INSTITUT DE LANGUES ET CIVILISATIONS ORIENTALES

MASTER 2 « INGENIERIE MULTILINGUE »

ANNEE 2013/2014

La détection des prédicats complexes hindi dans le cadre d'un outil d'aide à la lecture

Mémoire de recherche

présenté par
Satenik Mkhitaryan

dirigé par
François Stuck et Mathieu Valette

Date de soutenance **: septembre 2014**

Remerciements

Je remercie mes deux directeurs de mémoire et de stage, François Stuck et Mathieu Valette, de m'avoir donné l'opportunité de travailler au sein de l'ER-TIM sur ce projet intéressant et m'avoir guidée tout au long de ce travail.

Je tiens à remercier Annie Montaut pour son soutien infaillible durant mon parcours à l'Inalco. Je lui en suis vraiment reconnaissante. Je remercie également tous mes professeurs de hindi pour les connaissances solides transmises.

Merci à André Salem pour ses conseils toujours judicieux, à Jean-Michel Daube pour sa bonne humeur, à tous les professeurs impliqués dans le master « Ingénierie linguistique » pour leur enseignement riche et varié.

Merci aux autres étudiants du master pour l'ambiance de travail agréable et amicale, l'entraide et le véritable esprit d'équipe.

Enfin, je remercie le yoga qui m'a permis de rester zen dans toutes les situations.

1 Table des matières

2 Introduction

Les recherches menées dans ce travail portent sur deux domaines : le traitement automatique du hindi et l'Apprentissage des Langues Assisté par Ordinateur. Dans ce mémoire nous nous intéresserons aux prédicats complexes hindi en vue de leur détection automatique dans le cadre d'un dispositif d'aide à la lecture en cours de développement – DEJALU.

Un prédicat complexe est une expression composée d'un nom, adjectif, adverbe ou verbe et d'un verbe, mais qui se comporte comme un verbe simple. On peut comparer cette structure aux verbes à particule et aux constructions à verbe support de l'anglais. Ce type de construction est souvent qualifié de « pain in the neck for NLP » (Sag et al., 2002). Les prédicats complexes sont aussi réputés dans le milieu de l'enseignement des langues pour la difficulté qu'ont les apprenants à les appréhender. En effet, les apprenants ont tendance à interpréter chacun des composants du prédicat complexe (ce que fait Google Traduction pour certains prédicats complexes, cf. Annexe 1) ce qui les induit en erreur. Nous pensons qu'il est utile de détecter l'ensemble du prédicat complexe d'une part pour l'annoter et forcer le lecteur à considérer le prédicat complexe dans son intégralité et non chacune de ses composantes de façon séparée, et d'autre part pour pouvoir les exploiter pour développer d'autres fonctionnalités de DEJALU. Certaines études (Yasuda, 2010) ont d'ailleurs montré que l'annotation et la mise en valeur des verbes à particule de l'anglais dans un texte facilitaient l'apprentissage et permettaient par rapport à un texte non-annoté d'améliorer significativement les résultats des apprenants.

Dans ce mémoire, nous commencerons par présenter les caractéristiques générales de la langue hindi, avant de détailler son système verbal. L'état de l'art développé portera sur le traitement automatique du hindi en général et sur le volet plus spécifique de notre recherche à savoir la détection automatique des prédicats complexes hindi.

Dans la seconde partie, une description de DEJALU sera faite, précédée d'une brève présentation de quelques outils d'aide à la lecture afin de le situer parmi d'autres outils similaires. Dans un deuxième temps, en prenant en compte les éléments linguistiques ayant fait l'objet d'une présentation approfondie dans la première partie, nous tenterons de proposer une méthode de détection des prédicats complexes dans DEJALU.

Nous décrirons enfin les tests d'évaluation qui ont pu être réalisés, nous essaierons de tirer les conclusions de ces résultats et nous présenterons de possibles pistes d'amélioration.

3 La langue hindi

Avant de parler du traitement automatique du hindi, il apparaît essentiel, pour la compréhension de ce travail, de présenter les caractéristiques de la langue hindi, et notamment celles du système verbal.

3.1 Caractéristiques générales

Système d'écriture

Le hindi utilise le système d'écriture devanāgarī (देवनागरी), un système alphasyllabaire (les consonnes sont dotées d'une voyelle par défaut, la voyelle a (bref) en l'occurrence) qui s'écrit de gauche à droite, ne fait pas de distinction entre majuscules et minuscules. Les caractères sont attachés sous une barre horizontale (potence). L'alphabet est représenté sous forme d'un tableau phonologique où les sons sont classés en séries : les voyelles d'abord, comportant l'opposition brève/longue, et les consonnes ensuite, selon le lieu d'articulation (des vélaires aux labiales) et leur mode d'articulation (sourde, sourde aspirée, sonore, sonore aspirée, nasale**) :** क ka ख kha ग ga घ gha ङ ṅa. La série des sons empruntés au persan (fa, za, ġa, k͟ha) et à l'arabe (qa) est systématiquement représentée avec un point souscrit sous la lettre correspondant au son perçu comme le plus proche (क़, ख़, ग़, ज़, फ़). La devanagari utilise les signes de ponctuation de l'alphabet latin, sauf la fin de phrase qui est marquée par une barre verticale « । »[1]. Quant aux chiffres, elle dispose de ses propres symboles pour les représenter (० १ २ ३ ४ ५ ६ ७ ८ ९) mais les chiffres latins sont de plus en plus souvent employés.

Encodage des caractères

La devanagari est représentée dans le bloc Devanagari de l'Unicode (U+0900 à U+097F)[2]. Les ligatures[3] n'y sont pas. Pour réaliser une ligature, on a recours au signe ् appelé virāma ou halant, qui doit être inséré entre deux caractères que l'on veut ligaturer : ब (ba) + ् (virāma) + य (ya) = ब्य (bya)

Translittération

Il existe plusieurs systèmes de translittération. Celui qui sera utilisé dans ce travail est

[1] U+0964 DEVNAGARI DANDA à ne pas confondre avec la barre verticale « | » située dans le bloc BASIC_LATIN (U+007C).
[2] http://www.unicode.org/charts/PDF/U0900.pdf
[3] Une ligature est la fusion de deux graphèmes d'une écriture pour n'en former qu'un seul nouveau, considéré ou non comme un caractère à part entière. Exemple : o+e → œ (Wikipedia)

l'alphabet international pour la translittération du sanskrit IAST (International Alphabet of Sanskrit Transliteration, cf. Annexe 2) qui est de facto le standard académique.

Eléments de morphosyntaxe

Le hindi possède deux genres et deux nombres. L'adjectif varie moins que le nom, seulement ceux en –a se fléchissent, les autres adjectifs sont invariables.

Formellement on distingue deux cas en hindi, le cas direct et le cas oblique. Ce dernier est utilisé devant les postpositions qui elles indiquent la fonction du groupe nominal : ne ergatif[4], ko datif, se ablatif/instrumental, kā, kī, ke génitif etc.

L'ordre des mots canonique est SOV. Il est relativement contraint et les composants de la phrase ont plus ou moins une position fixe mais elle peut changer selon le contexte ou des besoins stylistiques. Le sujet, s'il n'est pas omis, est en première position et le verbe est final, donc les compléments sont antéposés au verbe. Les adjectifs précèdent les noms. Les prépositions, appelées postpositions, sont postposées. L'hindi est donc une langue centripète.

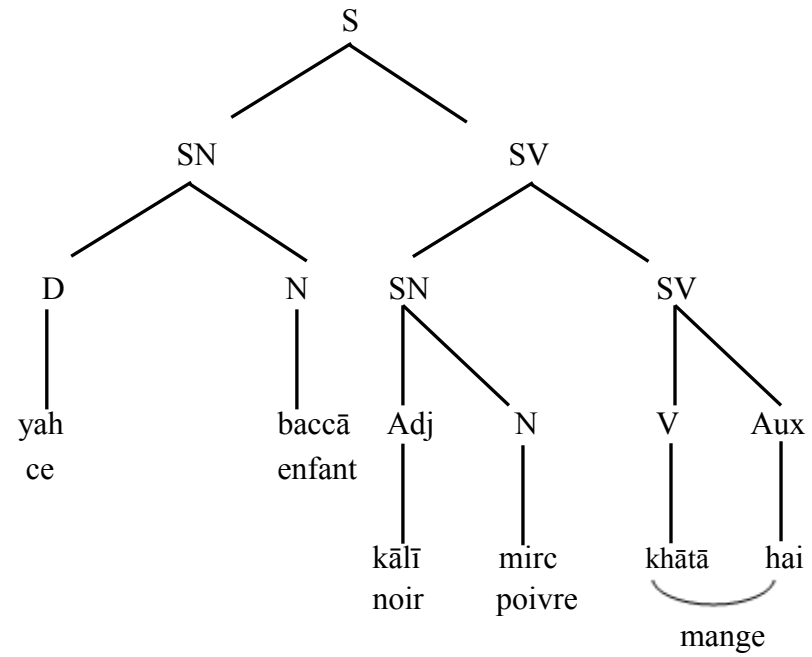
Ci-dessous la représentation d'une phrase par un arbre syntaxique qui illustre l'ordre des mots en hindi :

यह वच्चा काली मिर्च खाता है

Yah baccā kālī mirc khātā hai

DEM enfant noir.F poivre manger PRES.3S

Cet enfant mange du poivre noir.



[4]L'ergativité en hindi est déterminée par l'aspect accompli du verbe transitif qui s'accorde avec l'objet direct et non pas avec le sujet. Le patient direct est alors au cas oblique et il est suivi de la postposition ने ne.

4 Eléments de grammaire hindi pour la détection et la catégorisation des PC.

Après ce bref aperçu de la langue hindi, nous allons à présent détailler les connaissances linguistiques sur le sujet plus précis qui nous intéresse : les verbes en hindi. Ce chapitre est inspiré des travaux de A.Montaut [6] [7] [8].

4.1 La morphologie et le lexique verbal

Une grande partie des bases verbales se présente par couples transitif/intransitif auxquels s'ajoute le causatif ou le second causatif lequel augmente d'un argument la valence verbale.
Le transitif ou le causatif sont dérivés à partir de l'intransitif. Cette opération morphologique recourt à :

1. des modifications phonétiques dans la base verbale
   रुकना ruknā « s'arrêter » रोकना roknā « arrêter »

1. une suffixation de –ā ou –lā à la base verbale avec ou sans modification phonologique dans la base
   उठना uṭhnā « se lever » उठाना uṭhānā « lever »
   सोना sonā « dormir » सुलाना sulānā « endormir »

2. une substitution du verbe support intransitif par transitif dans les locutions verbales (होना honā « être », करना karnā « faire »)
   बंद होना band honā « être fermé » बंद करना band karnā « fermer »

Aux verbes corrélés transitifs/intransitifs s'ajoute un troisième verbe corrélé, le causatif ou le double causatif, formé avec le suffixe –vā.

बनना	bannā	être fait	intransitif
बनाना	banānā	préparer/fabriquer	transitif
बनवाना	ban**vā**nā	faire préparer/fabriquer	causatif

Ainsi, les verbes ont trois bases : intransitif, transitif, causatif ou bien transitif, causatif, double causatif, mais l'on trouve aussi des verbes qui en ont une seule, deux ou quatre. Il arrive aussi que certaines bases gardent la même forme qu'elles soient intransitives ou transitives (बदलना badalnā « changer »).

4.2 Temps, aspect, mode

4.2.1 Les formes finies

L'infinitif

L'infinitif est un nom verbal (toujours avec la désinence –nā au cas direct) qui peut remplir de nombreuses fonctions différentes en lui ajoutant diverses postpositions auquel cas le verbe se mettra à l'oblique (-e) comme les noms masculins se terminant par –ā.

