Aider la compréhension de la terminologie médicale par les non experts

Natalia Grabar

STL CNRS UMR8163

Vers la simplification de textes de spécialité

1. Contexte
2. Détection de mots/passages difficiles
3. Acquisition de ressources pour la simplification
4. Conclusion

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 : 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...)

Health Literacy

- Facilité à lire, comprendre et utiliser les informations de santé
- Dans différents contextes :
  - suivre les instructions de traitement
  - prendre les décisions (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 et al., 2013 ; Brin, 2014) :
  - mieux communiquer avec les patients
  - mieux guider les patients dans leurs parcours médical
- Établir une confiance mutuelle (Sorensen, 1996)
- Améliorer l'efficacité des soins médicaux

FALC : Facile à Lire et à Comprendre

- Objectif :
  - rendre compréhensibles les informations institutionnelles pour
    - le grand public
    - les personnes avec des pathologies
    - les personnes avec le handicap intellectuel
    - ...
- Motivation : législation Européenne en vigueur à partir de janvier 2015

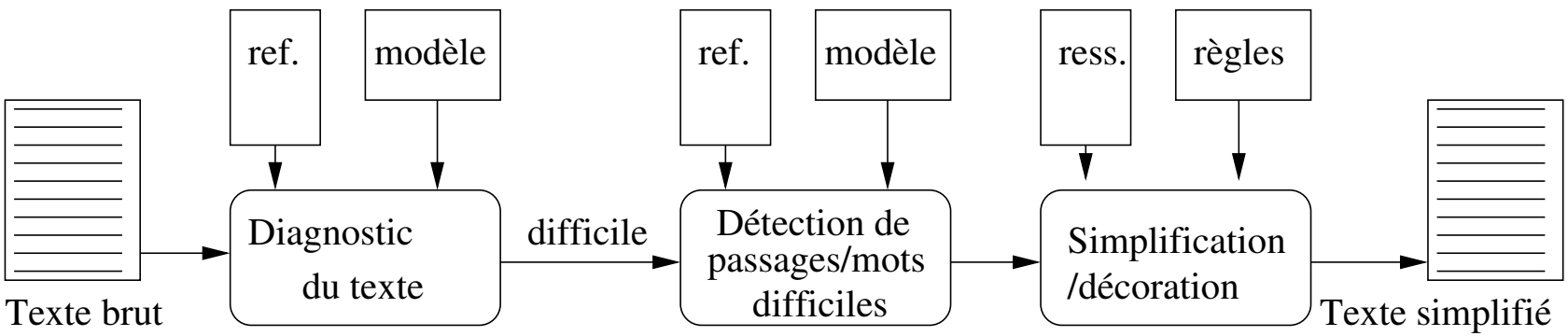
Ce que cela donne du côté des patients...

(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 triglicириliques*
- *Faut m'opérer du corps vitreux*

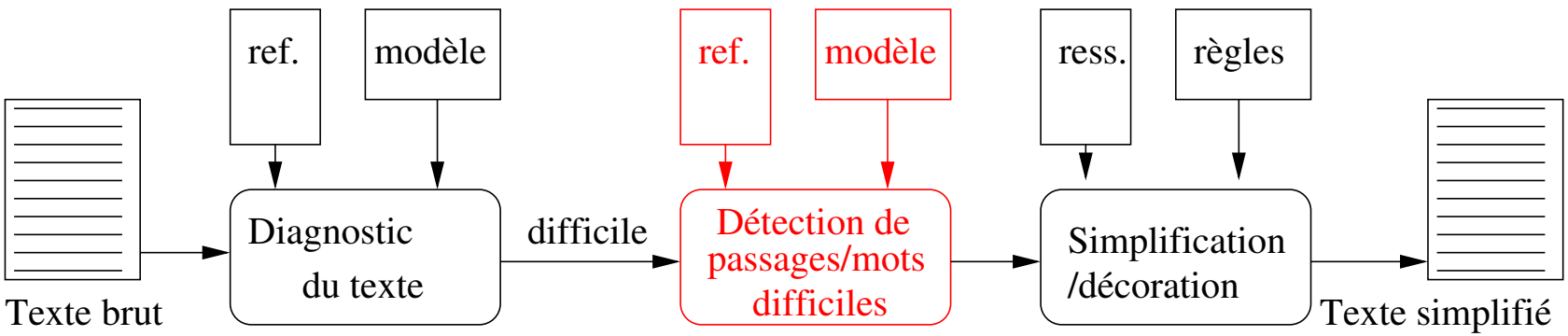
Objectifs

- Rendre les documents de santé mieux compréhensibles par les patients



Détection de mots/passages difficiles

1. Contexte
2. Détection de mots/passages difficiles
   (Grabar et al., 2014)
3. Acquisition de ressources pour la simplification
4. Conclusion



Détection de mots/passages difficiles

Objectifs : détecter les mots difficiles à comprendre

Histoire de la maladie
Le patient a été hospitalisé le 18/7/11 à PELLEGRIN pour un AVC ischémique dans le territoire profond de l'artère cérébrale postérieure droite, thrombolysé à H+3.

Le patient présente, comme déficit, une hypoesthésie gauche et une parésie gauche (force motrice à 1/5 au membre supérieur gauche et 2/5 au membre inférieur gauche), un NIHSS à 8, une désorientation tempora-spatiale et une vigilance fluctuante.
Dans les suites, est survenu un OAP post thrombolyse, probablement iatrogène (scanner injecté et NaCl afin de visualiser la zone de thrombolyse).

Le patient est donc transféré en réanimation : l'OAP est résolutif sous VNI et oxygénothérapie.

La majoration de l'insuffisance rénale nécessite 2 cures de dialyse.
Mr K. est ensuite transféré en post-réanimation devant l'évolution favorable et revient en service de neurologie à Pellegrin pour suite de la prise en charge.

Le 11/8/2011, le patient présente une douleur thoracique associée à une désaturation à 83 %, il est donc transféré en Unité de soins intensifs cardiologiques. Une embolie pulmonaire basale droite est mise en évidence par une scintigraphie pulmonaire. Une anticoagulation curative par CALCIPARINE est mise en place.

