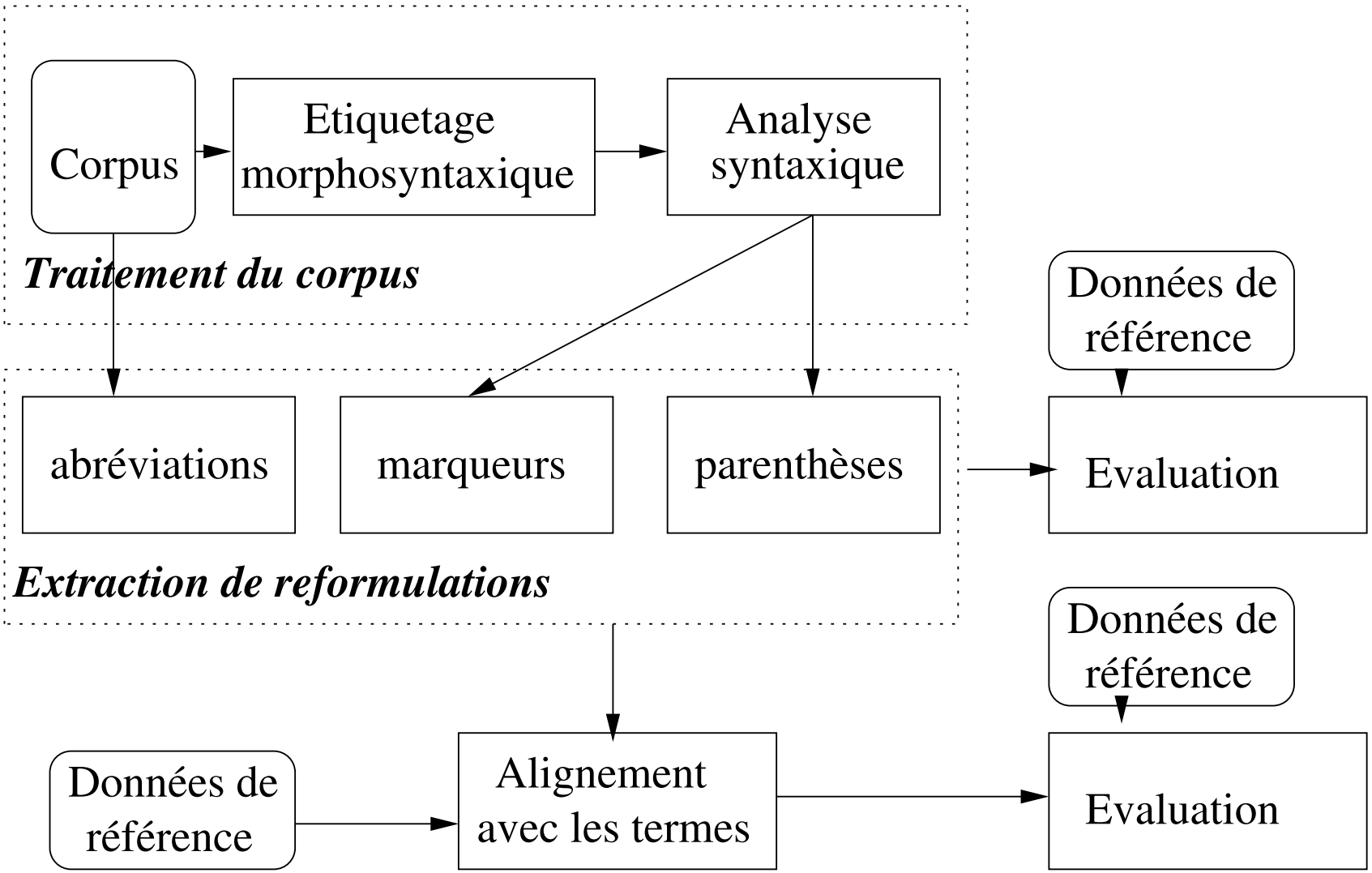
- Difficultés :
  - sémantique de parenthèses
  - détection de frontières
    - *se bouche (hémorroïdes)*
    - *de fer (hypochromie)*
    - *ni le taux d'hémoglobine (Hb)*
  - valeurs relatives, validité
    - *basse (inférieure à 13 g/dL)*
  - extractions non pertinentes :
    - *énergétique (carence plutôt liée au marasme)*

