I'm not robot	reCAPTCHA

Continue

Les analysis medicales nfs pdf free pdf files

À l'inverse, l'étude des terminologies médicales existantes peut aider à constituer des connaissances linguistiques, en particulier morpholexicales. 12Si l'on effectue le même type d'expérience sur les libellés d'une terminologie médicale, la proportion de mots inconnus augmente. Dans ce contexte, le traitement automatique des langues cherche à intervenir de plusieurs façons. Une deuxième voie pourrait être une reprise de l'approche "complète" de l'analyse de textes, mais dans une version plus "robuste". Fabre (1999) : Elementary dependency trees for identifying corpus-specific semantic classes. Ainsi, sur les 16132 termes de la version française du thésaurus MeSH, 20 % des 29604 occurrences de mots sont inconnus du DELAF (36 % des 13255 formes différentes). Columbia, Maryland, Defense Advanced Research Projects Agency: Morgan Kaufmann. (1999): A knowledge-free method for capitalized word disambiguation. ZELLIG construit un graphe dans lequel les mots dont la similarité est supérieure à un seuil donné sont liés par une arête. Methods of Information in Medicine, à paraître. Le tableau 3 montre des exemples d'expressions source et des termes normalisés correspondants dans le référentiel cible. Ce réseau constitue un modèle dans lequel on va trouver des représentations élémentaires pour des configurations de concepts.59Dans les expressions une angioplastie du segment II, une angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de Monsieur X, une angioplastie de Monsieur X, une angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie du segment II, une angioplastie de Monsieur X, une angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de Monsieur X, une angioplastie de Monsieur X, une angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de Monsieur X, une angioplastie de Monsieur X, une angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose, la relation sémantique précise qui lie l'angioplastie de la sténose qui lie l'angioplastie de la s représentée par le graphe suivant :63Les performances du prototype construit en fin de projet ont fait l'objet d'une évaluation formelle. IBM Journal of Research and Development 30 (1), 70-79. Grabar, N. 56Nous nous concentrons ici sur la partie concernant l'analyse du français. Computers and the Humanities 33 (3), 207-219. Harris, Z., M. En effet selon la tâche visée (aide au codage, etc.) et son besoin informationnel spécifique (repérage de diagnostics, d'actes, etc.), on pourrait cantonner la description des connaissances sur le domaine et la construction d'une représentation conceptuelle au champ sémantique correspondant à cette information cible. 50Nous avons testé cette procédure en prenant pour classes sémantiques les 11 catégories de la nomenclature SNOMED. La représentation des connaissances de la famille des réseaux sémantiques. Ces transformations, de nature morphologique, syntaxique et sémantique, sont contrôlées de façon à conserver une proximité suffisante entre termes normalisés et variantes. P. Nous avons employé pour cela une procédure de vote qui attribue au mot inconnu la classe la plus représentée parmi ses voisins. Nous avons testé plusieurs mesures de similarité entre ensembles de contextes. A Guide for Users. Ici encore, plusieurs champs sémantiques sont concernés. À chaque objectif correspond un type de terminologie. Comme nous l'avons noté plus haut, la plupart des grandes terminologies médicales existent en plusieurs langues. Deliverable report AIM-MENELAS 17, DIAM-SIM/INSERM U.194.Zweigenbaum, P. Peu d'analyseurs à large couverture étaient disponibles à l'époque pour le français. L'objectif du système d'analyse était la construction d'une représentation conceptuelle normalisée du contenu des textes : une représentation abstraite des variations d'expression à l'intérieur d'une langue (paraphrase) ou d'une langue (paraphrase) ou d'une langue à l'autre. Gottfried, T. Rothwell, J. Leur combinaison permet d'augmenter sensiblement cette précision, tout en conservant un très bon rappel. Une description étendue de ce travail dépasserait le cadre de cet article; nous en donnons les grandes lignes, et renvoyons le lecteur à Zweigenbaum (1998) pour une synthèse et à Zweigenbaum (1998) pour une synthèse et à Zweigenbaum (1998) raitement automatique de la langue médicale. medicine: an overview. À notre connaissance, et malgré différentes limitations, il s'agit toujours du seul exemple d'une évaluation d'un système de traitement automatique de la langue médicale produisant une représentation conceptuelle normalisée, dans laquelle la notation est effectuée automatiquement plutôt que par le jugement a posteriori d'un expert. In J.-I. 41Une liste de noms de micro-organismes est facile à obtenir : on en trouve une dans la SNOMED ou dans l'UMLS. Un mot inconnu étant donné, l'idée est de déterminer sa classe selon celle de ses voisins dans le graphe. 