Détection de mots/passages difficiles

Matériel

- Objet : termes médicaux (151 104)
- Source : Snomed International (Côté, 1993)
- Unité : mots (29 641)
  - lemmes (Treetagger, Flemm)
- Approche : catégorisation supervisée
- Données de référence :
  - annotation manuelle indépendante par 3 personnes :
    - A1, A2, A3
    - unanimité
    - majorité
  - catégories :
    1. *Je peux comprendre*
    2. *Je ne suis pas sûr*
    3. *Je ne peux pas comprendre*

Détection de mots/passages difficiles

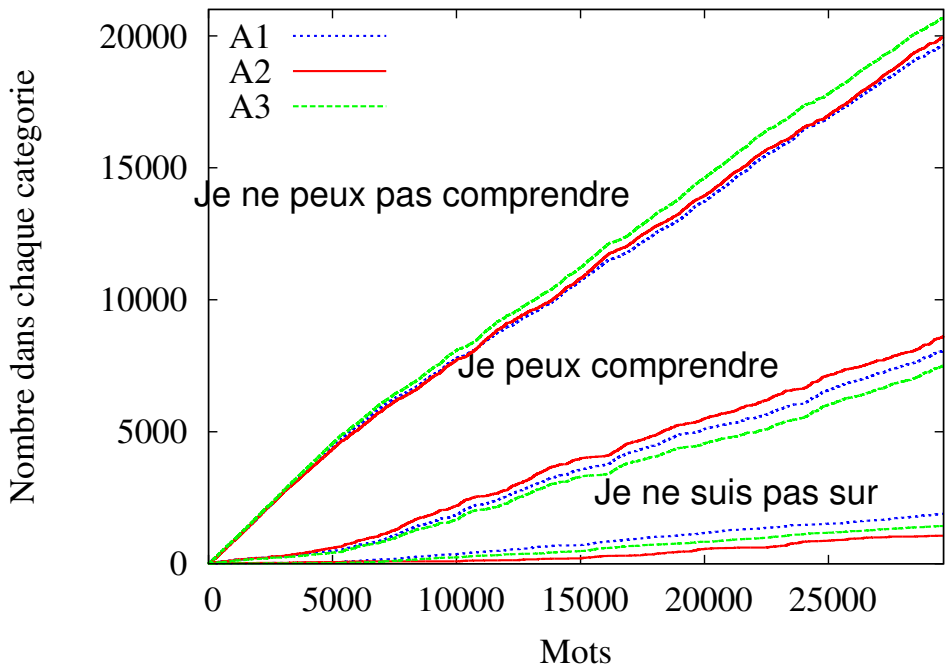
Matériel

- composés (*abdominoplastie*, *dermabrasion*)
- construits (*cardiaque*, *acineux*, *lipoïde*)
- simples (*acné*, *fragment*)

Détection de mots/passages difficiles

Annotation

Cat.	*A1 (%)*	*A2 (%)*	*A3 (%)*	*Unan. (%)*	*Major. (%)*
1.	8 099 (28)	8 625 (29)	7 529 (25)	5 960 (26)	7 655 (27)
2.	1 895 (6)	1 062 (4)	1 431 (5)	61 (0,3)	597 (2)
3.	19 647 (66)	19 954 (67)	20 681 (70)	16 904 (73,7)	20 511 (71)
Total	29 641	29 641	29 641	22 925	28 763



Accord inter-annotateur : Kappa Fleiss 0.735, Kappa Cohen 0.736

Détection de mots/passages difficiles

Descripteurs

24 descripteurs linguistiques et extra-linguistiques :

- *Catégories syntaxiques.* TreeTagger (Schmid, 1994) et Flemm (Namer, 2000) (noms, adjectifs, noms propres, verbes, abréviations) ;
- *Lexiques de référence.* TLFi et lexique.org ;
- *Fréquence sur un moteur de recherche* ;
- *Fréquence dans la terminologie medicale* ;
- *Nombre de types sémantiques* ;
- *Longueur de mots* (nombre de caractères et syllabes) ;
- *Nombre de bases et affixes.* Analyseur morphologique Dérif (Namer, 2003) ;
- *Chaînes initiales et finales.* 3 à 5 caractères ;
- *Nombre et % de consonnes, voyelles et autres caractères* ;
- *Scores de lisibilité classiques.* (Flesch, 1948) et Flesch-Kincaid (Kincaid, 1975).