Reformulations

Parenthèses : résultats

- *de ses anticoagulants (héparine)*
- *des antibiotiques (cyclines)*
- *avec des prélèvements (biopsie)*
- *par de toutes petites particules (nanoparticules)*
- *par une pH-métrie (pour mesurer l'acidité de l'oesophage)*
- *du duodénum (le début de l'intestin)*
- *un reflux gastro-oesophagien (reflux de l'acidité de l'estomac dans l'oesophage)*
- *d'un flou visuel au changement de position (éclipse visuelle)*
- *un radio du rachis (colonne vertébrale)*
- *d'une infection virale (rhume, grippe)*
- *à la fin du premier trimestre (8-12 sem)*

Reformulations



Reformulations

Alignement : méthode

- Alignement des segments extraits avec les termes
  - pertinence des extractions
    - *en fibres (pas trop vite sinon vous serez ballonnée)*
  - association avec les termes médicaux avérés
    - exploitation en recherche d'information, indexation, etc.
- Approches :
  - casse, ordre, normalisation morphologique
  - mots vides
  - taux d'alignement : segment, terme [40 ;100]
- Données de référence :
  - à partir des alignements 40/40
  - deux évaluateurs, consensus
  - valider les bonnes propositions

Reformulations

Alignement : résultats

- Proposition pertinente (alignement complet) :
  - *AINS : ains.C0003211/C-60300*
  - *anti inflammatoires non stéroïdiens : anti inflammatoires steroidiens.C0003211*
- Variation morphologique de *troubles fonctionnels intestinaux* (alignement partiel) :
  - *troubles gastrointestinaux fonctionnels/C0559031.T047.DISO*
  - *troubles gastro intestinaux fonctionnels/C0559031.T047.DISO*
- Proposition partielle (alignement partiel) :
  - *semaines amenorrhee : amenorrhee/C0002453.T047.DISO*
- Proposition compositionnelle de *cause de pus* :
  - *cause/C0085978.T078.CONC/...*
  - *pus/C0034161.T031.ANAT/...*
- Proposition non pertinente :
  - *LCR : ph lcr/C0853364* (trop précis)
  - *liquide cerebro : regime liquide/C-F2300*
- Aucune proposition :
  - *NFS : —*

Reformulations

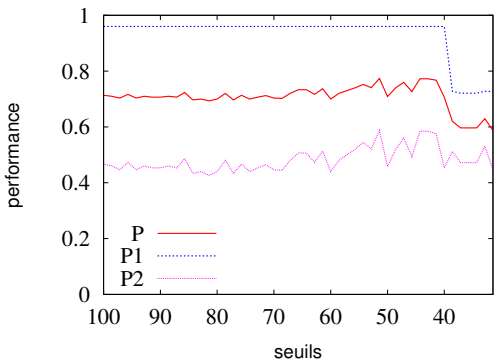
Alignement : résultats

	Développement			*Test*		
	Abrév	*Marq*	*Par*	*Abrév*	*Marq*	*Par*
nb occurrences	75	96	312	88 762	2 757	100 103
total	11	5	38	154	42	3 738
partiel	44	37	123	1 634	557	25 708
non alignés	20	54	150	6 318	1 937	60 928

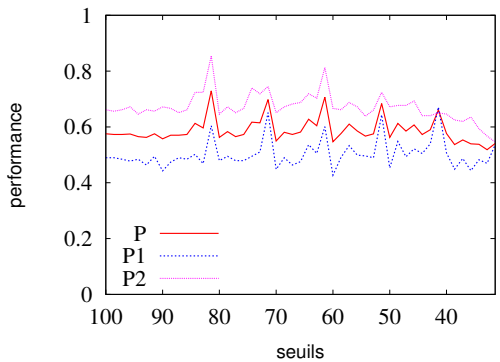
- deux segments alignés :
  - *d'une fibromyalgie : fibromyalgie.C0016053.T047.DISO*
  - *SPID (syndrome polyalgique idiopathique diffus) : syndrome polyalgique idiopathique diffus/C0016053.T047.DISO*
- un seul segment aligné :
  - *TSH : –*
  - *thyroïde : thyroide.C0040132.T023.ANAT*
- aucun segment aligné :
  - *HAS : –/Haute Autorité Santé : –*