38Les critères sur lesquels nous nous appuyons sont les suivants : commencer par une majuscule dans toutes ses occurrences dans la terminologie considérée; ne pas être un symbole ou une abréviation (SIDA, B12); ne pas être un nom de genre de micro-organisme, qui par convention est toujours capitalisé (Hæmophilus influenzæ); être invariant à travers plusieurs traductions (Sarcome de Kaposi de la peau, Kaposi's sarcoma of skin, Kaposi-Sarkom der Haut); apparaître à des positions spécifiées (_np) dans un patron lexico-syntaxique : 39contexte... (de ? Il est possible d'identifier encore davantage de couples initiaux. Il reste donc à mettre au point une stratégie de combinaison de ces deux types de méthodes, qui soit capable de détecter les cas où une simple reconnaissance de variantes lexicales suffit et ceux où une analyse plus complète est nécessaire.65Notons que des approches intermédiaires entre un simple traitement morpho-lexical et une analyse complète existent et font l'objet de recherches actives. Mémoire de HDR, Université de Nantes.Lindberg D.A.B., B.L. Humphreys, & A.T. McCray (1993): The Unified Medical Language System. Zweigenbaum (1999): A lexical method for assisted extraction and coding of ICD-10 diagnoses from free text patient discharge summaries. On trouve par exemple des relations de synonymie dans la SNOMED, des relations hiérarchiques (hyperonymie, méronomie) dans CIM-10, SNOMED et MeSH, des relations transversales (décomposition ou "référence") dans la SNOMED. D'autre part, les syntagmes nominaux du corpus ; et une exploitation de syntagmes verbaux, avec en particulier les relations verbe-objet et sujet verbe, renforcerait sans doute la modélisation de la distribution des mots. Cette aide au codage vise à faciliter le recueil des données nécessaires au traitement de l'information médicale. Le corpus de départ était celui rassemblé pour le projet MENELAS dans le domaine des maladies coronariennes (voir la section 4.2.), constitué de comptes rendus d'hospitalisation et de lettres de médecins hospitaliers à leurs collègues "de ville" (Zweigenbaum et Consortium MENELAS, 1995). De fait, aucune liste complète de ces noms propres n'est disponible sous forme électronique. Des heuristiques permettent de classer ces chemins ; elles prennent en compte, entre autres critères, la nature de la relation grammaticale présente. Il reprend les grandes classes du MeSH, mais aussi des sous-classes plus fines ; par exemple, l'anatomie (structures anatomiques complètement formées) est subdivisée en organes et parties du corps, cellules, tissus. Les connaissances morphologiques mentionnées à la section 3.2 sont alors utiles. (1998/1999) : Des lexiques pour la terminologie médicale. Par exemple (Nazarenko et al.,2000, figure 4), le mot bilan (qui a ici le sens de batterie d'examens) se trouve relié à examen, épreuve, contrôle, exploration, coronarographie (une sorte de radiographie dynamique du cœur et des artères coronaires) et plan ; ce dernier mot est un artefact issu de l'expression sur le plan Adj qui introduit souvent un adjectif précisant les caractéristiques d'un examen (sur le plan biologique, sur le plan angiographique, sur le plan angiographique, sur le plan électrique, etc.). Ce type de constatation contribue à justifier l'exploration de méthodes plus complexes d'analyse de textes. Beckett, & L. Northfield: College of American Pathologists. Defense Advanced Research Projects Agency (1996): MUC-6: Proceedings of the Sixth Message Understanding Conference. Elle se combine bien sûr à la dérivation, comme les deux séries ci-dessous le montrent :18La base CELEX recense ce type de données pour l'anglais médical, le lexique "Specialist" de l'UMLS et les programmes associés (le "générateur de variantes lexicales" lvg) mettent en relation les familles de mots dérivés. et P. Zweigenbaum (2000a): Automatic acquisition of domain-specific morphological resources from thesauri. Catach (1986): Conceptual graphs for semantics and knowledge processing. ZELLIG recense ainsi les différents contextes syntaxiques de chaque mot. Notre hypothèse est qu'ils ont un sens proche. La CIM-10, qui est gérée par l'Organisation mondiale pour la santé, existe en de nombreuses langues dont le français. L'ensemble de ces descriptions vise à restituer le réseau complexe de relations qui
existent entre les concepts du domaine. Methods of Information in Medicine 32 (2), 81-91. Mikheev, A. Le nombre de classes proposées va de quelques unités à une centaine. à paraître. Organisation mondiale de la Santé (1993) : Classification statistique internationale des maladies et des problèmes de santé connexes, Dixième révision. L'UMLS (Lindberg et al., 1993) définit 134 types sémantiques hiérarchisés. A. Ryckman, P. Nous avons donc recensé une série de critères complémentaires qui, lorsqu'ils sont satisfaits, tendent à montrer qu'un mot est un nom propre. Pour ce qui concerne les codes de la Classification internationale des maladies, chaque terme cible pertinent pour le domaine (une centaine de diagnostics et d'actes) a été représenté par un graphe conceptuel. & J. & C. En résumé, la précision obtenue est insuffisante pour proposer une classification automatique ; elle est cependant bien meilleure que ce que l'on obtiendrait en assignant à chaque mot la catégorie observée la plus probable (classe G) (Nazarenko et al., 2000).51Comme nous l'avons mentionné plus haut, l'un des services attendus du traitement automatique de la langue médicale est l'"aide au codage". Ce corpus compte 85000 mots. Pour ce qui concerne la tâche de réponse à 30 questions, le rappel et la précision moyens sont de 66 et 77 %. Bethesda, Maryland, National Library of Medicine. Nazarenko, A., P. L'étape d'analyse morpho-syntaxique a repris l'analyseur du projet Kalipsos du Centre scientifique d'IBM à Paris (Fargues et al., 1986). Cette tâche peut se concrétiser par le repérage, dans des textes, de termes normalisés auxquels sont associés des codes cible. Bachimont, & P. L'explosion actuelle de la quantité de documents textuels médicaux accessibles en ligne rend ce dernier aspect particulièrement important. 