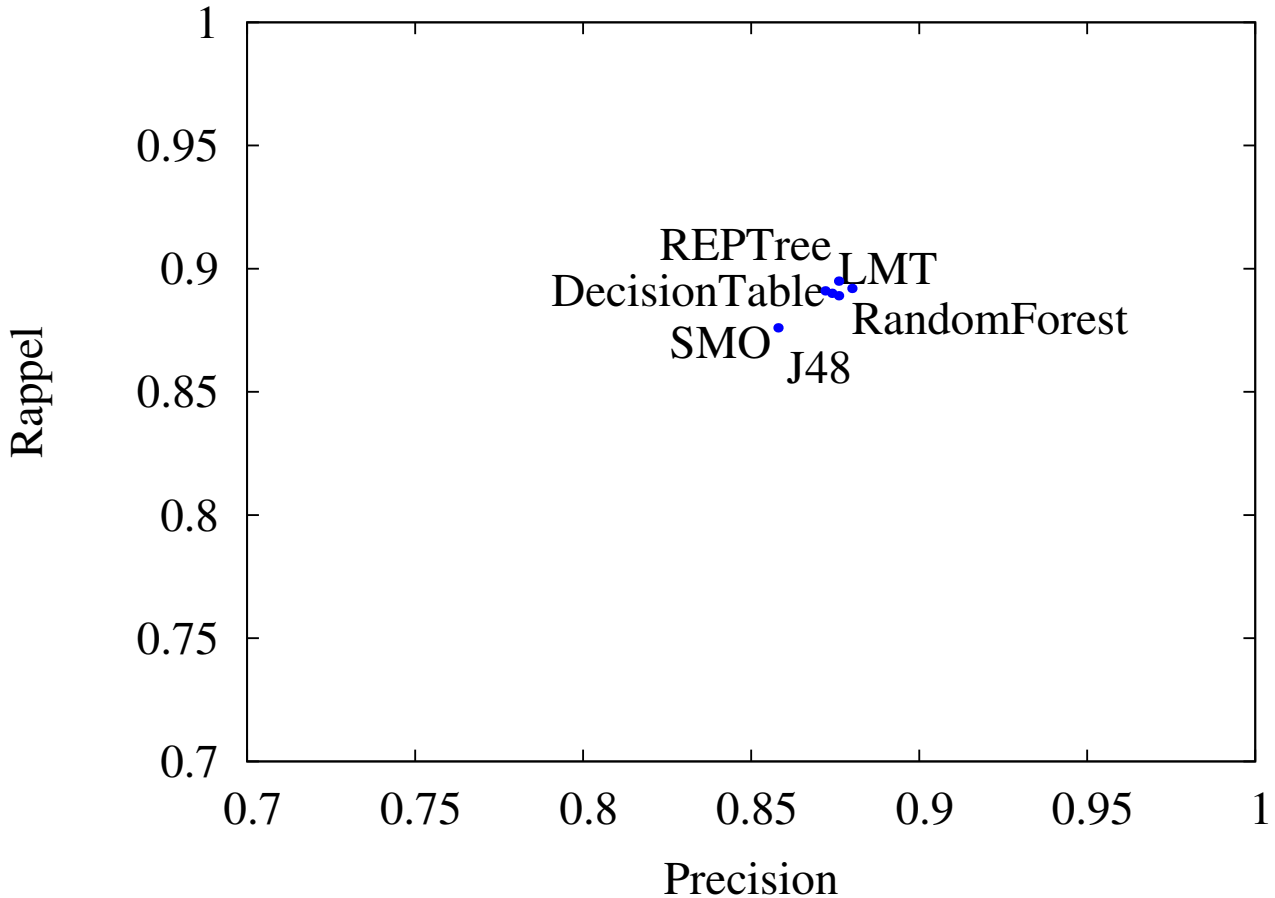
Détection de mots/passages difficiles

Catégorisation supervisée

- Catégorisation avec WEKA
- Cinq ensembles de référence :
  - 3 ensembles : annotations des trois annotateurs (29 641 mots),
  - ensemble *unanimité*, tous les annotateurs sont d'accord (22 925 mots),
  - ensemble *majorité*, accord majoritaire des annotateurs (28 763 mots).
- Distinction entre les mots compréhensibles et non-compréhensibles
- Pertinence des descripteurs
- Baseline : catégorisation dans la catégorie majoritaire

Détection de mots/passages difficiles

Résultats de la catégorisation



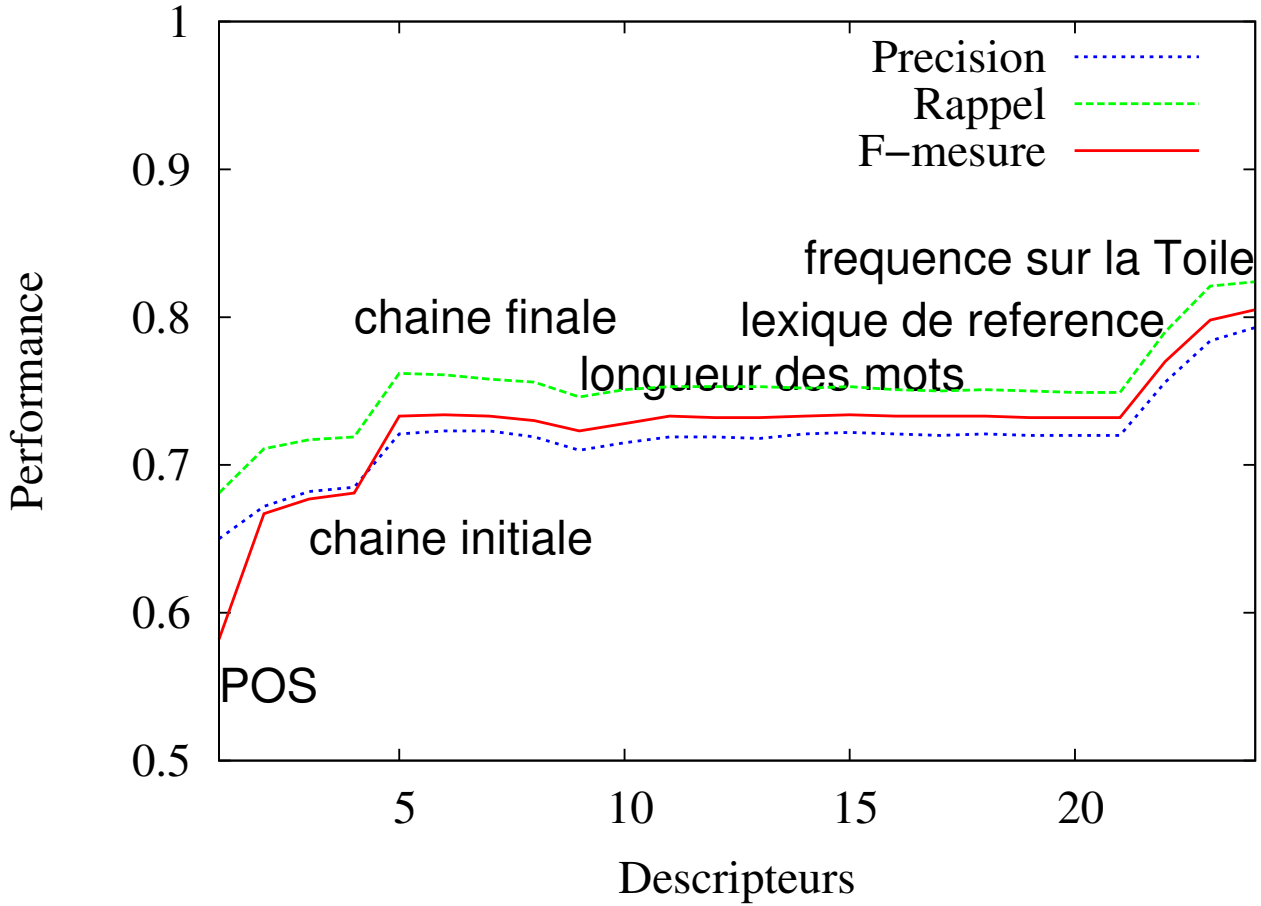
Détection de mots/passages difficiles

Résultats de la catégorisation J48

	A1	*A2*	*A3*	*Una.*	*Maj.*
$\mathcal{P}$	0.794	0.809	0.834	0.946	0.876
$\mathcal{R}$	0.825	0.826	0.862	0.949	0.889
$\mathcal{F}$	0.806	0.814	0.845	0.947	0.881
BL	0.66	0.67	0.70	0.74	0.71
gain	0.14	0.13	0.14	0.20	0.16