खाने से पहले हाथ धो

khāne se pahale hāth dho

Lave-toi les mains avant de manger

Les participes

Les deux participes du hindi se distinguent par l'aspect, perfectif (ou accompli) et imperfectif (ou inaccompli), et correspondent à peu près aux participes présent et passé du français. La flexion des participes est celle des adjectifs variables en –ā (-ā, -e, -i). L'aspect imperfectif (base+t+désinences adjectivales) correspond à une action en train de se dérouler. L'aspect perfectif (base+désinences adjectivales) correspond au résultat d'une action achevée. L'ajout des désinences adjectivales entraîne une transformation phonologique quand la base verbale se termine par une voyelle : le glide –y– s'intercale entre la base et la désinence –ā du masculin singulier, et les désinences –i et –e dans les usages non standard. Le participe de la copule (हुआ huā, हुए hue, हुई huī) peut s'adjoindre et renforcer les participes. Les participes peuvent avoir soit des fonctions adjectivales, soit adverbiales (le plus souvent à la forme oblique en –e invariable).

Emploi adjectival (le participe se place avant le nom)

बोलती (हुई) लड़की मेरी बहन है

Boltī (huī) laḍki merī bahan hai

La fille qui parle est ma sœur.

Emploi adverbial (le participe se place entre le nom et le verbe)

मेरी बहन बोलते (हुए) आ रही है

Merī bahan bolte (hue) ā rahī hai

Ma sœur vient en parlant

L'absolutif

L'absolutif (ou gérondif) se forme en suffixant –kar (parfois –ke) à la base et il permet de

coordonner un verbe secondaire au verbe principal. Si la forme est redoublée, on a affaire à une simultanéité. Si les deux procès sont représentés comme indissociables, le morphème *kar* est effacé : le jānā, le ānā au lieu de lekar jānā, lekar ānā.

Le syntagme verbal hindi a cinq paradigmes simples : impératif, subjonctif, irréel, aoriste, futur**.**

L'impératif a 3 formes correspondant aux 3 pronoms de deuxième personne et une forme de politesse. Le subjonctif est utilisé pour la première personne du pluriel. L'infinitif peut substituer à l'impératif. Ex : bol, bolo, bolie, boliegā « parle », « parlez ».
Le subjonctif est formé par l'affixation à la base des désinences personnelles (ūṃ, e, eṃ) et ne comporte ni marques d'aspect, ni marques de genre. Exemple : bolūṃ, boleṃ « que je parle », « que nous parlions ».
Le conditionnel ou irréel, quasi identique au participe inaccompli (la seule différence est la nasalisation au féminin pluriel), ne possède ni marques personnelles ni marques temporelles.
L'aoriste ne diffère du participe accompli que par la nasalisation au féminin. Il ne comporte ni marque de temps ni de personne ni d'aspect. Comme tous les temps de l'accompli en hindi, l'aoriste adopte la structure ergative.
Le futur est formé sur la base du subjonctif à laquelle s'affixe le morphème *-gā*. C'est le seul temps simple qui porte à la fois une marque de temps de personne et de genre/nombre.

4.2.2 Les formes composées

Deux morphèmes temporels représentent les temps : *hai*, « est », pour le présent, porte des désinences personnelles (ūṃ,o,ai,aiṃ) et *thā*, « était », pour le passé, porte seulement des désinences de genre et de nombre (thā, the, thī, thīṃ).
Ces morphèmes s'affixent à l'un ou l'autre des participes pour produire les quatre temps composés de l'indicatif : présent général, imparfait général, perfectif composé 1, perfectif composé 2.

On note deux particularités de l'emploi du présent : l'auxiliaire être est le plus souvent omis à la forme négative et, en cas d'omission de l'auxiliaire, le verbe est nasalisé.

लड़कियां गोश्त नहीं खातीं
Laḍakiyāṃ gośt nahīṃ khatīṃ
Les filles ne mangent pas de viande.

Comme dans beaucoup de langues, les deux présents, ainsi que le perfectif simple, ont parfois

une référence temporelle future.

अभी आता हूं / आ रहा हूं / आया
Abhi ātā hūṃ / ā rahā hūṃ / āyā
J’arrive tout de suite.

Pour produire les subjonctif et irréel composés 1 ou 2, dont la référence temporelle est dans les deux cas passée, les participes 1 ou 2 sont utilisés respectivement auxquels s'ajoutent le subjonctif ou l'irréel de « être ».

4.2.3 Les marqueurs d'aspect

Le plus fréquent des affixes aspectuels est l'actualisateur *rahā,* variable en genre et en nombre, suffixé à la base et auquel se suffixe ensuite le marqueur temporel hai ou thā pour produire le présent ou l'imparfait actualisé.

वह बोल रहा है
Vah bol rahā hai
Il est en train de parler

वह बोल रहा था
Vah bol rahā thā
Il était en train de parler

4.2.4 Les marqueurs d'aspect secondaire

Un certain nombre de verbes fournissent les marqueurs d’aspects secondaires.
Le duratif qui exprime un processus continu est formé du participe inaccompli suivi de l’auxiliaire रहना rahnā « rester » conjugué aux divers temps de l’indicatif, du subjonctif, de l’impératif.
Le duratif progressif représente l’action comme à la fois constante et augmentant d’intensité. Il est formé du participe inaccompli et l’un des auxiliaires fournissant cet aspect : चलना calnā « marcher », जाना jānā « aller », चला जाना calā jānā « partir », आना ānā « venir », चला आना calā ānā « venir ». Le fréquentatif exprime la répétition ou l’habitude, il est formé du participe accompli à la forme du masculin singulier, invariable, suivi de l’auxiliaire करना

karnā « faire » conjugué. L'entrée dans le procès est représentée par l'auxiliaire लगना lagnā « se coller » qui suit l'infinitif fléchi. L'inceptif est incompatible avec l'ergatif, le progressif et la négation. La sortie du procès se forme par la combinaison de la base verbale nue et de l'auxiliaire चुकना cuknā « être acquitté ». Comme dans le cas de l'inceptif, cette forme est incompatible avec l'ergatif, le progressif et la négation.

4.2.5 Les modalisateurs

La modalité du **pouvoir** se forme par deux auxiliaires de potentiel : सकना saknā et पाना pānā, construits sur la base verbale et rarement utilisé avec l'actualisateur (seul pānā est compatible avec le progressif), saknā exprime la possibilité générale (capacité, autorisation) alors que pānā exprime l'incapacité à vaincre un obstacle interne ou externe et est souvent utilisé dans les contextes négatifs.

La modalité du **devoir** se construit avec trois auxiliaires affixés à l'infinitif qui se distinguent par le sens : चाहिए cāhie « veuillez », invariable, est le moins marqué et exprime une obligation générique de l'ordre du conseil ; होना honā « être » une obligation ponctuelle ; पड़ना paḍnā « tomber » une contrainte forte. Tous les trois présentent une construction spéciale : le sujet est au datif (cas oblique postposé par को ko) et le syntagme verbal s'accorde avec l'objet si le verbe est transitif et au masculin singulier si le verbe est intransitif.
Le verbe देना denā « donner » fournit la modalité **permissive**, il est postposé à l'infinitif au cas oblique.
La modalité **acquisitive** est une forme rare et a la même construction que le permissif (pānā au lieu de denā) et peut être considéré comme la contrepartie passive du permissif.
La probabilité (le **présomptif**) est rendue à l'aide du morphème hogā, futur de la copule, qui se substitue au morphème temporel dans les formes conjuguées ou s'ajoute à la forme aoriste dans le cas de l'aoriste. La forme est compatible avec divers aspects, temps et modes.

4.2.6 Le passif

Pour former la voix passive on ajoute l'auxiliaire जाना jānā « aller » au participe accompli.

Tous les deux s'accordent en genre et en nombre avec le sujet de la phrase passive, donc l'objet de la phrase active. L'agent, quand il est exprimé, est suivi de la postposition ke dvārā.

पुलिस के द्वारा चोर पकड़े गए
Pulis ke dvārā cor pakaḍe gae
Les voleurs ont été arrêtés par la police

Le sujet du passif peut garder la postposition ko, s'il l'avait dans la phrase active. Dans ce cas le verbe ne s'accorde pas et prend la forme neutralisée du masculin singulier –a.

चोरों को पकड़ा गया
Coroṃ ko pakaḍā gayā
Les voleurs ont été arrêtés

4.2.7 Explicateurs verbaux

Une liste de verbes, qui varie de 12 à 35 verbes selon les auteurs, est utilisée en hindi pour marquer à la fois le perfectif (achèvement de l'action jusqu'à son terme) et l'attitude. La base du verbe principal est suivie d'un deuxième verbe conjugué qui lui ajoute diverses valeurs soit aspectuelles, soit directionnelles, soit subjectives.

Les explicateurs verbaux ne sont pas compatibles avec la négation, l'interrogation, l'inceptif, les aspects duratif et terminatif, l'absolutif, les deux auxiliaires de possibilité. Compte tenu ces quelques contraintes, on a attribué aux explicateurs une valeur perfectivante. Les explicateurs verbaux sont appelés également semi-auxiliaires perfectivants ou prédicats complexes aspectuels (Butt, 1995). Cette structure est parfois comparée aux verbes à particules de l'anglais (clean up). Les verbes les plus courants sont जाना jānā « aller », आना ānā « venir », देना denā « donner », देना lenā « prendre », उठना uṭhnā « se lever », पड़ना paḍnā « tomber », डालना ḍālnā « jeter », मारना mārnā « frapper »[5].

Exemple :
बच्चा सो गया है
baccā so gayā hai
enfant dormir aller PRES.3MS
L'enfant s'est endormi

[5] jānā « aller » transforme un état en processus inchoatif, denā « donner » et lenā « prendre » orientent l'action vers le sujet ou loin du sujet, uṭhnā « se lever » et paḍnā « tomber » exprime la soudaineté ou l'impulsivité de l'action, ḍālnā « jeter » et mārnā « frapper » donne une valeur brusque et violente.

4.2.8 Les prédicats complexes

Une grande partie des prédicats hindi est formée à l'aide d'un élément non verbal et d'un verbe et l'ensemble se comporte alors comme un verbe simple. Les nombreuses locutions qui suivent ce schéma constituent une catégorie très productive en hindi.

On peut classer les prédicats complexes selon la nature du premier formant (nom : pyār karnā « aimer » ; adjectif : khālī karnā « vider » ; adverbe : mālum honā « savoir »), les plus nombreux étant les expressions verbo-nominales, ou selon la nature du formant verbal (transitif ou intransitif, le plus souvent karnā « faire » et honā « être » respectivement). La substitution de karnā par honā correspond à un contraste actif/passif mais parfois, quand il s'agit de sentiments ou d'états, il s'agit d'un contraste entre processus involontaire et volontaire ou consciemment assumé.
Dans ce cas le sujet devient un expérient au cas oblique suivi de la postposition ko

मुझको परेशानी हो रही है
Mujhко pareśānī ho rahī hai
1S.DAT souci être PROG PRES.3FS
Je m'inquiète

Les prédicats complexes sont souvent en paires en concurrence selon l'origine du formant non verbal, sanskrite ou arabo-persane : pratīkśā / intazār karnā « attendre », āśā/ummīd karnā « espérer ». A cela s'ajoutent les prédicats complexes formés à partir d'emprunts à l'anglais, auquel cas le verbe est systématiquement recatégorisé comme formant non verbal avant de s'adjoindre au formant verbal hindi : यूज़ करना yūz karnā « utiliser », चैक करना caik karnā « vérifier ».

Le formant non verbal et le verbe sont relativement solidaires. Formellement ils ne sont dissociables que par quelques particules : तो to (opposition), ही hī (emphatique), भी bhī (aussi), नहीं nahīṃ, न na, मत mat (négation), mais parfois ils sont considérablement éloignés l'un de l'autre.

On peut faire la distinction entre deux types de prédicats complexes : coalescent et non coalescent.

Dans le premier type, le formant non verbal est relativement solidaire du formant verbal laissant une place vacante pour le patient externe qui peut être spécifiquement marqué.