10Le troisième objectif est l'indexation d'articles scientifiques à des fins de recherche bibliographique. Ladeira, et J. Par exemple, nous avons appliqué le lexique DELAF du LADL (Silberztein, 1993) à dix comptes rendus d'hospitalisation en pneumologie (Zweigenbaum, 1999). Ainsi, la nomenclature SNOMED répartit ses concepts en huit classes sémantiques ("modules"): Topographie (T), Morphologie (M), Fonction (F), Organismes vivants (L), Médicaments, produits chimiques et biologiques (C), Agents, activités physiques et forces naturelles (A), Métiers et professions (J), Contexte social (S), complétées par deux classifications: Maladies et diagnostics (D), Procédures et actes professionnels (P), et un répertoire de Qualificatifs et termes relationnels (G). Les résumés des articles de la littérature scientifique biomédicale internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base Medline, sont en ligne et "internationale, indexés selon le thésaurus MeSH dans la base MeS que le "codage" d'informations médicales spécifiques exprimées dans des textes peut aussi se concevoir comme une indexation de prêter davantage attention à la structure de la phrase entière ; et la notion de facteur de risque cardio-vasculaire mise en jeu au cinquième énoncé constitue une connaissance sur le domaine qui va au-delà d'une simple synonymie. & Consortium MENELAS: coding and information retrieval from natural language patient discharge summaries. Il est probable que dans un certain nombre de cas, des traitements lexicaux relativement simples du type de ceux indiqués dans la section précédente auraient suffi à déterminer les codes appropriés de la Classification internationale des maladies. à paraître. Habert, B. L'examen de ce graphe à des seuils donnés fait ressortir certains groupes de mots pour lesquels l'hypothèse est vérifiée de façon remarquable. 8Le premier objectif est la collecte d'un jeu relativement réduit d'informations à des fins statistiques, par exemple le diagnostic principal et les éventuels diagnostics associés. La règle ci-dessus identifie au sein de ces mots les couples suivants : 31 {méat, méatite} {médiastin, médiastinite} 32 {tympan, tympanite} {vagin, vaginite} 33 parmi lesquels {tympan, tympan, tympanite} {vagin, vaginite} 32 {tympan, tympanite} {vagin, vaginite} 33 parmi lesquels {tympan, tympanite} {vagin, vaginite} 32 {tympan, tympanite} {vagin, vaginite} 33 {tympan, tympanite} {vagin, vaginite} 33 {tympan, tympanite} {vagin, vaginite} 33 {tympan, tympanite} {vagin, vaginite} 34 {tympan, tympan, tympanite} {vagin, vaginite} 34 {tympan, tympan, tymp tympanite} et {déférent, déférentite} n'avaient pas encore été identifiés. 47ZELLIG part des arbres de constituants produits lors d'une analyse syntaxique préalable. Nous donnons ci-dessous des exemples de ces trois types de relations dans la nomenclature SNOMED (SAI signifie sans autre indication) : 21Synonymie : 22palais dur a pour synonyme voûte palatine 23Hiérarchie : 24sinusite, SAI est une sorte de maladie du sinus paranasal, SAI 25Référence : 26sinusite, SAI fait référence à sinus paranasal, SAI 27On voit que dans de tels couples de termes, il arrive que deux mots appartenant respectivement à l'un et à l'autre des deux termes soient en relation morphologique. Les arbres de constituants sont réduits en arbres de dépendance élémentaires (Habert et Fabre, 1999) suivant des patrons N Adj, NPrep N, etc. Nous avons effectué des expériences dans le cas de "mots simples" (par exemple, sténose, par opposition à des termes polylexicaux comme infarctus du myocarde). 6L'analyse de grands corpus médicaux peut aussi aider à mettre à jour ces terminologies pour suivre l'évolution des techniques et de la langue médicale, ou à construire de nouvelles terminologies pour des besoins spécifiques. Appliquée à des termes non prétraités, cette méthode propose des couples de mots liés non seulement par dérivation et composition, mais aussi par flexion. Dans les expériences que nous avons effectuées, cette technique double le nombre de couples identifiés et ne réduit que de 1 % le pourcentage de couples corrects. Dans une tâche d'aide au codage, ce type de méthode devrait apporter une meilleure précision que la méthode beaucoup plus lâche évoquée à la section 4.1 (Blanquet et Zweigenbaum, 1999). & N. Une approche exhaustive comme celle de MENELAS, qui demande le développement de grammaires et de bases de connaissances importantes, ne se justifie que pour les cas plus complexes. Comme on pouvait s'y attendre, l'examen d'une terminologie, qui concentre une grande partie des termes essentiels d'un domaine permet de recenser immédiatement un grand nombre de mots spécifiques à ce domaine. Le quatrième énoncé comporte plusieurs difficultés. (1995) pour une description extensive du projet. Linguisticae Investigationes (XXII), 383-395. Zweigenbaum, P., et Consortium MENELAS (1995) : MENELAS final report. Par exemple, le tableau 2 montre les contextes dans lesquels le mot décompensation a été trouvé dans les syntagmes nominaux du corpus : épisode de décompensation, décompensation cardiaque, etc. 