Détection de mots/passages difficiles

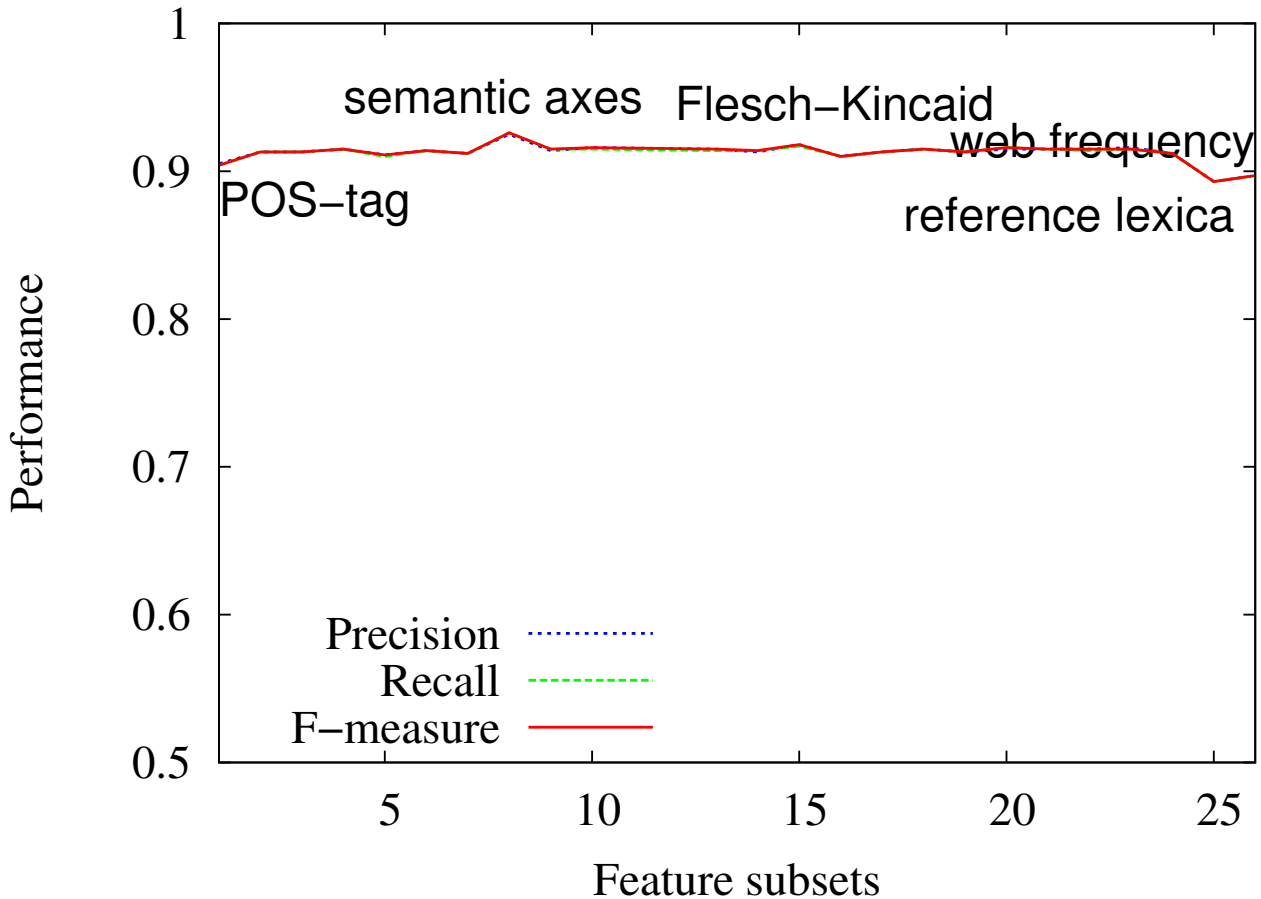
Ajout incrémental des descripteurs



- scores de lisibilité, fréquence dans la terminologie, nombre de types sémantiques

Détection de mots/passages difficiles

Take one out



- types sémantiques, scores de lisibilité

Détection de mots/passages difficiles

Texte brut

Histoire de la maladie
Le patient a été hospitalisé le 18/7/11 à PELLEGRIN pour un AVC ischémique dans le territoire profond de l'artère cérébrale postérieure droite, thrombolysé à H+3.

Le patient présente, comme déficit, une hypoesthésie gauche et une parésie gauche (force motrice à 1/5 au membre supérieur gauche et 2/5 au membre inférieur gauche), un NIHSS à 8, une désorientation tempora-spatiale et une vigilance fluctuante.
Dans les suites, est survenu un OAP post thrombolyse, probablement iatrogène (scanner injecté et NaCl afin de visualiser la zone de thrombolyse).

Le patient est donc transféré en réanimation : l'OAP est résolutif sous VNI et oxygénothérapie.

La majoration de l'insuffisance rénale nécessite 2 cures de dialyse. Mr K. est ensuite transféré en post-réanimation devant l'évolution favorable et revient en service de neurologie à Pellegrin pour suite de la prise en charge.

Le 11/8/2011, le patient présente une douleur thoracique associée à une désaturation à 83 %, il est donc transféré en Unité de soins intensifs cardiologiques. Une embolie pulmonaire basale droite est mise en évidence par une scintigraphie pulmonaire. Une anticoagulation curative par CALCIPARINE est mise en place.

Détection de mots/passages difficiles

Texte annoté

Histoire de la maladie
Le patient a été hospitalisé le 18 / 7 / 11 à PELLEGRIN pour un AVC ischémique dans le territoire profond de l' artère cérébrale postérieure droite , thrombolysé à H + 3 .

Le patient présente , comme déficit , une hypoesthésie gauche et une parésie gauche (force motrice à 1 / 5 au membre supérieur gauche et 2 / 5 au membre inférieur gauche) , un NIHSS à 8 , une désorientation tempora-spatiale et une vigilance fluctuante .
Dans les suites , est survenu un OAP post thrombolyse , probablement iatrogène (scanner injecté et NaCl afin de visualiser la zone de thrombolyse) .

Le patient est donc transféré en réanimation : l' OAP est résolutif sous VNI et oxygénothérapie .

La majoration de l' insuffisance rénale nécessite 2 cures de dialyse . Mr K . est ensuite transféré en post-réanimation devant l' évolution favorable et revient en service de neurologie à Pellegrin pour suite de la prise en charge .

Le 11 / 8 / 2011 , le patient présente une douleur thoracique associée à une désaturation à 83 % , il est donc transféré en Unité de soins intensifs cardiologiques . Une embolie pulmonaire basale droite est mise en évidence par une scintigraphie pulmonaire . Une anticoagulation curative par CALCIPARINE est mise en place .