Au passif et à l'ergatif, c'est le patient externe qui contrôle l'accord, comme dans les verbes simples.

मैं ने दरवाज़ा बंद किया
Maiṃ ne darvāzā band kiyā
1S ERG porte fermé faire.AOR.MS
J'ai fermé la porte

Dans le type non coalescent, le patient externe ne peut pas être spécifiquement marqué et il est construit comme génitif du formant nominal du prédicat. C'est avec le formant nominal que le verbe s'accorde à l'ergatif et au passif. Un petit nombre de locutions peut entrer dans les deux processus de construction (kī yād ānā/honā/karnā)

मैं राम का इंतज़ार कर रहा हूं
Maiṃ Rām kā intazār kar rahā hūṃ
1S Ram GEN attente faire PROG PRES.1MS
J'attends Ram

मैं ने राम का इंतज़ार किया
Maiṃ ne Rām kā intazār kiyā
1S ERG Ram GEN attente faire.AOR.MS
J'ai attendu Ram

Pour un sous type de non coalescents, l'accord au passif et à l'ergatif se fait avec le formant nominal, mais un participant externe peut occuper sa place d'argument, suivi des postpositions (se, par, ko), comme avec un verbe simple : bāt karnā « parler » construit son argument externe comme les verbes simples bolnā « parler » ou kahnā « dire ».

मैं ने उससे बात की
Maiṃ ne usse bāt kī
1S ERG 3S.SOC parole faire.AOR
Je lui ai parlé

La construction des locutions verbales (avec ou sans postposition) doit s'apprendre comme un fait de vocabulaire, car il n'y a pas de règle. On verra par la suite que cette catégorie de verbes ainsi que les explicateurs verbaux sont sources de difficultés dans le traitement automatique.

Après cette description relativement détaillée du fonctionnement des verbes en hindi, nous allons nous pencher sur les travaux qui ont été réalisés pour le traitement automatique du hindi en mettant l'accent sur le sujet de ce travail, l'identification automatique des prédicats complexes en hindi.

5 Etat de l'art

5.1 Sur le traitement automatique du hindi en général

Le traitement automatique du hindi et des langues indiennes en général est en plein développement dans plusieurs institutions partout en Inde. Les centres de recherche qui semblent les plus actifs dans ce domaine sont Indian Institut of Technology Bombay[6] (P.Bhattacharya, D.Chakrabarti,V.M.Sharma), International Institute of Information Technology Hyderabad[7] (Rajeev Sangal, Dipti Misra Sharma), Indian Language Technology Proliferation and Deployment center[8], ainsi que International Institute of Information Technology Kharagpur et International Institute of Information Technology Kanpur.

5.1.1 Traduction automatique

L'Inde étant un pays où une multitude de langues coexistent, plusieurs institutions se sont vite intéressées à la traduction automatique aussi bien entre les langues indiennes que de l'anglais vers des langues indiennes. Les approches sont différentes – basées sur des règles, des corpus parallèles, des lexiques, des méthodes statistiques etc.

Trois projets (AnglaMT, Anuvadaksh, Sampark) ont été réalisés grâce à la collaboration d'une quinzaine d'institutions de différentes régions de l'Inde (Hyderabad, Pune, Bombay, Chennai, Kharagpur, Allahabad, Tamil, Bangalore, Jadavpur etc.), financés par le gouvernement indien dans le cadre du projet TDIL (Technology Development in Indian Languages).

AnglaMT et Anuvadaksh[9] sont deux systèmes différents pour traduire de l'anglais vers des langues indiennes. Le premier est un système basé sur la grammaire hors contexte qui génère une pseudo langue intermédiaire pour ensuite traduire vers des langues indiennes (aussi bien indo-aryennes que dravidiennes) pour lesquelles il existe des modules de génération de texte. Quant à Anuvadaksh, c'est un système hybride qui combine plusieurs méthodes de traduction automatique : grammaire d'arbres adjoints, statistiques, règles d'analyse et de génération, corpus parallèles. Le système intègre également des modules de reconnaissance d'entités nommés, de désambigüisation lexicale et d'évaluation.

[6] IIT Bombay http://www.cfilt.iitb.ac.in/
[7] IIIT Hyderabad http://web2py.iiit.ac.in/research_centres/default/view_area/3
[8] ILTPDC http://tdil-dc.in/index.php?option=com_vertical&parentid=2
[9] http://tdil-dc.in/components/com_mtsystem/CommonUI/homeMT.php

Sampark[10] (« contact ») est un système de traduction automatique comprenant 6 paires de langues indiennes (Punjabi - Hindi, Hindi - Punjabi, Telugu - Tamil, Urdu - Hindi, Hindi - Urdu, Hindi - Telugu). D'autres paires sont en cours de développement. Sampark est basé sur une approche de type analyse-transfert-génération où chaque phase contient de nombreux modules. La langue source est analysée en utilisant la grammaire paninienne computationnelle (Computational Paninian Grammar), ensuite un transfert de vocabulaire et de structure vers la langue cible est effectué, et en dernier lieu la langue cible est générée. Au vue de la complexité du système et de l'hétérogénéité des modules, les auteurs ont utilisé un format très lisible de stockage d'analyse linguistique appelé Shakti Standard Format[11].
Nous pouvons citer aussi Anusaaraka[12], un projet commencé à Kanpur en 1995 par Rajeev Sangal et qui continue à Hyderabad, financé par le TDIL également. Anusaaraka traduit de cinq langues indiennes (Telugu, Kannada, Bengali, Punjabi, Marathi) vers le hindi. Ce système aussi utilise les principes de la grammaire de Panini et exploite les similarités des langues indiennes.

D'autres systèmes plus spécialisés existent : Mantra (MAchiNe assisted TRAnslation) pour traduire des textes administratifs, lettres, mémorandums, notifications, ordres etc. ; MaTra, développé à Bombay, pour traduire des nouvelles, avec un module qui détermine d'abord le type du texte entré pour utiliser le dictionnaire approprié.

Beaucoup d'autres systèmes existent dont nous citerons les noms sans entrer dans les détails[13] : AnglaBharti, Anubharti, Anuvaadak, Hinglish, Anubaad, Shakti, Shiva etc.

5.1.2 Etiqueteurs morphosyntaxiques

Plusieurs étiqueteurs ont été créés adoptant différentes approches soit basées sur des règles soit sur des méthodes statistiques (Modèle de Markov caché, arbres de décisions, CRF, entropie maximale etc.) combinées avec des règles. Mais seulement deux étiqueteurs sont disponibles en téléchargement libre.
Le premier, développé par les chercheurs de IIT Bombay, est basé sur CRF++. Il renvoie comme résultat la catégorie syntaxique uniquement, au format SSF. Les auteurs utilisent les

[10] http://sampark.org.in/sampark/web/index.php/content
[11] http://ltrc.iiit.ac.in/mtpil2012/Data/ssf-guide.pdf
[12] http://anusaaraka.iiit.ac.in/node/1018
[13] Pour plus de détails se référer à [9] et [11]

étiquettes définies par A. Bharati[14] (la liste des étiquettes est présentée en Annexe 3).

Le tableau ci-dessous présente quelques exemples de mots étiquetés :

Correct			Incorrect		
1	दिल्ली (Delhi)	NNP	18	मेट्रो (métro)	NNP
2	की (de)	PSP	23	किया \ (faire.P2)	NNP
3	सफ़लता (réussite)	NN	34	हुई। (être.P2)	NNP
4	#	SYM	39	चौबीस (vingt-quatre)	NNP
105	हैं। (sont, aux.)	VAUX	186	है। (est, aux.)	JJ
107	चार (quatre)	QC			

Un autre étiqueteur en python, basé sur l'étiqueteur TnT[15], a été développé par Siva Reddy de l'université d'Edinburgh. Pour les catégories, les auteurs utilisent les mêmes étiquettes que l'étiqueteur précédent, donc celles de A. Bharati.

Cet étiqueteur renvoie non seulement la catégorie morphosyntaxique mais aussi le lemme, le suffixe, le genre, le nombre, la personne, le cas.

Voici un exemple de sortie (l'exemple en gris est incorrect) :

Unité	Lemme	Cat	Suffixe	Cat	Genre	Nombre	Personne	Cas
चलते calte	चल cal	VM	ता	v	m	pl	any	
रिपोर्टर ripārtar	रिपोर्टर ripārṭar	NN	0	n	m	sg	3	d
कहानियों kahāniyoṃ	कहानी kahānī	NN	0	n	f	pl	3	o
नज़ारा nazārā	नज़ारा nazārā	VM		n	m	sg		d

Les détails sur l'étiqueteur sont présentés dans (Reddy et Sharoff 2011, modèle 5)[16].

Les auteurs ne donnent aucune information sur la précision des étiqueteurs. En l'absence de corpus annoté, nous n'avons pas pu les évaluer.

5.1.3 Hindi Wordnet et les prédicats complexes

[14] http://ltrc.iiit.ac.in/tr031/posguidelines.pdf
[15] http://www.coli.uni-saarland.de/~thorsten/publications/Brants-ANLP00.pdf
[16] http://sivareddy.in/papers/clia2011IndianCrossLang.pdf

Le Wordnet hindi a été créé sur la base du Wordnet anglais par les chercheurs de IIT Bombay. Il contient des noms, adjectifs, adverbes et verbes. Pour une entrée (en bleu), il fournit un ensemble de synonymes (en noir souligné), une définition (en vert), un exemple en contexte (en noir entre guillemets) et la position dans l'ontologie (liens en bleu). Le Wordnet hindi permet aussi d'établir des liens avec les Wordnets des autres langues indiennes disponibles (initiales des langues en bleu entre parenthèses).

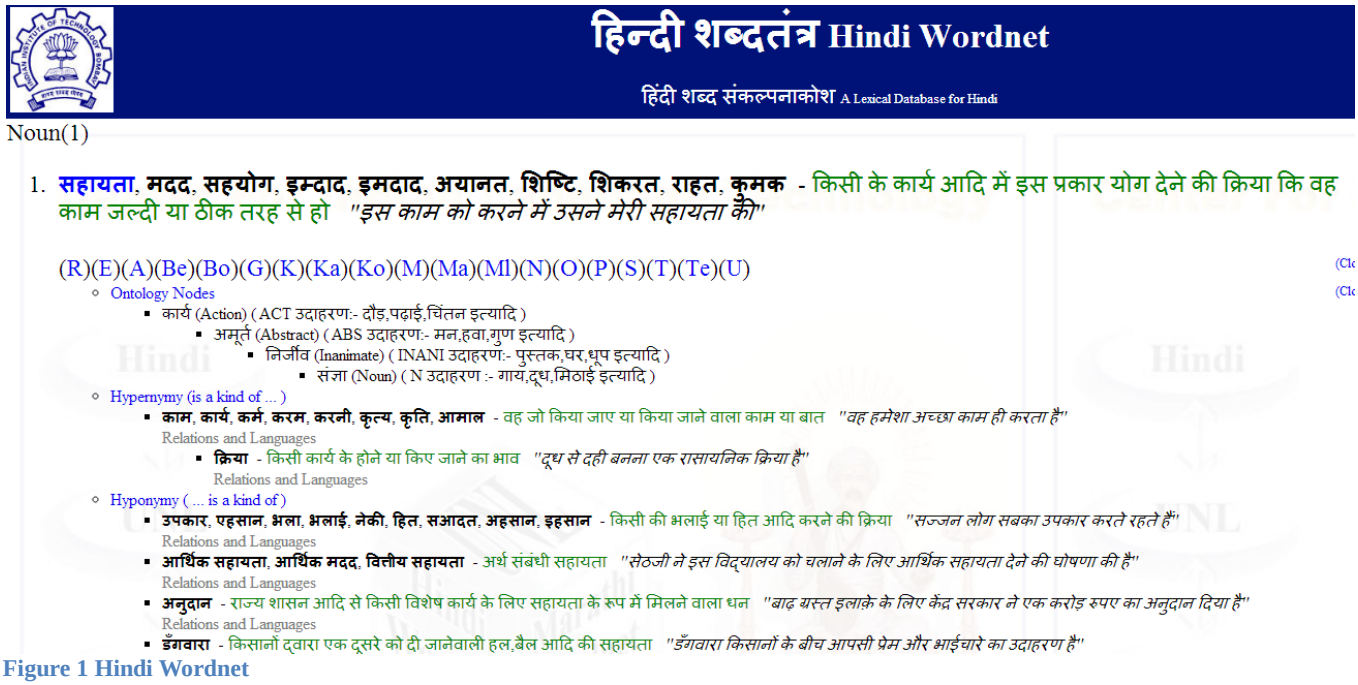


Figure 1 Hindi Wordnet

(Bhattacharya et al., 2006) s'intéressent à la façon de stocker les prédicats complexes dans le Wordnet hindi car Princeton Wordnet n'est pas adapté pour le hindi. En se limitant à deux types de PC, N+V et V+V, ils essayent de répondre à deux questions fondamentales, à savoir comment déterminer si, dans une combinaison de nom et d'un verbe, le nom est incorporé dans un prédicat complexe ou s'il s'agit d'un argument de verbe, et pour une combinaison de deux verbes, le deuxième verbe est un verbe aspectuel/modal ou un explicateur verbal.