48L'ensemble des contextes élémentaires d'un mot représente sa distribution syntaxique dans le corpus. Dans la mesure où le besoin informationnel est plus ouvert et le degré de détail utile peut être plus grand, une précision plus fine et la couverture d'un plus grand éventail de notions sont nécessaires. L. Le MeSH est traduit dans plusieurs langues ; sa version française a été écrite par l'INSERM, qui continue à en assurer la mise à jour.11Les lexiques électroniques "généraux" peuvent contenir de nombreux termes spécialisés. La dérivation, comme dans le reste de la langue, est employée : oesophage, oesophagien, oesophagien par les descriptions associées aux concepts de l'ontologie) un chemin reliant leurs concepts de définition (Bouaud et al., 1996). Amsterdam, John Benjamins. Chanod (1997): Incremental finite-state parsing. Dordrecht, Kluwer Academic Publishers. Jacquemin, C. J. L'identification formelle de variantes de termes est une première voie. Novembre 1995.En ligneFargues, J., M.-C. Par exemple, un compte rendu d'hospitalisation peut contenir l'expression Aspect de leucémie aiguë de type LAM 2; le terme normalisé correspondant (ici, dans un thésaurus d'hématologie) est LAM 0, LAM 1, LAM 2 (leucémie aiguë myéloblastique) auquel est associé le code 2904HM.52Les exemples cités montrent que la difficulté de la tâche est variable. Amsterdam, IOS Press. Dans la troisième
expression, carotidienne doit pouvoir être mis en correspondance avec carotide. Mémoire de HDR, Université Paris 13.En ligneZweigenbaum, P. F. Nous terminons sur des perspectives de recherches sur le thème de l'aide au codage (section 5).7Nous considérons que le recueil de l'information médicale a trois objectifs principaux. La Nomenclature systématique de la médecine (SNOMED Inter-nationale) (Côté et al., 1993) est un représentation plus exhaustive et plus détaillée que celle qui serait strictement nécessaire pour une tâche d'aide au codage (ici dans la Classification internationale des maladies). Article 1L'attraction traditionnelle de la médecine pour les technologies de pointe s'est emballée dans la seconde moitié du XXe siècle, et l'informatique ne fait pas exception : les travaux pionniers de Naomi Sager, dans les années soixante, sur l'analyse automatique de textes, ont été testés en vraie grandeur sur des textes médicaux (Sager et al., 1987); et les premiers "systèmes experts", au milieu des années soixante-dix, étaient des systèmes de diagnostic médical (Musen et van Bemmel, 1997). Mémoire de HDR, Université de Caen. Zweigenbaum, P. Computer Management of Narrative Data. La tâche la mieux identifiée est l'aide au codage, production de codes de terminologies normalisées à partir de textes. (1997) : Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus. Zweigenbaum, B. Genève, Organisation mondiale de la Santé. Sager, N., C. Dans la modélisation effectuée, ce qui est dilaté est un "segment" d'artère ((purported_obj) [Artery_Segment]) : bien que l'on parle d'angioplastie d'une artère, on ne dilate bien sûr pas toute l'artère mais seulement une portion de cette artère. Bourigault, C. Par exemple, à partir du couple {sinus, sinusite}, nous faisons l'hypothèse qu'une règle générale Œ? ite (le symbole Œ note la chaîne vide) met potentiellement en relation des formes en -ite avec de syntagmes nominaux identifiés par LEXTER, un extracteur de termes (Bourigault, 1993). 16Mentionnons enfin que nous avons préparé, semi-automatiquement ou manuellement, des lexiques pour les mots apparaissant dans plusieurs thésaurus de spécialité (Zweigenbaum, 1999).17Comme nous l'avons remarqué ci-dessus, le lexique médical a largement recours à des mots construits. Traitement automatique des langues 41 (3). Journal of the American Medical Informatics Association 7 (suppl.). L'architecture globale du prototype construit suit le découpage classique en morphosyntaxe, sémantique et pragmatique. Ces résultats doivent être pris avec précaution, que ce soit dans un sens positif ou négatif. C'est le cas pour l'évaluation de l'activité des services hospitalisation fondé sur la construction d'une représentation conceptuelle normalisée (section 4.2). La catégorie sémantique d'un mot inconnu pourrait ainsi être apparentée aux catégories sémantiques de mots connus de distribution semblable. 50n constate ainsi la coexistence de deux systèmes d'enregistrement des informations médicales : d'une part des textes, et d'autre part des données codées par des termes normalisés. Notons qu'il faudrait cependant, tout mot utile peut être abrégé dans un texte, et il serait vain de chercher à répertorier a priori dans un lexique toutes les abréviations possibles. Daladier, T. Une longueur minimale commune de 3 caractères donne de très bons résultats, une longueur minimale de 4 caractères des résultats excellents : de l'ordre de 98 % des couples de mots ainsi identifiés sont effectivement, identifies tonc des deux. 43Ces critères, pris individuellement, identifiés sont effectivement en relation morphologique : dérivation, composition, ou combinaison des deux. 43Ces critères, pris individuellement, identifiés sont effectivement en relation morphologique : dérivation, composition, ou combinaison des deux. 43Ces critères, pris individuellement, identifiés sont effectivement en relation morphologique : dérivation, composition, ou combinaison des deux. 43Ces critères, pris individuellement, identifiés sont effectivement en relation morphologique : dérivation, composition, ou combinaison des deux. 43Ces critères, pris individuellement, identifiés sont effectivement en relation morphologique : dérivation, composition, ou combinaison des deux. 43Ces critères, pris individuellement, identifiés sont effectivement en relation morphologique : dérivation des deux. 43Ces critères deux noms propres de la terminologie étudiée (bon "rappel": de 70 à 100 %), mais proposent également comme noms propres de nombreux mots qui n'en sont pas ("précision" médiocre: de 5 à 50 % des mots proposés sont effectivement des noms propres). 