Détection de mots/passages difficiles

Analyse des erreurs et des limites

- Entités nommées (*France*, *Indiana*, *Nancy*, *Tokyo*)
  - OK pour les annotateurs, KO pour la catégorisation
- Anatomie humaine (*cloacal*, *pubovaginal*, *nasopharyngé*, *mitral*, *diaphragmatique*, *inguinal*, *strontium*, *érythème*)
  - très souvent : incompréhensibles pour les annotateurs
- Composés (*antihémophile*, *pseudohémophilie*, *sclérodermie*, *hydrolase*, *orthotopique*, *tympanectomie*, *arthrodèse*, *synesthésie*)
  - considérés comme compréhensibles à tord
- Mots avec - (*intestin-côlon*, *semi-fermé*, *post-cataracte*, *non-réponse*, *non-érotique*, *celle-ci*, *sous-rétinien*)
  - considérés comme non compréhensibles à tord
- Fautes d'orthographe (*oreillle*, *épaiseur*)
- Formes fléchies et dérivées
- Entités syntaxiquement complexes (*AVC ischémique*, *embolie pulmonaire basale*, *scintigraphie pulmonaire*)

Détection de mots/passages difficiles

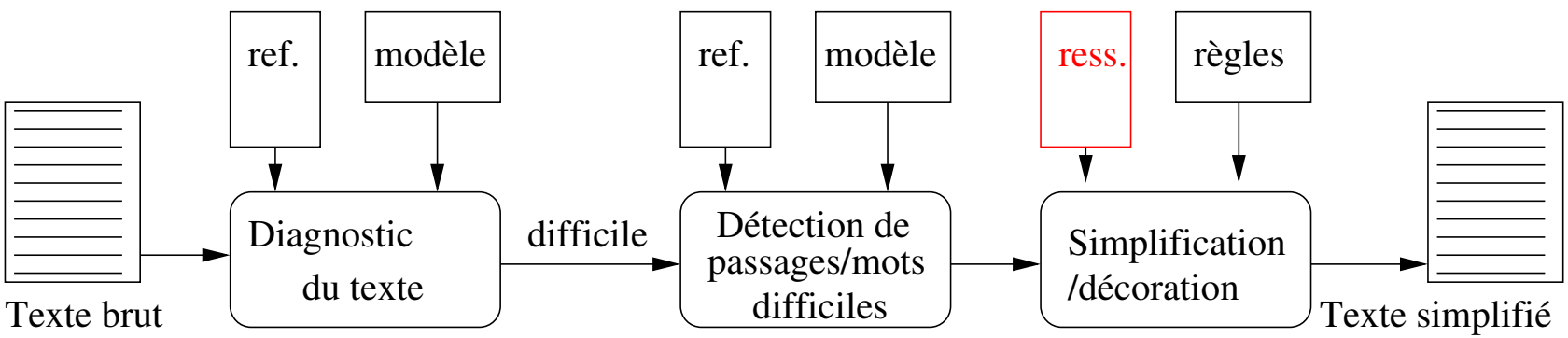
Bilan

- Catégorisation des mots en compréhensible ou non
- Apprentissage supervisé
- Bonnes performances
- Lien avec le texte
- Limites :
  - entités syntaxiquement complexes
  - formes fléchies et dérivées
  - orthographe
  - ...

Acquisition de ressources pour la simplification

1. Contexte
2. Détection de mots/passages difficiles
3. Acquisition de ressources pour la simplification

   (Grabar and Hamon, 2014 ; Grabar and Hamon, 2015 ; Antoine, 2015)
4. Conclusion



Acquisition de ressources pour la simplification

Motivation

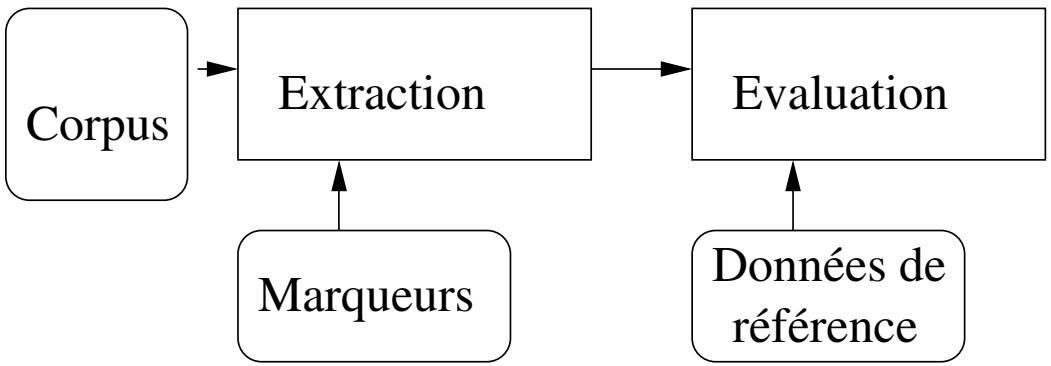
- Besoin d'avoir des ressources dédiées
- " Traduire " les termes difficiles
- Souvent, des glossaires {*difficile*, *facile*}
  - {*myocardial infarction*, *heart attack*}, {*abortion*, *termination of pregnancy*}, {*acrodynia*, *pink disease*} (Zeng et al., 2006)
  - {*consommation régulière*, *consommer de façon régulière*}, {*gêne à la lecture*, *empêche de lire*}, {*évolution de l'affection*, *la maladie évolue*} (Deleger & Zweigenbaum, 2008)
  - {*retard de cicatrisation*, *retarder la cicatrisation*}, {*apports caloriques*, *apport en calories*}, {*calculer les doses*, *doses sont calculées*}, {*efficacité est renforcée*, *renforcer son efficacité*} (Cartoni & Deléger, 2011)
  - {*myocardial infarction*, *heart attack*}, {*SBP*, *systolic blood pressure*}, {*atrial fibrillation*, *arrhythmia*}, {*hypercholesterolemia*, *high cholesterol*}, {*mental stress*, *stress*} (Elhadad et al., 2007)

Acquisition de ressources pour la simplification

Expériences

1. Contextes définitoires
   (Grabar and Hamon, 2014 ; Grabar and Hamon, 2015)
2. Compositionnalité morphologique des termes
3. Reformulations

Contextes définitoires



Contextes définitoires

Matériel

- Termes :
  - Snomed International (Cote, 1997), partie française d'UMLS (Lindberg et al., 1993)
    - mots des termes
    - pas de nombres
- Corpus :
  - Wikipédia, Portail de la Médecine
    - version de janvier 2015
    - 18 434 articles
    - 15 235 219 occurrences

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 (Pery-Woodley et al., 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