Pour répondre à la première question (type N+V), ils proposent les tests suivants:

1. Ajout d'une marque d'accusatif au nom
2. « Constituency tests » (déplacement du nom, question et coordination)
3. Ajout de modificateurs au syntagme nominal

1. Ajout d'une marque d'accusatif au nom

Dans un prédicat complexe on ne peut ajouter de marque d'accusatif alors que c'est possible

si le nom est l'argument du verbe.

राम ने चाय ली
Rām ne cāy lī,
Ram a pris le thé

राम ने उस चाय को लिया जो
Rām ne us cāy ko liyā jo…
Ram a pris le thé qui …

राम ने जम्हाई ली
Rām ne jamhāī lī
Ram a bâillé (a pris un bâillement)

*राम ने उस जम्हाई को लिया जो
*Rām ne us jamhāī ko liyā…
*Ram a pris le bâillement qui…

2. « Constituency tests » (déplacement, question et coordination)

2.1 Déplacement du nom

En général les noms en hindi peuvent apparaître dans des positions non canoniques. Si une combinaison N+V n'accepte pas de tels déplacements, c'est un PC.

उसने सुबह उठकर चाय ली
Usne subah uṭhkar cāy lī
Il a pris du thé en se réveillant le matin

चाय उसने सुबह उठकर ली
Cāy usne subah uṭhkar lī
Le thé, il l'a pris en se réveillant le matin

उसने प्रतियोगिता में भाग लिया
Usne pratiyogitā meṃ bhāg liyā
Il a pris part dans une compétition

* भाग उसने प्रतियोगिता में लिया
*bhāg usne pratiyogitā meṃ liyā
*part, il l'a prise dans une compétition

2.2 Question-réponse

Considérons ces deux phrases :

a) राम ने जम्हाई ली

Rām ne jamhāī lī
Ram a bâillé (a pris un bâillement)

b) वह बाज़ार से फल लाया
Vah bāzār se phal lāyā
Il a apporté des fruits du marché

Les questions sur les actions de ces phrases révèlent une autre propriété des noms incorporés.

राम ने क्या किया ?
Rām ne kyā kiyā
Qu'a fait Ram ?

वह क्या लाया / उसने क्या किया ?
Vah kyā layā/usne kyā kiyā ?
Qu'a-t-il apporté / Qu'a-t-il fait ?
A la différence des noms incorporés, pour la phrase b) les questions peuvent isoler soit le nom (phal) soit tout le syntagme verbal (phal lāyā).

2.3 Coordination

La coordination n'est pas possible entre les noms incorporés, alors que c'est parfaitement possible pour les arguments du verbe.

लोग चाय और नामकीन ले रहे थे
Log cāy aur nāmkīn le rahe the
Les gens prenaient du thé et du salé.

*लोग निन्द और जम्हाई ले रहे थे
*log nind aur jamhāī le rahe the
*les gens prenaient du sommeil et du bâillement

3. Ajout de modificateurs au nom

Les arguments du verbe peuvent être modifiés par des adjectifs, numéraux, déterminants etc., mais pas les noms incorporés.

मैं ने आज बहुत चाय पी
Maiṃ ne āj bahut cāy pī
J'ai bu beaucoup de thé aujourd'hui

उसने ज़ोर से धक्का मारा
Usne zor se dhakkā mārā

Il a frappé fort
उसने मेरी बहुत मदद की
Usne merī bahut madad kī
Il m'a beaucoup aidé

Dans la première phrase, le modificateur *bahut* modifie le nom (cāy), mais dans les suivantes *zor se* et *bahut* modifient tout le syntagme verbal et pas les noms.

Verbes conjoints (compound verbs)

Parmi les 5 séquences de type V+V, seuls sont considérés les trois types suivants comme des PC car ils se comportent comme un verbe simple et des adverbes ou la marque de négation ne peuvent pas s'intercaler entre les composants :

V-obl+lagnā (inceptif)

V-inf+ paḍnā (modalité du devoir)

V-base+explicateur verbal

Les deux autres séquences n'ont pas été retenues car les tests ont montré que l'on pouvait placer des adverbes ou marque de négation entre les composants de la séquence :

a) Usne mujhe khat likhne ko **nahīṃ** kahā
b) Vah nahākar **jaldī se** ā gayā

Pour identifier les verbes conjoints, les auteurs reprennent les tests de Butt[17] et Paul[18] :

1. Adverbes
2. Négation
3. Nominalisation
4. Passivisation
5. Causativisation
6. Déplacement

Ils affirment que ces séquences sont le résultat de processus de dérivation et par conséquent elles doivent être stockées dans des bases de données lexicales.

Nous verrons par la suite que les différents tests mentionnés dans cette section ont été repris par d'autres chercheurs pour l'identification des PC.

[17] Butt, M. 1995. "The Projection of Arguments: Lexical and Compositional Factors", A. Alsina et al. (eds.), Complex Predicates. CSLI Publications, Stanford.
[18] Paul, S. 2004. An HPSG Account of Bangla Compound Verbs with LKB Implementation, Ph.D. Dissertation. CALT, University of Hyderabad.

5.1.4 Autres

D'autres outils et ressources ont été créés (dont certains sont disponibles en téléchargement) : analyseur morphologique, outils de translittération, corpus annotés, OCR, reconnaissance vocale, conversion de texte en parole, correcteurs orthographiques, reconnaissance de l'écriture manuscrite, moteur de recherche cross-lingue etc.

5.2 Détection automatique des prédicats complexes

De nombreux acteurs de TAL se sont intéressés à la détection des PC pour réaliser différentes tâches de traitement automatique. Ils ont adopté différentes approches: certains utilisent des corpus parallèles, d'autres des corpus monolingues, certains appliquent des méthodes statistiques, d'autres les combinent avec des règles.

5.2.1 Les différentes approches

5.2.1.1 Corpus parallèle

(Mukerjee et Raina, 2006) ont essayé de traiter la question en utilisant un corpus parallèle anglais-hindi[19] et l'étiquetage morphosyntaxique du corpus anglais avec Brill Tagger. Leur but est de créer une base de données de PC. Les types de PC considérés sont adj+V, nom+V, adv+V et V+V.

Cette méthode n'utilise pas beaucoup les caractéristiques du hindi, elle peut donc être appliquée à d'autres langues.

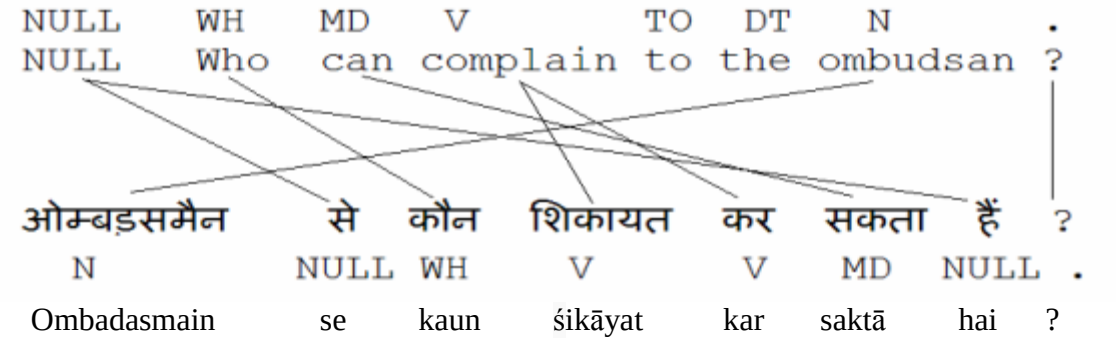
Selon les auteurs, les méthodes basées sur des règles ne sont pas très efficaces car il n'existe pas vraiment de règle pour distinguer les PC de constructions semblables mais qui n'appartiennent pas à la catégorie des PC. Avec quelle règle pourrait-on distinguer ānumati denā (donner son approbation, approuver) de kitāb denā (donner un livre) ? Même s'il existe des règles, elles découlent de propriétés sémantiques : un livre est un objet physique et peut être donné. Cependant, en utilisant un corpus parallèle, un prédicat complexe peut être traduit par un verbe simple et les PC peuvent être détectés à l'aide de la projection des étiquettes morphosyntaxiques. Par la suite, des contraintes linguistiques sont appliquées pour déterminer si le groupe détecté est un PC. Celles-ci consistent en une liste de verbes supports (light verbs) susceptibles d'apparaître avec des composants adjectival, nominal, adverbial ou verbal d'un

[19] Il s'agit du corpus parallèle EMILLE qui comporte 200 000 mots, construit par le gouvernement de la Royaume-Uni pour les immigrants.

PC.

Dans cette méthode, les auteurs alignent tout d'abord les corpus au niveau des mots. Les phrases en anglais sont étiquetées et les étiquettes sont projetées sur les phrases en hindi. Les auteurs ont observé que les mots étiquetés comme verbes par projection dont les étiquettes originelles sont nom, verbe, adjectif ou adverbe en hindi et qui sont suivis d'un verbe support, sont souvent des PC.

Illustration avec un exemple de projection :



Dans l'exemple ci-dessus, on voit que le PC śikāyat kar donne V+V par projection. Etant donné, que śikāyat est un nom en hindi, le tout est identifié comme un PC. Par ailleurs, on remarque que les unités n'ayant pas d'équivalent dans l'une ou l'autre langue sont marquées NULL.

Cette méthode est cependant limitée car elle ne détecte que les PC ayant été traduits par un seul verbe en anglais. Par exemple, *javāb de (réponse donne)* sera traduit en anglais par «answer » ou par « give answer ». Dans le deuxième cas, le PC ne sera pas détecté.

Sur 4400 phrases, 1439 PC ont été identifiés : N+V 788, A+V 107, Adv+V 18, V+V 526.

Cette méthode ne détecte pas certains PC, mais ceux qui sont détectés semblent être corrects. La précision est de 83% et le rappel 46%.

Les erreurs sont dues :

- aux prédicats complexes du hindi traduits par des prédicats complexes en anglais (faux négatifs)
- à l'insuffisance du dictionnaire hindi utilisé pour un grand nombre de lexèmes (faux positifs).
- à l'homographie : par exemple, le mot kī peut être considéré soit une marque de possessif, soit comme le participe accompli du verbe karnā « faire ». Dans 0,2% des

cas, il n'a pas été correctement étiqueté.