Reformulations

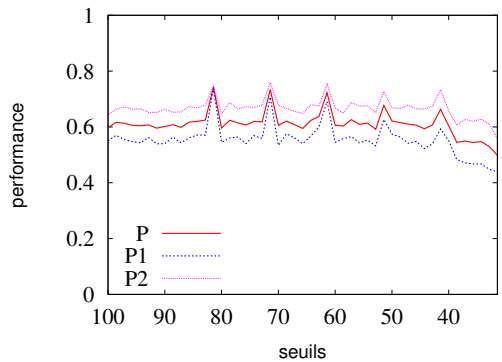
Alignement : résultats



abréviations



marqueurs



parenthèses

- meilleurs seuils (segment/terme) : 80/100, 70/100
- segment 2 : souvent plus facile à aligner

Reformulations

Typologie

- Typologie de l'état de l'art [Schwartz & Hearst, 2003]
- Difficulté de classifier avant les alignements (trop de bruit)
- Reformulations avec marqueurs :
  - synonyme :
    - *l'interruption naturelle ou accidentelle de la grossesse, c'est-à-dire, un avortement spontané*
  - définition :
    - *la contractilité myocardique, c'est-à-dire, la capacité des cellules musculaires myocardiques à se contracter en réponse à un potentiel d'action*
- Reformulations avec parenthèses :
  - synonyme :
    - *nerveux (hystérie)*
  - définition :
    - *une scoliose (courbure de la colonne vertébrale)*
  - relation cause à effet :
    - *d'ulcère tropical (moisissures de la jungle)*

Reformulations

Conclusion

- Exploitation de reformulation pour l'acquisition du vocabulaire
- 3 méthodes :
  - abréviation : inspiré d'un algorithme existant
  - marqueurs, parenthèses : observations des données
- Alignement avec une terminologie
- Résultats :
  - meilleurs résultats avec les abréviations (74, 94%)
  - bonne couverture avec les parenthèses
  - bonne pertinence avec les marqueurs
  - taux d'alignement : 65% - 313 (dev) ; 17% - 31 833 (test)

Reformulations

Conclusion

- Reformulations dans les corpus à destination du grand public
  - réponses des médecins dans les forums de discussion
  - Wikipédia
- Extraction de segments
  - différents types de segments
- Complémentarité de méthodes
- Alignement avec la terminologie médicale

Conclusion et Travaux futurs

Comparaison entre les approches

	type terme	nb. para	précision
définitions	tout type	1 028	0,52, 0,68
morphologie	composés	1 128	0,76, 0,86
abréviations	abréviations	42, 8 106	0,74/0,94
marqueurs	tout type	96, 2 710	0,24/0,98
parenthèses	tout type	305, 92 971	0,23/0,68

- propositions souvent différentes
- faible recouvrement
- lien avec les terminologies

Comparaison avec les travaux existants

	type terme	nb. para	précision
[Zeng *et al.*, 2006]	tous	CHV	
[Elhadad & Sutaria, 2007]	tous	152	0,58
[Deléger & Zweigenbaum, 2008]	m-synt.	65, 82	0,67, 0,60
[Cartoni & Deléger, 2011]	m-synt.	109	0,66
définitions	tout type	1 028	0,52, 0,68
morphologie	composés	1 128	0,76, 0,86
abréviations	abréviations	42, 8 106	0,74/0,94
marqueurs	tout type	96, 2 710	0,24/0,98
parenthèses	tout type	305, 92 971	0,23/0,68

- morpho-syntaxique :
  - {*consommation régulière*, *consommer de façon régulière*}
- performances comparables, meilleure couverture
- lien avec les terminologies

Comparaison avec les travaux existants

- `DériF` [Namer, 2003] :
  - glose en langage artificiel pour tout terme analysé
  - notre méthode : la couverture dépend du contenu des corpus
- *myocarde* :
  - *"(Partie de – Type particulier de) coeur en rapport avec le(s) muscle"*
  - *muscle du coeur*
- *desmorrhexie* :
  - *"rupture (du – liée au) ligament"*
  - *rupture des ligaments*

Intégration

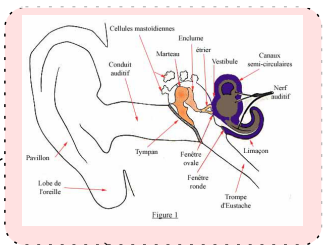
- Ajout d'informations

réparation

La myoplastie est une réfection chirurgicale d'un muscle.

inflammation de l'oreille

La pétrosite est une ostéite de la partie profonde du rocher (pyramide pétreuse)



presque toujours consécutive à une otite moyenne.