57Les étapes d'analyse sémantique (construction d'une représentation du sens pour chaque phrase) et pragmatique (mise en œuvre de connaissances sur le domaine, prise en compte du contexte) ont été réalisées dans un analyseur "sémantique" mis au point dans notre équipe pendant le projet. 20Nous désignons par "terminologie structurée" une terminologie dans laquelle des relations sémantiques relient certains termes. Sur 37 comptes rendus d'hospitalisation pris dans le corpus MENELAS, la tâche de codage a obtenu un rappel et une précision moyens situés entre 44 et 63 % selon que ses résultats étaient mesurés à l'aune de l'un ou l'autre de deux codeurs humains. Des efforts considérables ont de ce fait été consacrés à la construction de "systèmes de codage" pour ces données : terminologies, thésaurus, classifications et autres nomenclatures systématiques, qui sont employés pour enregistrer les données médicales. (1987) sont au nombre de 55 (ce nombre a varié selon les époques). 64Le projet MENELAS est chronologiquement antérieur à nos autres travaux sur le lexique et la terminologie médicale. à paraître.Bouaud, J., B. Christensen (eds.), Advances in Health Telematics, 82-89. In D. Par exemple, [un segment de] l'artère coronaire est le lieu sur lequel s'exerce la dilatation, alors que Monsieur X possède une artère dont un segment va être dilaté; quant à la sténose, c'est un état pathologique (rétrécissement) subi par un segment d'artère. La Classification internationale des maladies (CIM-10) (Organisation mondiale de la Santé, 1993) en est l'exemple le plus courant. Reading Mass., Addison Wesley. Silberztein, M. On peut se poser une classe sémantique pour un terme "inconnu" de la terminologie d'origine? (1999): Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur, analyse syntaxique automatique non combinatoire. L'hypothèse de base, dans l'esprit des travaux de Harris et al. Ce type de méthode, testée sur un corpus de cent comptes rendus d'hospitalisation en hématologie, obtient un excellent rappel (elle trouve la quasi-totalité des codes attendus), mais demande la participation active du "codeur" humain pour sélectionner le bon code parmi ceux proposés (précision brute faible) (Blanquet et Zweigenbaum, 1999). L'utilisation de relations hiérarchiques (SNOMED, CIM-10) ou transversales (SNOMED) donne également de très bons résultats (Grabar et Zweigenbaum, 2000b).35Un certain nombre de termes médicaux ont été forgés d'après un nom propre. De plus, les règles obtenues incluent des contraintes syntaxiques, par exemple la règle d'ajectivation de la question est "projetée" sur la représentation du texte; si cette projection est possible, le texte contient une réponse à la question, et la représentation du texte contient une réponse est déterminée par la représentation du texte. Des comptes rendus hospitaliers dans les dossiers de patients aux livres de cours en passant par les "références médicales opposables" et les protocoles diagnostiques et thérapeutiques, le texte est partout. & P. On trouve ainsi dans les couples de termes ci-dessus les couples de mots suivants : 28Hiérarchie, référence : 29Synonymie : 30En retournant cette observation, nous formulons l'hypothèse que dans le contexte contraint de deux termes en relation sémantique l'un avec l'autre, si un mot figurant dans le premier terme et un mot figurant dans le premier terme et un mot figurant dans le premier terme et un mot figurant dans le second terme possèdent une chaîne de caractères initiale commune "suffisamment longue", ces deux mots sont en relation morphologique (ils sont construits à partir d'une même base). Nous avons constaté que 88 % des 5896 occurrences de mots trouvées dans le DELAF, et qu'elles incluaient nombre de mots spécialisés (acidose, adénopathie, amylasémie, etc.). Enfin, le dernier énoncé mentionne plusieurs facteurs de risque cardiovasculaire en les nommant spécifiquement, mais sans employer le terme lui-même. 4Pourtant, les méthodes informatiques et statistiques de traitement de l'information médicale, qu'elles servent au soin des patients, à la recherche clinique ou à la gestion des établissements de santé, reposent sur le recueil de données normalisées. Nous les présentons brièvement ici tour à tour. Notre travail sur ce thème a été mené en collaboration avec Benoît Habert et Adeline Nazarenko (Nazarenko et al., 2000), en utilisant le système d'analyse de corpus ZELLIG (Habert et Fabre, 1999). La tâche que nous abordons en est une version particulière : identifier si un mot est un nom propre ou pas, et ce à partir de termes isolés plutôt que de phrases entières. Habert, & J. Elle atteint un score conjoint de 86 % de précision et 88 % de rappel sur la CIM-10 et 98 % de précision et 97 % de rappel sur la SNOMED.44Les modélisations du domaine médical recensent un ensemble de grandes classes sémantiques dans les quelles peuvent être rangées les notions principales mises en jeu dans les textes médicaux. Le thésaurus MeSH comprend 16 classes ("chapitres") dont certaines recouvrent celles de la SNOMED: anatomie (= topographie), êtres vivants, diagnostics, produits chimiques, etc. Le problème vient du fait que la casse seule (le fait de commencer par une majuscule) n'est pas un critère absolu. In M. Chaque