(Mukerjee et Raina) abordent aussi la question des PC discontinus. Comme nous l'avons vu précédemment, les composants du PC peuvent être séparés et même parfois considérablement éloignés l'un de l'autre, ce qui pose à coup sûr des difficultés pour l'identification automatique. Ce phénomène existe dans beaucoup d'autres langues et de nombreux chercheurs se sont penchés sur la question (Villavicencio et al. 2004).
Les auteurs n'ont pas essayé d'identifier les PC discontinus mais ils pensent que la projection des étiquettes morphosyntaxiques peut donner de bons résultats car les systèmes basés sur des règles ne sont pas très efficaces pour ce genre de constructions.

(Mahesh et Sinha, 2009) identifient aussi les PC en utilisant un corpus parallèle, mais à la différence de (Mukerjee et Raina), ils projettent le sens du verbe support dans un corpus parallèle, s'appuyant sur l'idée que le sens du prédicat complexe est différent du sens du verbe support.
Leur méthode est composée des phases suivantes : a) alignement des phrases hindi-anglais b) création d'une liste de verbes supports et de leurs traductions en anglais c) génération de toutes les formes morphologiques pour chacun des verbes supports, d) exécution de l'algorithme de détection pour chaque phrase alignée.
L'algorithme de détection utilise une liste de « exit words » (ex : ne, lekin, se, apne, jiskā, tumhārā) qui ne peuvent former des PC, ainsi qu'une liste de « stopwords » (ex : nahīṃ, na, bhī, pahale, bād meṃ) qui peuvent s'intercaler entre les composants du PC.

L'algorithme consiste à :
1. chercher un verbe support et ses formes dans la phrase en hindi et marquer sa position ;
2. chercher sa traduction dans la phrase correspondante en anglais ;
3. si la traduction du verbe a été trouvée, passer à la phrase suivante. En l'absence de résultat, relever les mots à gauche du verbe support dans la phrase hindi ;
4. si l'on rencontre un stopword dans les mots à gauche du verbe support, ignorer et continuer ;
5. si le mot n'est pas un stopword, arrêter le déplacement à gauche et récupérer le mot ;
6. si le mot récupéré est un exit word, passer à la phrase suivante. Dans le cas contraire, le mot et le verbe support forment un PC.

Aucun étiqueteur morphosyntaxique n'est utilisé dans ce processus. Ainsi, la nature du mot à gauche du verbe support n'est pas connue.

Cette méthode a été testée sur plusieurs corpus EMILLE et a atteint une F-mesure entre 88% et 97%, ce qui est un bon résultat compte tenu de sa simplicité. Elle se révèle plus efficace que le système de Mukerjee et Raina. En effet, celui-ci n'arrive pas à détecter les PC tels que khushi hona « to feel happy », salaah lena « to seek advice », tandis que celui de Mahesh et Sinha y arrive parfaitement.
C'est le meilleur résultat parmi les méthodes traitant autant de types de PC. Nous verrons plus loin que (Chakrabarti et al.) ont obtenu un meilleur résultat mais leur système ne traite que les cas V+V et ils ne précisent pas le rappel.

5.2.1.2 Corpus monolingue

(Begum et al. 2011) ont tenté de traiter les locutions verbales uniquement (N/Adj + V) en utilisant un corpus hindi. Ils ont atteint une précision de 85,28%.
Tout d'abord, (Begum et al.) présentent quelques conditions pour déterminer manuellement si telle ou telle combinaison de nom et de verbe est un PC ou pas (ils reprennent certaines conditions présentées en détails dans le chapitre Hindi Wordnet) et leur pertinence pour l'identification automatique.
Les tests sont les suivants :

1. Test de coordination (D1)
2. Test de réponse (D2)
3. Relativisation (D3)
4. Ajout de la marque d'accusatif (D4)
5. Ajout du pronom démonstratif (D5)

Après avoir mené une enquête auprès de 20 locuteurs natifs pour s'assurer de la pertinence de ces conditions, les auteurs sont passés à la phase d'automatisation de l'identification des PC à l'aide d'un outil statistique. Ils utilisent deux corpus annotés qui font partie de Hyderabad Dependancy Treebank. Le corpus d'entraînement contient 4500 phrases annotées manuellement et le corpus de test comporte 1800 phrases.

Les auteurs définissent 7 attributs répartis en 3 catégories : catégorie lexicale, catégorie binaire et catégorie basée sur les cooccurrences qui vont aider à classer les expressions N/Adj+V comme locution verbale ou expression littérale.

1. Verbe (f1)

Certains verbes sont plus enclins à apparaître comme verbe support servant à former des locutions verbales (karna « faire »). D'autres, au contraire, apparaissent plutôt dans des expressions littérales.

2. Objet (nom, adjectif) (f2)

Certains objets ont de grandes chances de former des locutions verbales comparés à d'autres.

3. Catégorie sémantique des objets (f3)

Certains travaux donnent de l'importance à la catégorie sémantique du nom/adjectif pour identifier les locutions verbales. Ils extraient la catégorie sémantique dans le Wordnet hindi. Il y en a 83 au total (ex : Abstraction, Etat, Objet physique).

4. Indicateur de postposition (f4)

C'est un booléen qui indique s'il y a une postposition entre un nom/adjectif et un verbe.

5. Indicateur de démonstratif (f5)

C'est un booléen qui indique si la paire nom/adj-verbe est précédée d'un démonstratif ou pas.

6. Fréquence d'un verbe correspondant à un objet particulier (f6)

Si un nom/adjectif apparaît souvent avec un verbe particulier, il est probable que cette paire forme une locution verbale.

7. Indicateur de l'argument du verbe (f7)

Cet attribut calcule le nombre moyen de postpositions qui apparaissent devant une paire nom/adj-verbe. Si une expression apparait avec un grand nombre de postpositions, cela signifie que l'argument de son verbe est probablement satisfait car chacune de ses postpositions est précédée d'un nom/adjectif qui peut potentiellement être l'argument du verbe. Ainsi, cette paire a plus de chance d'être une locution verbale.

Enfin, en utilisant l'entropie maximale[20] et ces attribut, les auteurs ont procédé à une classification binaire des expressions N/Adj+V en LV ou non LV.

Après des essais avec différents attributs, le meilleur résultat donne la combinaison des 6 premiers attributs. Il est intéressant de noter que l'ajout de (f7) fait baisser la précision de 7,78 %.

Dans le futur, (Begum et al.) vont essayer d'automatiser les D1 et D3, à condition d'avoir un plus grand corpus. Certaines conditions (D2) ne peuvent pas être automatisées, mais elles

[20] Maximum entropy toolkit http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

peuvent être un complément pour l'outil statistique.

(Chakrabarti et al. 2008) se concentrent sur les PC de type V+V en utilisant toujours un corpus hindi.
Parmi les séquences V+V, seul un sous-ensemble est un véritable PC, plus précisément les constructions suivantes : inf+padnā, inf-e+lagnā, base+V2.
Les deux premières structures, nommées *verbes conjoints syntaxiques* par les auteurs, sont invariables et prévisibles. Dans la troisième structure, le choix du deuxième verbe n'est pas prévisible. Les auteurs ont nommé cette structure *verbes conjoints lexicaux (lexical compound verbs)* et proposent d'inclure les verbes de cette catégorie dans la base lexicale en utilisant une méthode afin de les extraire automatiquement.
Leur algorithme est le suivant : si un verbe apparaît sous sa forme de base et qu'il est suivi d'un des verbes de la liste prédéfinie, alors les deux verbes forment un verbe conjoint lexical.
Par la suite, les auteurs ont demandé à dix locuteurs natifs de construire des phrases avec les séquences extraites. S'ils y arrivent, les séquences sont enregistrées comme des verbes conjoints lexicaux.

Au départ, les tests portant sur deux corpus différents atteignent des précisions de 70% et de 79%. La perte en précision est due à l'ambiguïté morphosyntaxique de certains mots, à la passivisation et aux usages idiomatiques. Après avoir pris des mesures pour réduire le nombre d'erreurs (sans en préciser le contenu), ils atteignent 98% de précision.

6 DEJALU

6.1 Apprentissage des Langues Assisté par Ordinateur (ALAO)

Avec le développement des technologies d'information et de communication, le e-learning, défini par l'Union européenne comme « l'utilisation des nouvelles technologies multimédias de l'Internet pour améliorer la qualité de l'apprentissage en facilitant d'une part l'accès à des ressources et à des services, d'autre part les échanges et la collaboration à distance », s'est imposé comme une solution souple et accessible dans de nombreux domaines que ce soit dans un cadre institutionnel (école, université), professionnel ou privé. Utiliser un ordinateur pour apprendre aujourd'hui ne surprend personne et l'apprentissage des langues n'y déroge pas. Un ordinateur permet de présenter les notions à acquérir sous diverses formes : exercices, tutoriels, jeux, simulation. En ce qui concerne les exercices, il s'agit d'exercices traditionnels comme les textes à trous, les questionnaires à choix multiples, les exercices d'appariement de mots, de remise en ordre de mots, etc. Les apprenants doivent cependant produire des phrases complètes, dont la complexité évolue au fil de leur progression. Des techniques basiques peuvent corriger des fautes simples, comme la comparaison de chaînes et la reconnaissance de patrons, mais la complexité et la variété des erreurs commises nécessitent un traitement plus intelligent. Dans un tutoriel, l'ordinateur simule un tuteur travaillant individuellement avec l'étudiant. Les jeux sont un bon moyen de présenter de manière ludique des notions jugées rébarbatives, comme par exemple les règles de grammaire ou l'apprentissage du vocabulaire. On qualifie de simulation les logiciels qui tentent de confronter l'apprenant à une situation de communication de la vie réelle. De plus, l'ordinateur est plus disponible qu'un enseignant. Les élèves peuvent progresser à leur propre rythme et sélectionner les exercices qu'ils veulent faire. Ils sont moins intimidés face à une machine que face à une classe ou à un enseignant chargé de les évaluer.

Le processus de conception d'un logiciel est très complexe. Les grandes étapes, résumées par (L'Haire, 2011), sont :

i. analyse des besoins, définition d'un public cible, étude de marché, identification des prérequis, spécification des objectifs pédagogiques, choix de la plateforme informatique, fixation du temps imparti pour l'accomplissement des tâches par l'apprenant, etc. ii. scénario pédagogique détaillé; iii. maquette complète de l'interface et des détails; iv. réalisation; v. test complet du logiciel et corrections; vi. documentation et maintenance du logiciel.

L'Apprentissage des Langues Assisté par Ordinateur (ALAO)[21] est un domaine très prolifique de l'utilisation de l'ordinateur pour la didactique et les applications du TAL à l'ALAO, appelé alors Apprentissage Intelligemment Assisté par Ordinateur (ALIAO), sont nombreuses.

La communauté ALAO est bien organisée disposant des associations spécialisées telles CALICO (Computer Assisted Language Learning and Instruction Consortium), EUROCALL (European Association for CALL), IAALT (International Association for Language Learning Technology), des manifestations scientifiques majeures (WorldCall, Eurocall, Jalt Call), des journaux spécialisés (Alsic, LLT, ReCall, Call, Calico, e-FLT, etc.), des laboratoires de recherche etc.

6.2 Outils d'aide à la lecture

Décrivons maintenant quelques outils d'aide à la lecture.
Dès les années 90, plusieurs projets ont vu le jour :

GLOSSER[22] (Nerbonne et al. 1998) au départ était un outil d'aide à la lecture pour les néerlandais apprenant le français. Par la suite d'autres paires de langues ont été ajoutées (anglais-hongrois, anglais-bulgare, anglais-estonien). GLOSSER est composé des éléments suivants : un analyseur morphosyntaxique (Locolex, Rank Xerox, Grenoble 1995) doté d'un désambiguïsateur stochastique, un dictionnaire considérant le contexte (context-dependent) et un concordancier bilingue avec une recherche indexée pour la rapidité (les tailles des corpus dépassent les 8 Mo).