Bilan

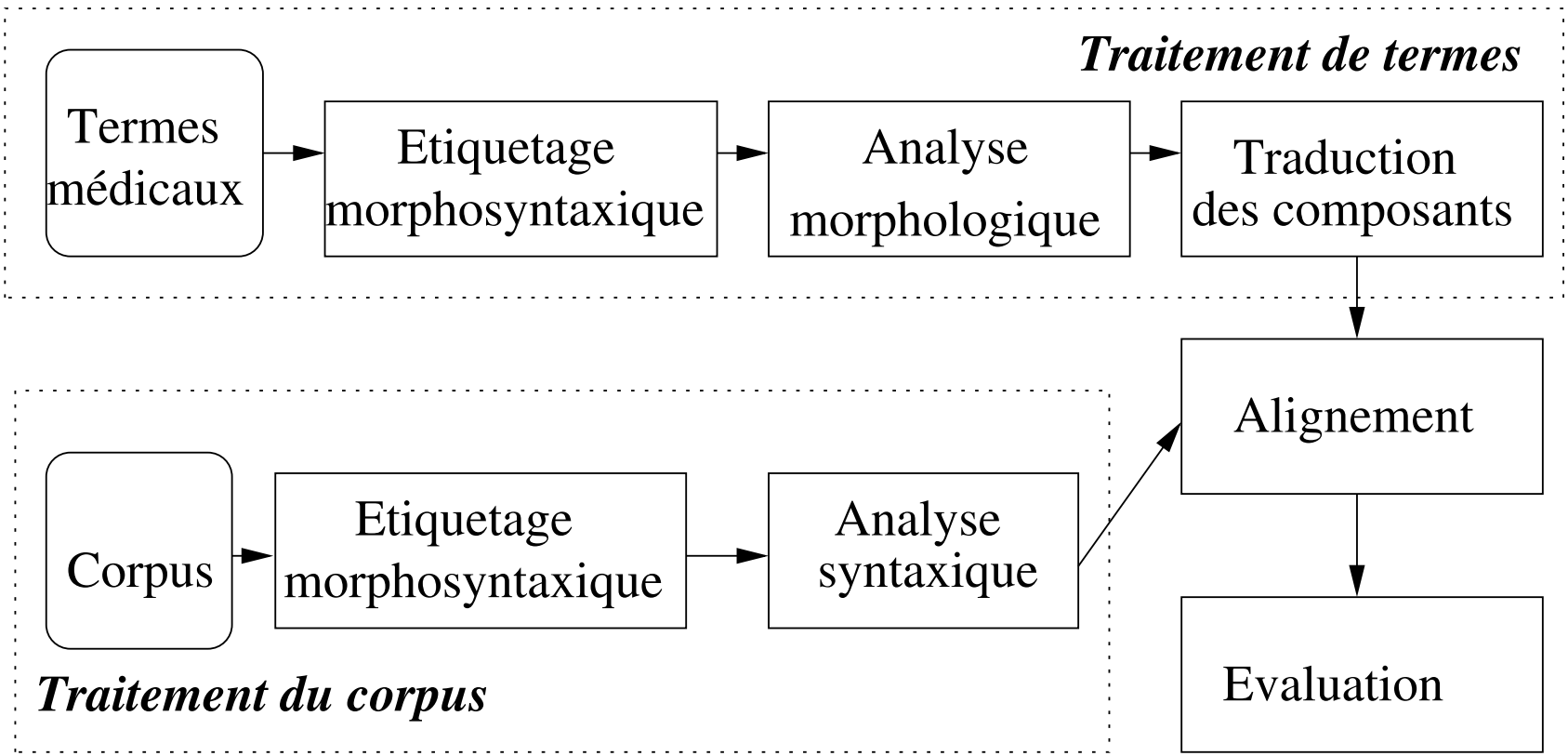
- Acquisition de définitions de termes médicaux
- Différents types de termes
  - non construits, affixés, composés néoclassiques
- Résultats :
  - jusqu'à 1 028 termes
- Précision :
  - stricte : 52,5 %
  - lâche : 68 %

Acquisition de ressources pour la simplification

Expériences

1. Contextes définitoires
2. Compositionnalité morphologique des termes
   (Grabar and Hamon, 2014 ; Grabar and Hamon, 2015)
3. Reformulations

Composition morphologique



Composition morphologique

Matériel

- Termes :
  - Snomed International (Cote, 1997), partie française d'UMLS (Lindberg et al., 1993)
    - mots des termes
    - pas de nombres
- Corpus :
  - Wikipédia, Portail de la Médecine
    - version de janvier 2015
    - 18 434 articles
    - 15 235 219 occurrences
- Ressources linguistiques :
  - liste de mots de vides
  - morphologique : 163 823 paires de mots (dérivations, flexions)

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, 2009)
  - *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) :

  $E1$: terme et syntagme complets dans l'alignement :
  - {*myo pathie*, *maladie du muscle*}

  $E2$: terme complet, syntagme partiel :
  - {*myo pathie*, *maladie du muscle cardiaque*}

  $E3$: terme partiel, syntagme complet :
  - {*myopathie*, *la maladie*}

  $E4$: 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 morphologique : 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 et al., 2014)

Composition morphologique

Bilan

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

Acquisition de ressources pour la simplification

1. Contextes définitoires
2. Compositionnalité morphologique des termes
3. Reformulations

   (Antoine, 2015)

Reformulations

Hypothèse

- Paraphrase : un même concept exprimé avec des moyens linguistiques différents :
  - *Google a acheté Youtube* $\rightarrow$ *Youtube a été vendu à Google*
- Reformulation : redire différemment ce qui a déjà été dit (Richard, 2008)
- Présence de reformulations :
  - indique les mots/termes difficiles
  - offre les indices pour l'extraction