critère isolé est insuffisant, mais une combinaison appropriée de ces critères est un indice fort de détermination des noms propres. In Actes de RIAO 2000 : Accès à l'Information Multimédia par le Contenu, Paris, CID, 765-784. Grabar, N. (1993) : An endogeneous corpus-based method for structural noun phrase disambiguation. A., D. Plus précisément, nous avons montré que l'on peut, en étudiant les mots présents dans une terminologie structurée, recenser des couples de mots possédant une relation morphologique (Zweigenbaum et Grabar, 2000). En revanche, il n'existe pas pour le français de base de données morphologiques sur la dérivation et la composition. 37La reconnaissance des noms propres est un thème de recherche actif dans le domaine de l'extraction d'information (Defense Advanced Research Projects Agency, 1996; Mikheev, 1999). 2Les entreprises déja anciennes de systématisation de terminologies médicales ont donné le jour à des bases de données terminologies médicales ont donné le jour à des bases de données terminologiques gigantesques, avec plus d'un million de terminologies médicales ont donné le jour à des bases de données terminologies médicales ont données terminologies médicales données terminologies don Lindberg et al., 1993). Mattick Jr, A. & J.-P. Nous appliquons cette règle à une liste de mots qui se trouvent dans la SNOMED et des mots qui se trouvent dans la SNOMED et de la SNOMED et de la SNO plus d'une analyse pour chaque phrase) de l'ordre de 70 %. 3Malgré cette emprise croissante de la technologie, le vecteur privilégié de l'information et de la connaissance médicale reste la langue. La figure 2 montre une partie de la description qui accompagne le type de concept pour angioplastie (une intervention consistant à dilater une artère). Enfin, des méthodes meilleures que la méthode de vote pourraient être employées pour la prédiction de la catégorie d'un mot inconnu. Nous avons testé cette hypothèse sur les termes de la nomenclature SNOMED en examinant la relation de synonymie. Parmi les 12 % restants, on trouve quatre classes principales : noms de médicaments (Biocidan, Clamoxyl, etc.), abréviations (NFS, PaO, temp, etc.), mots spécialisés dont la plupart sont construits par dérivation ou composition savante à partir de bases connues (hépatitique, hémodiafiltration, etc.), et noms propres (de patients, de médecins, d'hôpitaux, etc.). dépendances élémentaires établissent des contextes d'occurrence de mots dans les syntagmes nominaux du corpus (noms, adjectifs, noms propres et participes passés). Journal of the American Medical Informatics Association 6 (suppl). Bodenreider, O. Friedman, & M. Palotay, R. Par exemple, Blanquet et Zweigenbaum (1999) amorcent le repérage des expressions à coder sur les noms de tête des termes normalisés (ici, leucémie, sténose, etc.). L'invariance à la paraphrase des résultats obtenus dans les tâches visées (codage, réponse à des questions) constitue un test minimal de cette normalisés (ici, leucémie, sténose, etc.). L'invariance à la paraphrase des résultats obtenus dans les tâches visées (codage, réponse à des questions) constitue un test minimal de cette normalisés (ici, leucémie, sténose, etc.). noms propres dans ces terminologies. Nous avons par ailleurs étiqueté (catégorie syntaxique et nombre) et lemmatisé les termes de la SNOMED; nous obtenons alors des couples où seules dérivation et composition sont représentées. Chaque concept important du domaine (plus de 500) est accompagné d'une description des concepts qui lui sont liés et des relations qui assurent ces liens. 36La compilation d'un lexique des noms propres employés dans les termes médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus large de constitution de ressources lexicales pour le traitement de la langue médicaux s'inscrit dans notre projet plus la large de constitution de ressources lexicales pour le traite plus la large de constitution de ressources lexicales pour le traite plus la large de constitution de ressources lexicales plus la la large de constitution de ressource lexicales plus la large de le projet MENELAS. 54Dans le projet MENELAS (Zweigenbaum et Consortium MENELAS, 1995), nous avons cherché à mettre en œuvre une analyse de comptes rendus d'hospitalisation du sens des phrases et des textes, et travail à partir de cette représentation. En cela, les problématiques (tudiées dans cet article sont également pertinentes pour l'indexation à des fins de recherche d'information. Il peut s'agir de figures scientifiques (maladie de Parkinson), de lieux (fièvre de Lassa), de personnages historiques (syndrome de Münchhausen) ou mythologiques (tendon d'Achille). 45Les termes d'une terminologie sont catégorisés selon les classes de cette terminologie. Jacquemin, et M.