Voici une capture du résultat obtenu en cliquant sur la forme *offrirent* :

[21] Le terme en anglais le plus employé est Computer- Assisted Language Learning (CALL) mais l'on trouve aussi TELL (Technology-Enhanced Language Learning), qui inclut les autres technologies que l'ordinateur, CBI (Computer Based Instruction), CAI (Computer-Assisted ou Computer-Aided Instruction), ITS (Intelligent Tutoring System), CALLLT (CALL using Language Technologies) et NBLT (Network- based language teaching).
[22] http://www.let.rug.nl/glosser/Glosser/

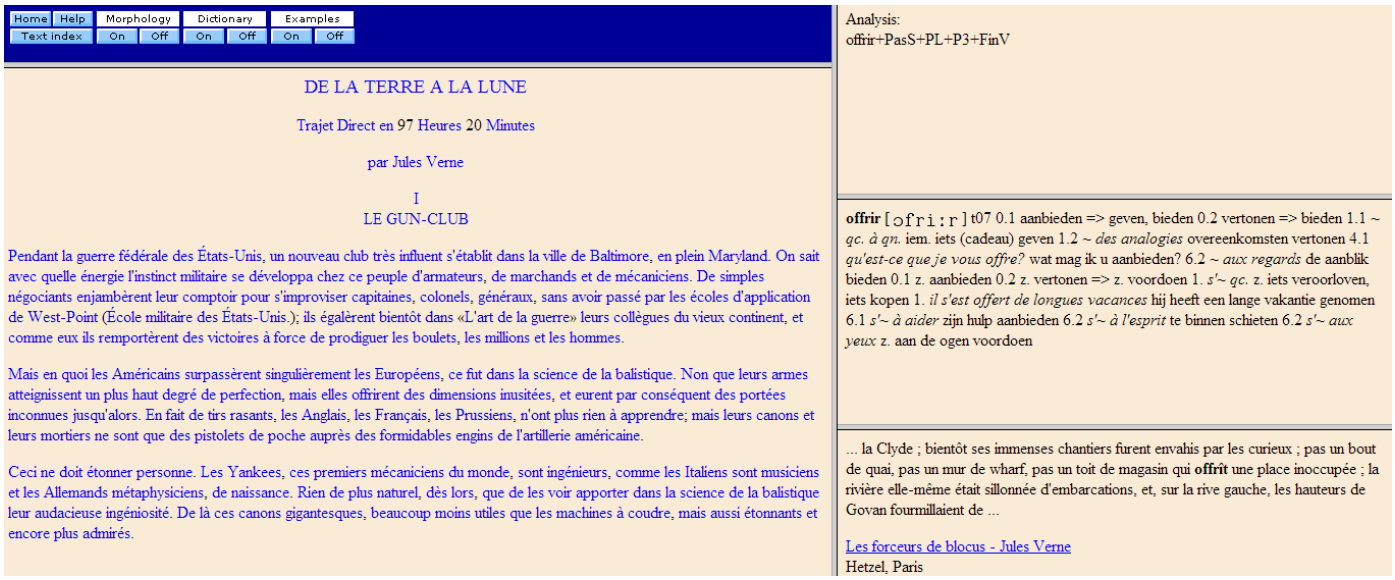


Figure 2 Résultat obtenu en cliquant sur la forme *offrirent*

COMPASS[23] (Breidt & Feldweg 1997) est un projet collaboratif entre des institutions françaises, allemande et anglaise[24]. COMPASS est similaire à GLOSSER mais à la différence de ce dernier, il donne aussi la traduction de l'expression à laquelle appartient le mot sur lequel l'utilisateur a cliqué, toujours en considérant le contexte. Les couples de langue disponibles sont anglais-français et allemand-anglais.

Plus récemment, le nombre de projets similaires a considérablement augmenté :

RAFALES[25] (Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère de Spécialité; Zampa, 2004, 2005) est un logiciel d'aide à la lecture et d'acquisition de vocabulaire spécialisé. Son but est de fournir aux apprenants des textes adaptés leur permettant d'élargir leur espace sémantique. Pour ce faire, le système évalue les connaissances de l'apprenant en se basant sur l'Analyse sémantique latente (LSA)[26] et lui propose un texte correspondant à son niveau permettant d'acquérir progressivement de nouveaux mots par la lecture. En citant les travaux de Vygotsky, les auteurs affirment que « si les textes que le prototype fournit à l'apprenant sont trop proches ou trop éloignés de ce qu'il

[23] http://www.sfs.uni-tuebingen.de/Compass/
24 Rank Xerox Research Centre, Grenoble, France ; Institut für Arbeit und Organisation, Fraunhofer Gesellschaft, Stuttgart, Germany ; Seminar für Sprachwissenschaft, Universität Tübingen, Germany ; Department of Marketing, Advertising and Public Relations, Bournemouth Univ., United Kingdom ; Langues Étrangères appliquées, Université Lyon 2, France
[25] http://www.inrp.fr/biennale/6biennale/Contrib/affich.php?&NUM=125
[26] http://alsic.revues.org/339

connaît déjà, il n'acquerra que peu de connaissance ». Il faut donc trouver cette distance optimale et définir avec LSA ce que les auteurs appellent une Proximité Optimale d'Acquisition, une notion qui se fonde sur la zone proximale de développement (ZPD)[27] de Vygostky. Cependant, à la différence de ZPD, la médiation humaine en est absente.

Le prototype **Didalect** (Hermet et al., 2006) fournit une aide à la lecture du français pour apprenants de niveau intermédiaire à avancé. Il est basé sur un corpus de texte pour lequel il existe des exercices de compréhension. Le système utilise l'analyseur Xerox Incremental Parser (XIP), qui procède à une analyse par morceaux, ainsi qu'un dictionnaire de synonymes et un dictionnaire relationnel. Pour évaluer la réponse de l'apprenant, le système utilise une à trois phrases sélectionnées dans le texte par le concepteur de la question comme réponse ; il bâtit trois listes de mots, (i) celle des mots communs à la réponse de l'apprenant et à la phrase modèle, (ii) celle des mots présents uniquement dans la réponse de l'apprenant et (iii) celle des mots présents uniquement dans la phrase modèle. Les synonymes, hyperonymes et hyponymes sont traités grâce au dictionnaire ad hoc. La réponse est évaluée sur le plan sémantique et syntaxique, en vérifiant la grammaticalité de la phrase et la présence de constituants obligatoires. C'est une étude de faisabilité de reconnaissance et évaluation automatique des réponses libres à un certain type de questions ouvertes destinées à évaluer la compréhension des textes.

REAP (Reader Specific Lexical Practice)[28] propose aux apprenants des textes adaptés à leur niveau et à leurs goûts. L'utilisateur passe un test préliminaire de vocabulaire en cochant les mots qu'il connaît déjà. Après chaque lecture, un test permet d'évaluer et d'actualiser le niveau de l'apprenant. L'étudiant répond aussi à des questions concernant ses centres d'intérêt. Après les questionnaires, REAP présente une sélection de textes récoltés sur le web. Pendant la lecture du texte, REAP souligne les mots qu'il veut faire apprendre à l'étudiant. Les autres mots sont mis en surbrillance lorsque la souris les survole et l'étudiant peut cliquer dessus pour en obtenir la définition. A la fin de la lecture, REAP pose une série de questions à l'apprenant afin d'évaluer l'effet de la lecture sur l'apprentissage du vocabulaire. Ce processus se poursuit jusqu'à ce que l'apprenant souhaite arrêter. REAP possède également une fonctionnalité qui évalue automatiquement la lisibilité des textes en se basant sur les

[27] Cette théorie suggère que les enfants sont aptes à mieux apprendre les problèmes et à s'améliorer davantage autour d'un enfant plus expérimenté, d'un parent ou d'un enseignant, plutôt que d'un enfant à leur niveau cognitif. Cela encourage donc l'apprentissage en milieu scolaire à ce stade de la vie. La zone proximale de développement augmente nettement le potentiel d'un enfant à apprendre plus efficacement (Wikipedia)

[28] http://reap.cs.cmu.edu

caractéristiques lexicales (difficulté sémantique des mots) et syntaxiques (présence de certaines constructions).

ALPHEIOS[29] est un projet pour le grec ancien, le latin et l'arabe (d'autres langues sont en cours de développement). C'est une extension de Firefox et par conséquent elle peut être appliquée à n'importe quel texte sur le web. ALPHEIOS donne des informations morphosyntaxiques et des traductions du mot cliqué. A la demande de l'utilisateur, le système peut aussi proposer un tableau de flexions et stocker le mot dans une liste de mots appris consultable à tout moment.

Figure 3 Résultat obtenu en cliquant sur la forme *perpauci*

NaviLire (Lundiquist et al.) est un outil didactique pour l'enseignement des langues étrangères. Il est basé sur le principe d'une navigation thématique au sein de textes préalablement didactisés manuellement avec un outil d'annotation spécifique (NoteTool). L'outil se base sur le principe de la navigation dans les textes. Les apprenants sont invités à faire des exercices de repérage des unités textuelles, notamment celles qui contribuent à la cohérence du texte et la manière dont elles sont agencées dans le texte (anaphores, connecteurs, discours rapporté etc).

[29] http://alpheios.net

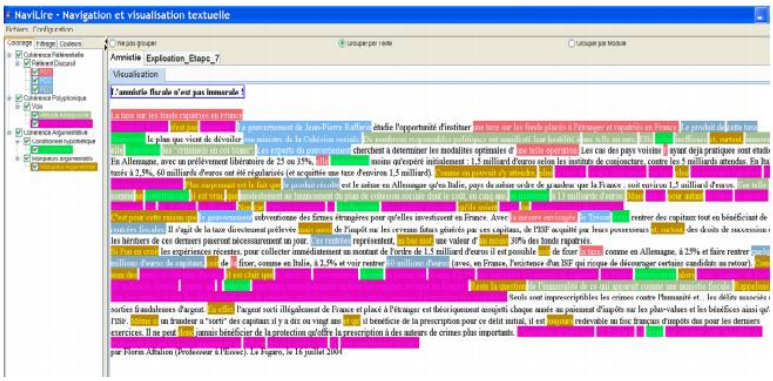


Figure 4 Exemple d'annotation

WERTi (Meurers et al. 2010) est une autre extension de Firefox, cette fois-ci pour des langues européennes modernes : anglais, allemand, espagnol. Le principe de ce projet est d'attirer l'attention du lecteur sur certaines formes afin de lui permettre d'acquérir des propriétés spécifiques de la langue. Les unités mises en évidence (considérées comme difficiles pour les apprenants de l'anglais comme langue étrangère) sont les articles, les déterminants, les prépositions, « noun countability », le choix entre le gérondif et l'infinitif précédé de « to », les phrases interrogatives et les verbes à particules. Le système possède 3 types d'action : coloration des catégories choisies, identification automatique du motif en cliquant dessus et coloration, génération de texte à trous avec ou sans choix multiples.

La capture suivante est l'illustration de la fonctionnalité de la mise en valeur des verbes à particules :

Meanwhile, doctors in London have begun treating a Briton who **contracted** Ebola **in** Sierra Leone.

William Pooley, a 29-year-old volunteer nurse, was flown home in a specially equipped military aircraft on Sunday and **taken to** Hampstead's Royal Free Hospital.

He volunteered to go to **care for** victims of the Ebola outbreak.

Figure 5 Coloration des verbes à particules

ARET (Arabic Reading Enhancement Tool) est un outil développé par (Maamouri et al. 2012) pour utilisation en classe. L'outil intègre un analyseur morphologique pour l'arabe moderne afin de fournir des informations grammaticales sur les mots, ainsi qu'un corpus indexé pour le concordancier. ARET utilise aussi un outil commercial de conversion de texte en parole pour l'arabe moderne. ARET a deux sous-parties : ARFT (Outil de facilitation de la lecture en arabe) et ARAT (Outil d'évaluation de la lecture en arabe).