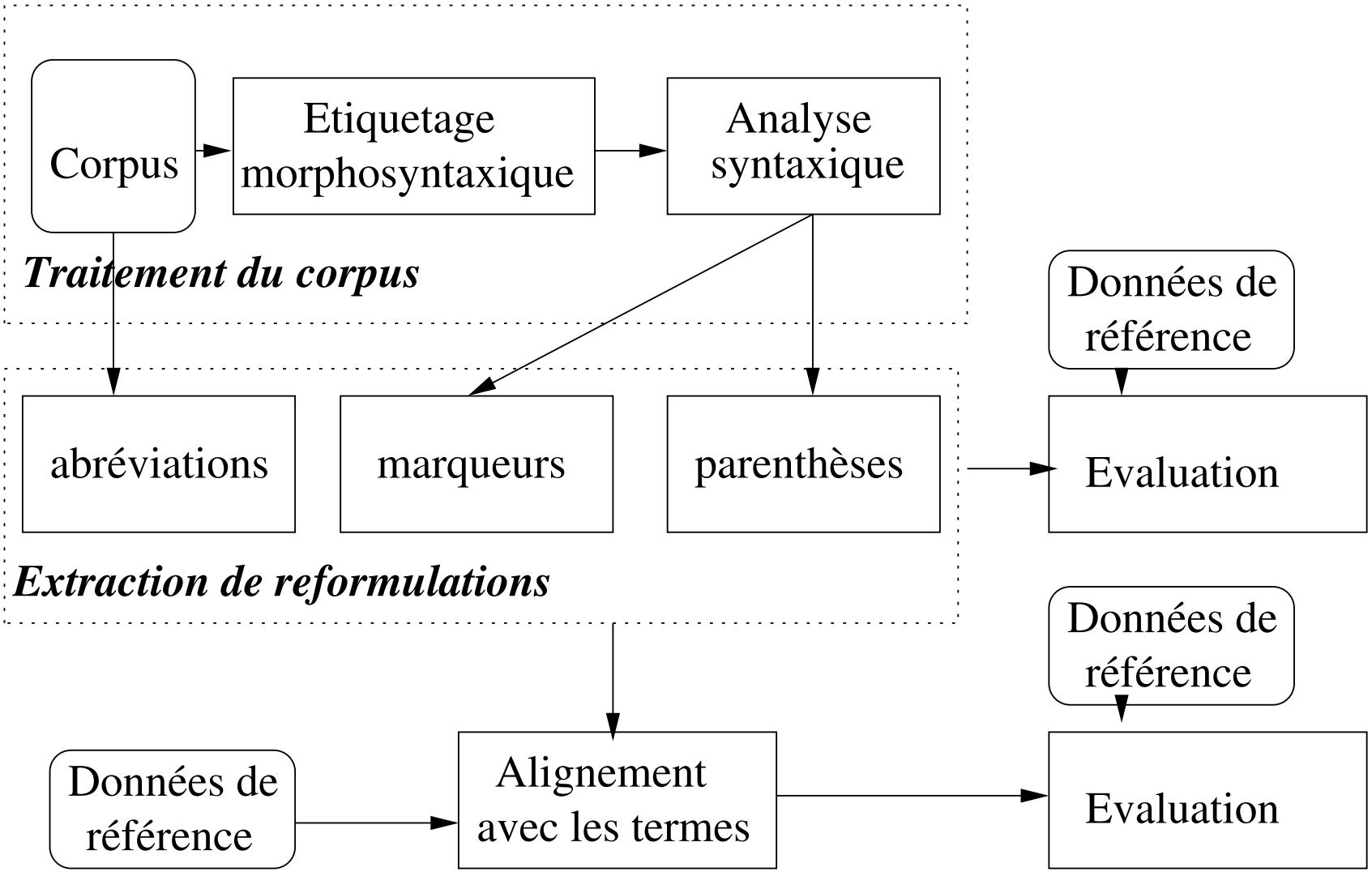
Reformulations

Corpus et ressources

- Corpus : monolingues simples, discours médical
  - développement : forum *masante.net*
    - 6 139 réponses, 315 362 occurrences
      Cher(e) pseudonyme, réponse du médecin. *Bien cordialement. Ceci n'est pas une consultation médicale et n'a pas pour objet de la remplacer.*
  - test : Wikipédia, Portail de la Médecine
    - 18 434 articles, 15 235 219 occurrences
- Ressources linguistiques :
  - liste de mots de vides
  - morphologique : 163 823 paires de mots (dérivations, flexions)
- Terminologie médicale en français :
  - UMLS : Unified Medical Language System (Lindberg et al, 1993)
  - SNOMED Int : Systematized Nomenclature of Medicine (Côté, 1993)

Reformulations

Schéma général de l'approche



Reformulations

Extraction de siglaisons et de leurs formes étendues

- Inspiré de (Schwartz et al, 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

Extraction de reformulations avec marqueurs

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
- Récupération du concept et de la reformulation :
  - informations syntaxiques

Reformulations

Extraction de reformulations avec marqueurs

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

Vous ne devez pas employer de savons ou des laits sophistiqués, c'est-à-dire, contenant plusieurs composants

Reformulations

Extraction de reformulations avec marqueurs

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

Vous ne devez pas employer de savons ou des laits sophistiqués, c'est-à-dire, contenant plusieurs composants

Reformulations

Extraction de reformulations avec parenthèses

concept (reformulation)
avec des prélèvements (biopsie)

- Pré-traitement
- Étiquetage et analyse morpho-syntaxique de Cordial (Laurent et al, 2009)
- Déclencheur : parenthèses
- Filtres pour limiter le bruit :
  - *un problème hormonal (thyroïde, surrénale)*
- Extraction :
  - concept : informations syntaxiques
  - reformulation : entre parenthèses

Reformulations

Extraction de reformulations avec parenthèses

forme	*lemme*	*POS*	*POSMT*	*GS*	*type GS*	*Prop*
une	un	DETIFS	Da-fs-i	13	N	2
gastroscopie	gastroscopie	NCFS	Ncfs	13	N	2
avec	avec	PREP	Sp	16	H	2
des	de le	DETDPIG	Da-.p-i	16	H	2
prélèvements	prélèvement	NCMP	Ncmp	16	H	2
(	(	PCTFAIB	Ypo	-	-	2
biopsie	biopsie	NCFS	Ncfs	18	N	2
)	)	PCTFAIB	Ypc	-	-	2
.	.	PCTFORTE	Yps	-	-	-

une gastroscopie avec des prélèvements (biopsie)

Reformulations

Évaluation des extractions

- Préparation des données de référence des 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

Alignement avec une terminologie médicale

- Comparaison des segments extraits avec les termes de la terminologie médicale :
  - évite les extractions non pertinentes :
    - *en fibres (pas trop vite sinon vous serez ballonnée)*
  - fait ressortir les segments pertinents et exploitables
- Méthode :
  - casse, désaccentuation, normalisation morphologique
  - suppression des mots vides
  - choix du taux d'alignement : segments, termes

Reformulations

Évaluation de l'alignement

- Données de référence :
  - à partir de l'alignement aux seuils 40/40, corpus de développement
  - deux annotateurs, consensus
- Mesure d'évaluation : précision
- Définition des seuils sur le corpus de développement
- Application de ces seuils sur le corpus de test

	Alignement
Abréviations	0,208
Marqueurs	0,714
Parenthèses	0,817

Reformulations

Résultats : Extractions des siglaisons

	Développement			*Test*		
	Abrév	*Marq*	*Par*	*Abrév*	*Marq*	*Par*
nb occ.	75	96	312	88 762	2 757	100 103
nb types/ref.	42	96	305	8 106	2 710	92 971