-C. Des analyseurs robustes commencent à voir le jour pour le français (Vergne, 1999; Aït-Mokhtaret Chanod, 1997); il serait intéressant de textes médicaux se trouve dans Spyns (1996). Bibliographie En ligneAït-Mokhtar, S. D'une part, on parle de resténose, ce qui pour le codage cible doit ici être considéré comme une sorte de sténose. Zweigenbaum (2000b): A general method for sifting linguistic knowledge from structured terminologies. Methods of Information in Medicine 35 (4-5), 285-301. Vergne, J. Zweigenbaum (1996): Processing metonymy: a domain-model heuristic graph traversal approach. L'Homme (eds.), Recent Advances in Computational Terminology. Elle recense de l'ordre de cent mille concepts répartis dans une dizaine d'axes sémantiques orthogonaux: anatomie, atteintes morphologiques, atteintes fonctionnelles, êtres vivants, etc. 49Si l'on connaît les classes sémantiques des mots du corpus, on peut les noter sur le graphe produit. MENELAS Final Edited Progress Report. Zweigenbaum, P. Nous avons donc entrepris la construction d'une telle base pour le français médical. d') np ((, np)* et np)? Landau, A. De plus, comme souvent dans les langues techniques, la composition savante est un mode privilégié de construction de nouveaux mots. Springer-Verlag.National Library of Medicine (1986): Medical Subject Headings. Il est fréquent que les comptes rendus d'hospitalisation contiennent des phrases longues et emploient un style moins tenu, ce qui peut rendre difficile le travail d'analyseurs syntaxiques fondés sur des grammaires formelles. Bouaud (2000): Corpus-based extension of a terminological semantic lexicon. Ce projet a, lorsque c'était possible, réutilisé des lexiques et analyseurs existants; il a aussi donné lieu à la conception de nouveaux outils. Grabar (2000): A contribution of medical terminology to medical language processing resources: Learning morphological knowledge from thesauri. En effet, la fréquence relative des mots des différentes classes sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas
réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs sémantiques de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble de la SNOMED ne sont pas réellement suffisantes pour couvrir l'ensemble des champs de la sont pas réellement suffisantes pour couvrir l'ensemble de la sont pas réellement suffisantes pour couvrir l'ensemble de la sont pas réellement suffis au l'ensemble de la sont pas réellement suffis au l'ensemble de la sont partie de la sont partie de la sont partie de catégoriser les patients dans des classes disjointes, et les terminologies employées sont structurées en classifications hiérarchiques. Nous présentons ensuite nos travaux sur la constitution de ressources lexicales pour le traitement automatique de la langue médicale (section 3). Harris (1989): The Form of Information in Science, Analysis of Immunology Sublanguage. Nous abordons la problématique de l'aide au codage (section 4.1). Elle a été appelée pour 8749 relations grammaticales et a exploré 247877 chemins, soit une moyenne de 28 chemins par appel et 904 par phrase. Un sous-ensemble de la SNOMED, le Répertoire d'anatomopathologie, a été traduit en français ; la traduction de la SNOMED entière est en cours. Brochu (eds.) (1993): The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International. Laires, M. Dugourd, & L. Par ailleurs, les étapes préalables d'analyse syntaxique, en particulier l'étiquetage syntaxique, en particulier l'étiquetage syntaxique, n'étaient pas exemptes d'erreurs, qui ont été propagées dans la suite des traitements. Au vu du contenu de ce thésaurus, il semble que le besoin de précision soit moindre, mais que le domaine à couvrir soit plus large : par exemple, le MeSH inclut un axe "géographie" (pays, villes) mais pas la SNOMED. (1993) : Dictionnaires électroniques et analyse automatique de textes : le système INTEX. 19Plutôt que de construire d'emblée cette base lexicale manuellement, nous avons cherché à mettre au point une méthode permettant de mettre en évidence automatiquement un grand nombre de mots construits. Les classes sémantiques mises au point dans les travaux de Sager et al. Dans la première expression, les mots leucémie aiguë LAM 2 suffisent à identifier le terme cible correct de façon unique. En comparaison, si l'on évalue l'un des codeurs par rapport à l'autre, rappel et précision sont de 59 et 71 %. Par exemple, pour l'angioplasty et le concept Human Being :61Du fait de la richesse du réseau constitué, plusieurs chemins, voire plusieurs milliers, peuvent être trouvés entre deux concepts. Les mots dont les ensembles de contextes sont similaire ont des sens proches. Le principe est d'appliquer des transformations qui permettent de passer des termes normatifs aux variantes effectivement trouvées en corpus. 15Nous présentons ci-dessous une méthode pour recenser les dérivations et compositions à l'œuvre dans le lexique médical à partir de l'étude de terminologies structurées. Harris, & S. Ce modèle est organisé autour d'une "ontologie" du domaine : une hiérarchie de 1800 types de concepts structurée par la relation sorte-de (la figure 1 en montre un extrait), une hiérarchie de plus de 300 types de relations, et des contraintes sur les concepts plus spécifiques. 9Le second objectif est la description d'informations plus variées devant être consignées dans le dossier d'un patient. 