L’apprenant peut accéder aux textes de Al-Kitaab par volume, chapitre ou page. Il peut cliquer sur n’importe quel mot pour obtenir une analyse morphologique complète, la traduction de la forme dépendant du contexte, ainsi que la traduction du lemme associé et celles des formes ayant la même racine. ARFT a aussi trois autres fonctionnalités importantes : i) l’outil ajoute ou enlève les signes diacritiques du passage affiché, ii) il fournit un concordancier et iii) si la souris survole un mot sans signe diacritique, une fenêtre pop-up apparaît donnant la forme avec les signes diacritiques.

Figure 6 Exemple d’annotation d’un texte en langue arabe

6.3 Qu’est-ce que DEJALU ?

DEJALU est un projet en cours de développement au sein de l'Equipe de Recherche Textes, Informatique, Multilinguisme (ER-TIM). Il vise à élaborer un outil web pour l’apprentissage des langues favorisant la pratique précoce et intensive de la lecture, mais il s’agit d’une lecture enrichie automatiquement d’informations linguistiques permettant au lecteur de pratiquer une lecture autonome et de développer ses compétences lexicales en réception, ses stratégies de lecture et de compréhension, sa réflexion métalinguistique. L’application vise

différentes fonctionnalités simples et intuitives d'analyse et d'annotation textuelles qui varient selon les langues (mise en évidence de certaines parties du discours, concordance des unités lexicales, surlignage de certaines structures considérées comme difficile pour les apprenants, phonétiseurs, mesure de lisibilité etc.). L'application sera dans l'idéal utilisable sur différents supports afin de multiplier les situations de lecture.

DEJALU propose des textes mais le but final serait que l'utilisateur puisse soumettre ses propres textes conformes à ses centres d'intérêt, à ses nécessités professionnelles et à son niveau. Les textes sont enrichis à la demande.

Au jour d'aujourd'hui, l'application, encore fruste, se concentre sur 4 langues (anglais, hindi, hongrois, thaï) typologiquement très différentes, dans le but d'explorer les types d'annotation (fonctionnalités) utiles à l'apprenti-lecteur. L'objectif ultime de DEJALU est de proposer, pour des langues suffisamment dotées d'outils de TAL, une méthodologie permettant "simplement" leur intégration dans ce dispositif d'aide à la lecture.
La capture d'écran ci-dessous est celle de la sobre et fonctionnelle interface de l'application :

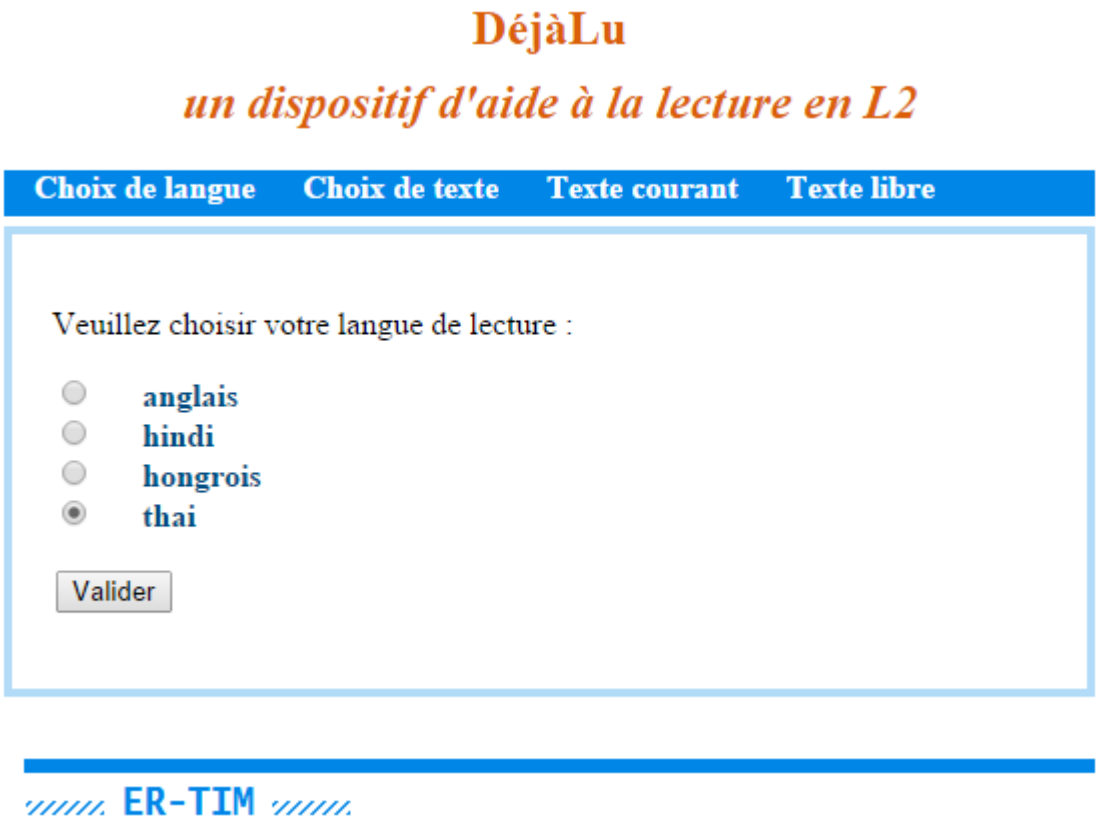
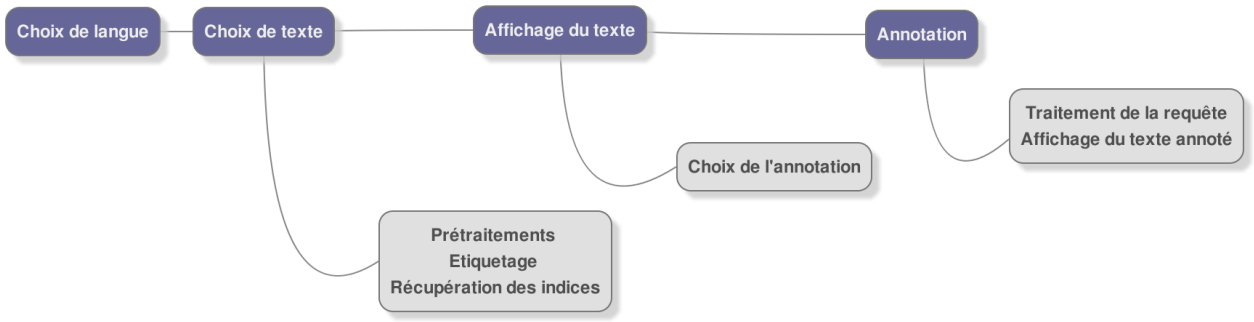


Figure 7 DEJALU - Page d'accueil

6.4 Architecture de DEJALU

Nous pouvons représenter la structure de DEJALU par le schéma suivant :



L'utilisateur est invité à choisir une langue puis un texte. Dès que le texte est choisi, un certain nombre de traitements sont effectués : prétraitements (suppression des blancs et sauts de lignes multiples, des caractères invisibles), étiquetage (un étiqueteur est lancé et le texte est regénéré à partir de la sortie de l'étiquetage), récupération des indices (les informations sur les indices des mots et leurs étiquettes sont stockées au format json).

Après ces traitements, dont le temps d'exécution varie selon la longueur du texte et la langue choisie, le texte est affiché. L'utilisateur peut alors choisir ce qu'il veut annoter. Les scripts correspondants à la requête sont exécutés et le texte annoté est affiché.

La capture suivante illustre la fonctionnalité d'annotation des prédicats complexes en hindi :

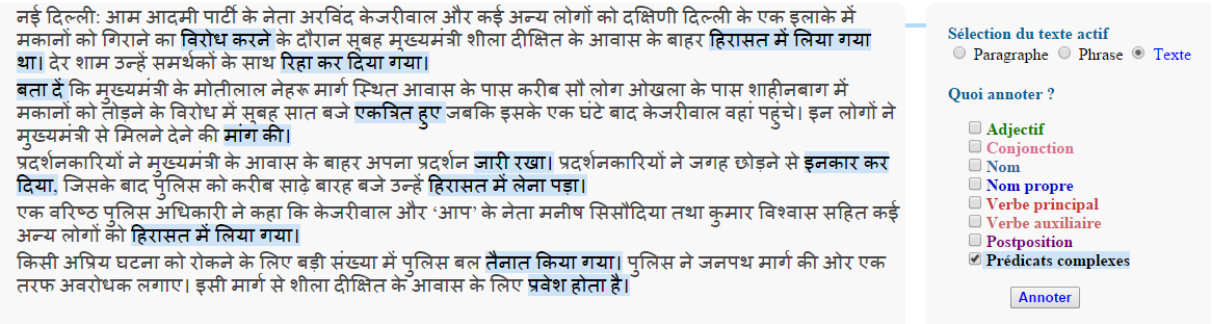


Figure 8 DEJALU - annotation des prédicats complexes hindi

Le dispositif DEJALU s'organise autour d'une architecture client-serveur.

Les diverses demandes de l'utilisateur (choix de langue, de niveau, de texte, d'annotation particulière, etc.), prises en charge par un module javaScript AJAX, sont envoyées au serveur sous forme de requêtes HTTP normalisées (type / session / paramètres).

Côté serveur, une application CGG-Perl intercepte les requêtes et délègue leur traitement à des modules Perl spécifiques. Les données issues de ces traitements (les réponses aux

requêtes) sont renvoyées, en format JSON, au navigateur pour affichage.
Schématiquement, on peut représenter le processus de la manière suivante :

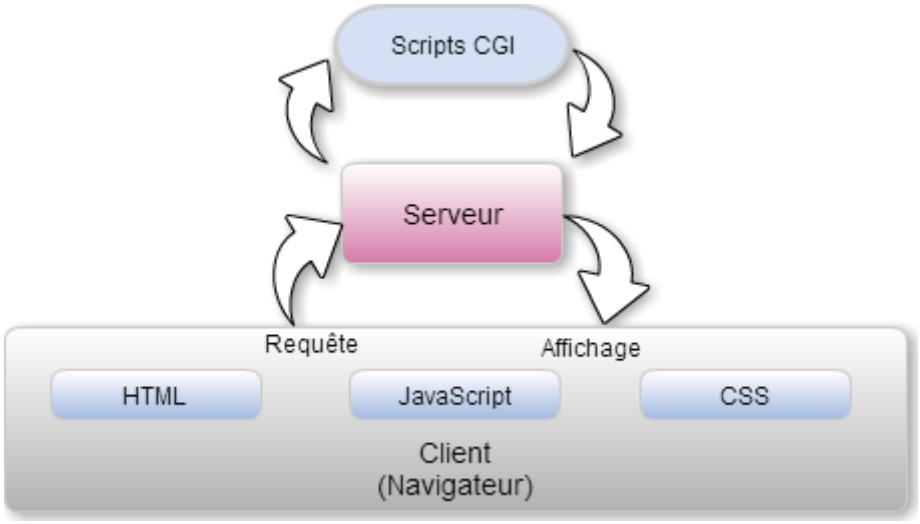


Figure 9

6.5 La détection des prédicats complexes hindi dans DEJALU

Nous avons vu précédemment dans le chapitre 5.2 que la détection des prédicats complexes dans les travaux antérieurs se faisait soit à l'aide de corpus parallèles soit avec des méthodes statistiques. Pour le premier, nous ne disposons pas de bi-textes ni d'outil d'alignement automatique. Pour le deuxième, des corpus annotés de grande taille sont nécessaires mais nous n'y avons pas accès. Nous avons vu aussi que la détection des PC avec des règles n'était efficace que pour certains types de PC.
Nous avons adopté une méthode similaire à celle de (Meurers et al. 2010) pour la détection des verbes à particules de l'anglais dans l'outil Werti, méthode qui consiste à utiliser une liste de verbes à particule en les considérant comme un phénomène lexical. Contrairement à l'anglais, le hindi est une langue morphologiquement riche. Même en ayant une liste de PC, leur détection révèle de nombreuses difficultés. Il faut tenir compte de tous les temps/aspects/modes ainsi que des transformations morphologiques propres à certains verbes et temps.