- 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)*

Résultats

Extractions des reformulations avec marqueurs

	Développement			*Test*		
	Abrév	*Marq*	*Par*	*Abrév*	*Marq*	*Par*
nb occ.	75	96	312	88 762	2 757	100 103
nb types/ref.	42	96	305	8 106	2 710	92 971

- Trois marqueurs :
  - *c'est-à-dire* : 80/1 929
  - *a-Autrement dit* : 8/145
  - *encore appelé(e)(s)* : 8/86
- Difficultés : détection de frontières
  - *une toxi-infection, c'est-à-dire, qu' elle peut*
  - *Une salpingite, c'est-à-dire, une inflammation des trompes est possible*
  - *en, c'est-à-dire, au contact du sang circulant*
  - *des dilations des canaux galactophores, c'est-à-dire qui fabriquent le lait*

Résultats

Extractions des reformulations avec parenthèses

	Développement			*Test*		
	Abrév	*Marq*	*Par*	*Abrév*	*Marq*	*Par*
nb occ.	75	96	312	88 762	2 757	100 103
nb types/ref.	42	96	305	8 106	2 710	92 971

- Difficultés :
  - frontière des concepts :
    - *une greffe de valve prothétique (valve mécanique artificielle)*
    - *se bouche (hémorroïdes)*
  - extractions non pertinentes :
    - *énergétique (carence plutôt liée au marasme)*

Évaluation des extractions

Précision, rappel et F-mesure des extractions pour chaque méthode

	Abréviations			*Marqueurs*			*Parenthèses*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
exact	0.74	0.74	0.74	0.24	0.24	0.24	0.23	0.23	0.23
inexact	0.94	0.94	0.94	0.98	0.98	0.98	0.68	0.68	0.68

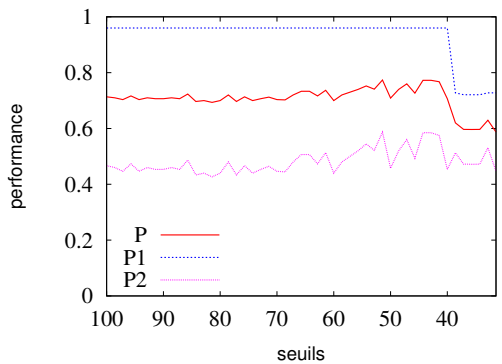
- corpus de développement
- données de référence consensuelles
- script d'évaluation : DEFT 2015, tâche 3
- Bilan :
  - fiabilité des extractions de siglaisons
  - pertinence des reformulations avec marqueurs
  - bruit des reformulations avec parenthèses
  - recouvrement entre marqueurs et parenthèses : 0,007%

Alignement avec la terminologie

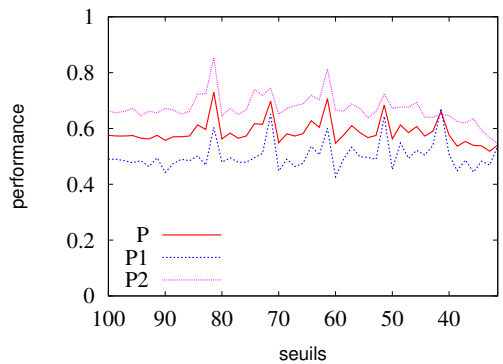
- Types d'alignement de termes :
  - proposition pertinente :
    - *communication interventriculaire : communication interventriculaire.C0018818...*
  - proposition avec variation morpho-syntaxique :
    - *troubles gastrointestinaux fonctionnels/C0559031.T047.DISO*
    - *troubles gastro intestinaux fonctionnels/C0559031.T047.DISO*
  - proposition partielle :
    - *semaines amenorrhee : amenorrhee/C0002453.T047.DISO*
  - proposition compositionnelle (*cause de pus*) :
    - *cause/C0085978.T078.CONC/...*
    - *pus/C0034161.T031.ANAT/...*
  - Proposition non pertinente :
    - *LCR : ph lcr/C0853364*
    - *liquide cerebro : regime liquide/C-F2300*
  - Aucune proposition : *NFS : —*

Alignement avec la terminologie

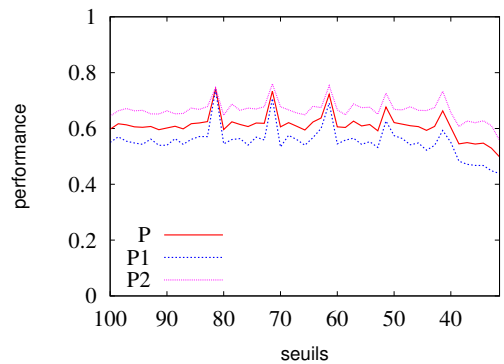
définition des seuils des alignements



(a) Abréviations



(b) Marqueurs



(c) Parenthèses

- corpus de développement
- données de référence consensuelles
- précision des segments alignés :
  - par segment, moyennes

Alignement avec la terminologie

	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 : –*

Typologie des reformulations

- Typologie de l'état de l'art (Bhagat et al, 2013)
- 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

Discussion

- Exploitation de reformulation pour l'acquisition du vocabulaire
- 3 méthodes :
  - abréviation : inspiré de l'algorithme proposé par (Schwartz et al, 2003)
  - 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

Discussion

- 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

Acquisition de ressources pour la simplification

Bilan

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

Acquisition de ressources pour la simplification

Bilan

Comparaison avec les travaux existants

	type terme	nb. para	précision
(Zeng et al., 2005)	tous	CHV	
(Elhadad et al., 2007)	tous	152	0,58
(Deleger et al., 2008)	m-synt.	65, 82	0,67, 0,60
(Cartoni et al., 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

Acquisition de ressources pour la simplification

Bilan

Comparaison avec les travaux existants

- `DériF` (Namer, 2009) :
  - 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*

Conclusion et Travaux futurs

Conclusion générale

- Différents aspects menant vers la simplification de textes
  - documents de spécialité
  - médecine
- Méthodes
  - diagnostic de textes
  - diagnostic de passages/mots non compréhensibles
- Ressources
  - plusieurs méthodes
  - évaluation des extractions
  - alignement avec les terminologies

Travaux futurs

- Améliorer la détection de mots incompréhensibles :
  - entités syntaxiquement complexes
  - formes fléchies et dérivées
  - orthographe
- Augmenter la couverture des paraphrases :
  - d'autres corpus
  - ressources supplétives plus couvrantes
  - d'autres méthodes pour extraire des paraphrases
  - gérer les paraphrases concurrentes
  - combinaison avec les images
- D'autres langues
- Simplification lexicale de textes médicaux
- Évaluation avec des utilisateurs