inflammation de l'oreille

affection de la peau (dermatose)

augmentation de volume

cellule de la peau

La mastocytose est une hyperplasie des mastocytes dont les manifestations peuvent prédominer au niveau de la peau (Urticaire pigmentaire).

affection de la peau (dermatose)

Intégration

- Remplacement, substitution
  {*lombalgie*, *douleurs lombaires*}, {*hépatite*, *inflammation du foie*}
  - *Les lombalgies inflammatoires provoquent une douleur de type inflammatoire*
    Les douleurs lombaires inflammatoires provoquent une douleur de type inflammatoire
  - *La lombalgie est une affection coûteuse pour le système de soins de santé et est un motif fréquent d'absentéisme.*
    La douleurs lombaires est une affection coûteuse pour le système de soins de santé et est un motif fréquent d'absentéisme.
  - *Hépatite C : lutter contre le virus et ses résistances*
    Inflammation du foie C : lutter contre le virus et ses résistances

Conclusion générale

- Acquisition de ressources
  - pour expliquer les termes techniques
- Méthodes dédiées à différentes manifestations
  - paraphrases, reformulations...
- Corpus pour le grand public
- Méthodes complémentaires
- Résultats intéressants et exploitables

Travaux futurs

- Augmenter la couverture des paraphrases et reformulations :
  - d'autres corpus
    - parallèles (Cochrane, notices de médicaments, Wiki/Viki)
    - monolingues
  - ressources supplétives plus couvrantes
  - d'autres méthodes pour extraire des paraphrases
  - d'autres unités syntaxiques (propositions)
- Alignement plus complet
- Diffusion de la ressource
- D'autres langues
- Simplification lexicale de textes médicaux
  - projet ANR CLEAR (*Communication, Literacy, Education, Accessibility, Readability*)


AMA (1999).
Health literacy : report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association.
JAMA, **281**(6), 552–7.


Antoine, E. & Grabar, N. (2016).
Exploitation de reformulations pour l'acquisition d'un vocabulaire expert/non expert.
In *TALN 2016*.


Berland, G., Elliott, M., Morales, L., Algazy, J., Kravitz, R., Broder, M., Kanouse, D., Munoz, J., Puyol, J. & et al, M. L. (2001).
Health information on the internet. accessibility, quality, and readability in english ans spanish.
JAMA, **285**(20), 2612–2621.


Biran, O., Brody, S. & Elhadad, N. (2011).
Putting it simply : a context-aware approach to lexical simplification.
In *ACL*.


Boyer, C., Baujard, O., Baujard, V., Aurel, S., Selby, M. & Appel, R. (1997).
Health on the net automated database of health and medical information.
Int J Med Inform, **47**(1-2), 27–9.


Brin-Henry, F. (2014).
Éducation thérapeutique du patient aphasique et son conjoint.
Rééducation orthophonique, **256**, 9–20.


Cartoni, B. & Deléger, L. (2011).
Découverte de patrons paraphrastiques en corpus comparable : une approche basée sur les n-grammes.
In *TALN*.


Claveau, V. & Kijak, E. (2014).
Generating and using probabilistic morphological resources for the biomedical domain.
In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3348–3354.


Cohen, J. (1960).
A coefficient of agreement for nominal scales.
Educational and Psychological Measurement, **20**(1), 37–46.


Côté, R. A., Rothwell, D. J., Palotay, J. L., Beckett, R. S. & Brochu, L. (1993).
The Systematised Nomenclature of Human and Veterinary Medicine : SNOMED International.
Northfield : College of American Pathologists.


Deléger, L. & Zweigenbaum, P. (2008).
Paraphrase acquisition from comparable medical corpora of specialized and lay texts.
In *AMIA 2008*, pp. 146–50.


d'Ivernois, J.-F., Gagnayre, R. & *et al* (2011).

Compétences d'adaptation à la maladie du patient : une proposition [*The patient's psychosocial skills : a proposal*].
Educ Ther Patient/Ther Patient Educ, **3**(2), S201–S205.


Elhadad, N. & Sutaria, K. (2007).
Mining a lexicon of technical terms and lay equivalents.
In *BioNLP*, pp. 49–56.


Glasgow, R. E., Kurz, D., King, D., Dickman, J. M., Faber, A. J., Halterman, E., Woolley, T., Toobert, D. J. & et al, L. A. S. (2012).
Twelve-month outcomes of an internet-based diabetes self-management support program.
Patient Education and Communication, **87**(1), 81–92.