13 Parmi les quatre classes de mots inconnus identifiées plus haut, celle des médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose pas de problème particulier : les bases de données sur les médicaments ne pose particulier : les bases de données sur les médicaments ne pose particulier : les bases de données sur les médicaments ne pose particulier : les bases de données sur les des noms de substances qui les composent. La combinaison qui donne le meilleur équilibre entre précision et rappel tout en leur assurant des valeurs élevées se définit par (("majuscule" et non "symbole" et non "micro-organisme") et ("invariant" ou "patron")). Zweigenbaum (2000) : Stratégies d'identification de noms propres à partir de nomenclatures médicales parallèles. Comme les terminologies médicales sont multiples et de grande taille, nous avons cherché à mettre au point une méthode générique et adaptable pour l'aide au recensement des noms propres employés dans une terminologie (Bodenreider et Zweigenbaum, 2000). C'est aussi le cas dans le deuxième énoncé pour les mots sténose interventriculaire antérieure. Les couples montrés sur le tableau 1 ont effectivement été repérés par cette méthode à partir de la relation de synonymie dans la nomenclature SNOMED. L'un des principaux problèmes traités par cet analyseur était celui de la composition sémantique : étant donné deux représentations partielles pour deux segments de phrase en relation syntaxique, comment calculer la représentation pour le segment complet.58Le principe que nous avons adopté pour cette représentation. Lyman (eds.) (1987): Medical Information Processing. Selon le réglage des différents paramètres de la méthode, la proportion de mots bien classés varie de 50 à 70 %. van Bemmel (1997): Handbook of Medical Informatics. Cette représentation a également été utilisée pour répondre à des questions sur le contenu des textes. Leur forme va du nom simple, comme dans la plupart des exemples précédents, aux noms composés (sonde de Swan-Ganz, amyotrophie de Charcot-Marie-Tooth), coordonnés (syndrome de Fitz-Hugh et Curtis, maladie de Pierre Marie et Sainton), précédés de prénoms (signe pupillaire de Marcus Gunn), ou de particules (neurofibromatose de von Recklinghausen, syndrome de Van der Hoeve, thyroïdite de de Quervain, point de McBurney). Ce chemin est une relation ou une séquence de relations et de concepts intermédiaires. Nous examinons également la question des mots dans un corpus. In ANLP97, Washington, DC, 72-79. Blanquet, A. 62Une fois un texte analysé, les informations recherchées sont repérées directement dans sa représentation conceptuelle. Par ailleurs, le recensement a priori des terminologies pour un domaine donné, voire à compléter ou réorganiser les terminologies existantes. Tsujii (ed.), Proceedings of the 16th COLING, Copenhagen, 137-142. Bourigault, D. La question qui se pose est celle de la taille qui serait nécessaire pour qu'une proportion suffisante des termes présents dans une terminologie donnée soit représentée dans un corpus. Paris, Masson. Sowa, J. al/A.34 Nous avons pratiqué des expériences similaires sur l'anglais (SNOMED, CIM-10) et le russe (SNOMED) avec un taux de précision similaire (Grabar et Zweigenbaum, 2000a). Nous commençons par donner un aperçu des grandes terminologies médicales et de leurs fonctions (section 2). 42Nous avons dans ce travail employé les versions française, anglaise et allemande de la CIM-10, et les versions française et anglaise du Répertoire d'anatomopathologie de la SNOMED. Les termes cible sélectionnés sont ensuite ceux qui partagent le maximum de mots avec l'expression source, en tenant compte de proximités morphologiques (carotide-carotidienne). Cette méthode de composition sémantique par recherche de chemins dans un graphe a été testée sur 274 phrases tirées du corpus MENELAS. London, Addison-Wesley.
En ligneSpyns, P. D'autre part, la première proposition contient une négation, et l'information correspondante (sténose de l'artère interventriculaire antérieure (IVA)) ne doit pas être codée. Les notions désignées sont variées : maladies (maladie de Parkinson, aphasie de Wernicke, aphasie de Broca), signes et symptômes (réflexe de Bainbridge, signe de Romberg), examens (test de Coombs, coloration de Gram), appareils médicaux (valve de Starr-Edwards, broche de Kirschner), actes diagnostiques ou thérapeutiques (opération de Billroth, manœuvre de Heimlich, indice d'Appar), micro-organismes (bacille de Hansen, virus d'Epstein-Barr) ou localisations anatomiques (tubercule de Lisfranc, trompe de Fallope, ganglion de Gasser, cul de sac de Douglas, bronche de Nelson, polygone artériel de Willis). Les premiers exemples peuvent être traités par des méthodes lexicales. 55MENELAS s'est déroulé sous la forme d'un projet européen réunissant de 1992 à la mi-1995 douze partenaires et abordant l'analyse de comptes rendus en français, anglais et néerlandais. H. In Proceedings of the 6 EACL, Utrecht, 81-86. Burnage, G. Nijmegen, Centre for Lexical Information, University of Nijmegen, Centre for Lexical Information, University of Nijmegen, Centre for Lexical Information, University of Medicine, 1986) qui est utilisé à la National Library of Medicine américaine pour indexer la base bibliographique Medline. (1990): CELEX. 14Certaines abréviations courantes figurent dans les terminologies ou les dictionnaires imprimés. Étant donné une terminologie, des travaux comme ceux de Jacquemin (1997) visent à reconnaître en corpus des variantes de ses termes. 40par exemple, maladie de Parkinson.