Glavas, G. & Stajner, S. (2015).
Simplifying lexical simplification : Do we need simplified corpora ?
In *ACL-COLING*, pp. 63–68.


Golay, A., Lagger, G. & Giordan, A. (2007).
Motivating patient with chronic diseases.
Journ of Med and the Person, **5**(2), 57–63.


Grabar, N. & Hamon, T. (2016).
Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux.
TAL, **57**(1), 85–109.


Gross, O. & Gagnayre, R. (2013).

Hypothèse d'un modèle théorique du patient-expert et de l'expertise du patient : processus d'élaboration.
Recherches qualitatives, **15**(HS), 147–165.

Guilbert, M. (2014).
C'est grave docteur ?
Europe : Les Éditions de l'Opportun.

Kim, Y.-S., Hullman, J., Burgess, M. & Adar, E. (2016).
Simplescience : Lexical simplification of scientific terminology.
In *EMNLP*, pp. 1–6.

Laurent, D., Nègre, S. & Séguéla, P. (2009).
L'analyseur syntaxique Cordial dans Passage.
In *TALN 2009*.

Le Bot, M.-C., Schuwer, M. & Élisabeth Richard (dir.) (2008).
La reformulation : Marqueurs linguistiques – Stratégies énonciatives.
Rennes : Rivages linguistiques.

Lindberg, D., Humphreys, B. & McCray, A. (1993).
The unified medical language system.
Methods Inf Med, **32**(4), 281–291.

McCray, A. (2005).
Promoting health literacy.
J of Am Med Infor Ass, **12**, 152–163.

McCray, A., Loane, R., Browne, A. & Bangalore, A. (1999).

Terminology issues in user access to web-based medical information.
In *AMIA Symposium 1999*, pp. 107–7.

Namer, F. (2003).
Automatiser l'analyse morpho-sémantique non affixale : le système DériF.
Cahiers de Grammaire, **28**, 31–48.

Patel, V., Branch, T. & Arocha, J. (2002).
Errors in interpreting quantities as procedures : The case of pharmaceutical labels.
Int journ med inform, **65**(3), 193–211.

Péry-Woodley, M. & Rebeyrolle, J. (1998).
Domain and genre in sublanguage text : definitional microtexts in three corpora.
In *LREC*, pp. 987–992.

Risk, A. & Dzenowagis, J. (2001).
Review of internet information quality initiatives.
Journal of Medical Internet Research, **3**(4).

Schwartz, A. S. & Hearst, M. A. (2003).
A simple algorithm for identifying abbreviation definitions in biomedical text.
In *Pacific Symposium on Biocomputing*, pp. 451–456.

Specia, L., Jauhar, S. & Mihalcea, R. (2012).
Semeval-2012 task 1 : English lexical simplification.
In **SEM 2012*, pp. 347–355.

Sørensen, M. H. (1996).

Turchin's Supercompiler Revisited - An operational theory of positive information propagation.
Master thesis, University of Copenhagen, Copenhagen, Denmark.


Tapi Nzali, M., Bringay, S., Lavergne, C., Opitz, T., Azé, J. & Mollevi, C. (2015).
Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux.
In *IC 2015*.


Tran, T., Chekroud, H., Thiery, P. & Julienne, A. (2009).
Internet et soins : un tiers invisible dans la relation médecine/patient ?
Ethica Clinica, **53**, 34–43.


Williams, M., Parker, R., Baker, D., Parikh, N., Pitkin, K., Coates, W. & Nurss, J. (1995).
Inadequate functional health literacy among patients at two public hospitals.
JAMA, **274**(21), 1677–1682.


Wubben, S., van den Bosch, A. & Krahmer, E. (2012).
Sentence simplification by monolingual machine translation.
In *ACL*, pp. 1015–1024.


Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C. & Lee, L. (2010).
For the sake of simplicity : Unsupervised extraction of lexical simplifications from Wikipedia.
In *NAACL*, pp. 365–368.


Zeng, Q. & Tse, T. (2006).

Exploring and developing consumer health vocabularies.
JAMIA, **13**, 24–29.

Zeng, Q. T., Tse, T., Divita, G., Keselman, A., Crowell, J. & Browne, A. C. (2006).
Exploring lexical forms : first-generation consumer health vocabularies.
In *AMIA 2006*, pp. 1155–1155.

Zhu, Z., Bernhard, D. & Gurevych, I. (2010).
A Monolingual Tree-based Translation Model for Sentence Simplification.
In *COLING 2010*, pp. 1353–1361.