Description de la liste

Un grand nombre de PC a été stocké dans le Wordnet hindi. Nous disposons d'une liste de 6284 verbes qu'il est nécessaire de trier compte tenu que les expressions peuvent être

composées d'un, deux, trois, quatre, cinq ou six formants et qu'elles ne sont pas forcément des PC.

Voici quelques exemples :

unigramme	करना	karnā	faire
bigramme	हैरान करना	hairān karnā	étonner
trigramme	थप्पड़ रसीद करना	thappad rasīd karnā	gifler
quadrigramme	फिर से प्राप्त करना	phir se prāpt karnā	obtenir à nouveau
pentagramme	हाथ पर हाथ रखकर बैठना	hāth par hāth rakhkar baithnā	rester sans rien faire
hexagramme	एड़ी चोटी का पसीना एक करना	edī cotī kā pasīnā ek karnā	faire du zèle

Les unigrammes sont des verbes simples. Les pentagrammes et les hexagrammes sont des expressions idiomatiques. Les quadrigrammes, quant à eux, sont soit des expressions idiomatiques, soit des verbes simples avec des arguments. Nous ne retenons donc que les bigrammes et les trigrammes.

Après le tri, nous obtenons une liste de 3711 verbes. Nous possédons aussi une liste de verbes simples avec des explicateurs verbaux qui sont les plus fréquemment utilisés avec ces verbes-là (422 au total). Nous les considérons comme des prédicats complexes aspectuels et les ajoutons dans notre liste.

Il faut aussi noter que nous trouvons parfois des prédicats complexes accompagnés d'explicateurs verbaux, des formes plus ou moins figées, comme रद्द कर देना radd kar denā « éliminer », स्वीकार कर लेना svikār kar lenā « accepter », etc. Etant donné que tous les cas possibles ne sont pas spécifiés dans la liste et que c'est un processus relativement arbitraire et imprévisible, il est donc nécessaire de traiter les explicateurs verbaux à part.

De même pour la marque de négation qui est parfois incluse dans la liste (फ़र्क न पड़ना phark na padnā « ne pas faire de différence »). Elle nécessite donc un traitement supplémentaire.

Les variantes orthographiques sont prises en compte : les mots contenant des lettres empruntées à l'arabe ou au persan sont présents deux fois, avec et sans le point ajouté en-dessous de la lettre (ख़ुश/खुश करना khush karnā « rendre heureux ») ; les mots écrits avec ou

sans ligature (ख़तम/ख़त्म करना khatam/khatm karnā « terminer ») ; les prononciations différentes des emprunts à l'arabe ou au persan मुक़ाबला / मुक़ाबिला करना muqabalā/muqabilā karnā « concourir ».

Au passage, nous remarquons qu'il y a aussi dans la liste des emprunts à l'anglais (ब्लॉक कर देना blok kar denā « bloquer », टेक ऑफ करना tek of karnā « décoller »).

Le premier formant (les deux premiers formants pour les trigrammes) est le plus souvent invariable. Il peut être variable s'il s'agit d'un adjectif, d'un nom ou d'un participe variables.

Ces variations sont parfois présentes dans la liste :

नारा_लगाना nārā lagānā	नारे_लगाना nare lagānā	crier des slogans
अनसुना_करना ansunā karnā	अनसुनी_करना ansunī karnā	ignorer
पूरा_करना purā karnā	पूरी_करना purī karnā	réaliser

Dans la méthode que nous présentons ci-dessous, les variations du premier formant ne sont pas traitées, car leur traitement aurait nécessité de disposer d'un lemmatiseur robuste ce qui n'existe pas encore pour le hindi.

Description de la méthode

Ainsi, comment détecter avec une expression régulière, de manière optimale, les prédicats complexes d'un texte en se servant d'une liste de PC, en prenant en compte les contraintes suivantes:

1. transformations morpho-phonologiques dans la base verbale
2. temps/aspects/modes
3. voix
4. particules entre les composants (parfois dans la liste parfois non)
5. grande distance entre les composants
6. explicateurs verbaux
7. variantes orthographiques des désinences verbales

Les transformations morpho-phonologiques dans la base verbale sont possibles dans les cas suivants :

- pour former le participe passé, l'impératif, le subjonctif (et par conséquent le futur) de certains verbes (karnā, lenā, denā, honā, jānā)

- quand la base verbale se termine par une voyelle, le glide –y– s'intercale entre la base et la désinence

Les temps/aspects/modes et la voix sont récapitulés dans le schéma suivant classifié selon le type de formation (par commodité pour les expressions régulières). Le présomptif étant compatible avec divers aspects, temps et modes, nous ajoutons un astérisque quand c'est le cas. Les transformations morpho-phonologiques portent sur le participe accompli, l'impératif, le subjonctif et le futur uniquement. Les cases concernées sont encadrées en noir.

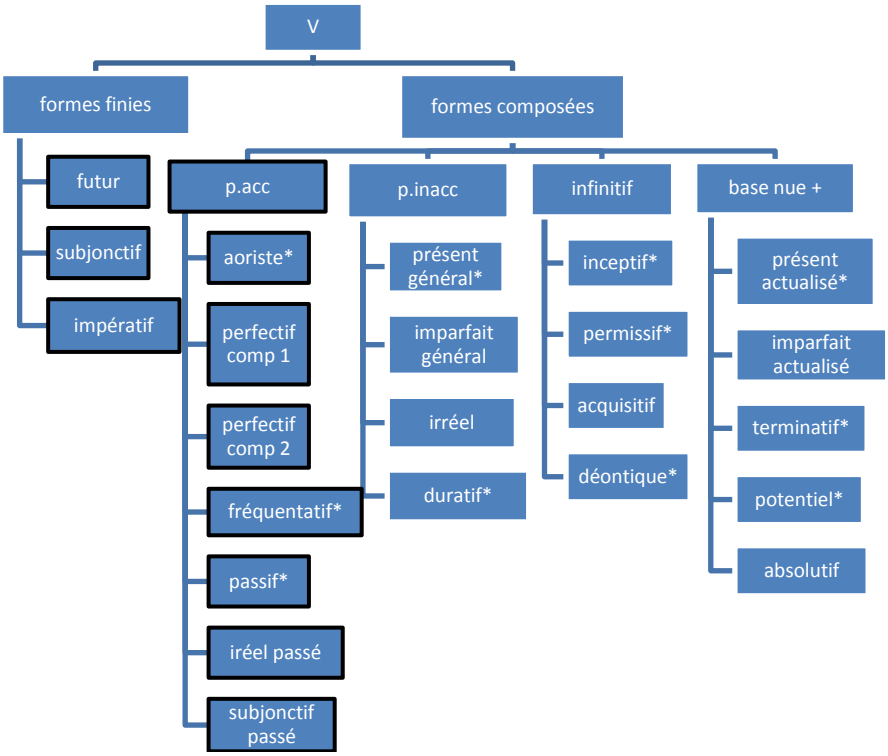


Figure 10 Temps, aspects, modes, voix

Les variations paradigmatique et syntagmatique du prédicat en hindi peuvent être schématisées de la manière suivante :

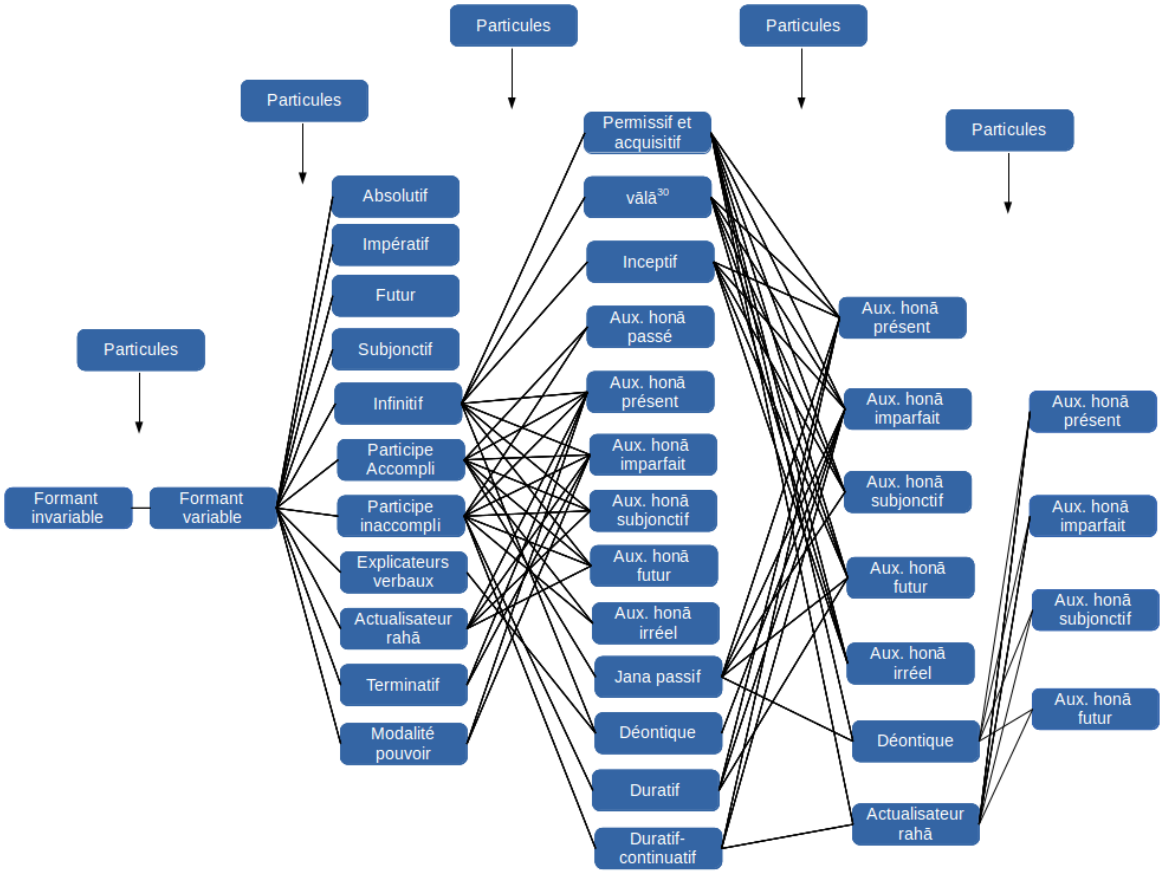


Figure 11 Variations paradigmatiques et syntagmatiques

Nous avons ici fait le choix de ne représenter que les combinaisons les plus courantes et non l'ensemble des combinaisons possibles pour des questions de lisibilité.

A l'aide des schémas ci-dessus et des tableaux, récapitulant la formation de différents temps, aspects, modes et voix, présentés dans l'Annexe 4, nous proposons l'expression régulière suivante qui permet la détection des PC ayant un formant verbal régulier (pour les formants verbaux irréguliers cette expression régulière a été modifiée légèrement):

```
(?:$fr_gauche)+?($inv (?:$particules)? ?$var(?:$classe0)? ?(?:$particules)? ?
(?:(?:$A)(?:$classe1))? ?(?:$particules)? ?(?:(?:$B)(?:$classe2))? ?(?:
$particules)? ?(?:(?:$C)(?:$classe3))?)(?:$fr_droite)
```

L'automate associé à cette expression régulière est présenté dans l'Annexe 5.

Les différentes catégories sont regroupées dans des variables. Elles sont commentées dans le tableau ci-dessous :

$fr_gauche	Il s'avère nécessaire de définir des frontières à gauche afin d'éviter de capturer des formes partielles : sans frontières, *santuṣṭ hona* aurait été identifié dans la forme *asantFuṣṭ hona*. Les frontières à gauche sont : début de ligne, espace,

	ponctuations.
$inv	La partie invariable du prédicat complexe est stockée dans la variable $inv qui est suivie obligatoirement d'une espace.
$var	Cette variable contient la base nue du formant verbal du PC qui peut être variable ou rester sous sa forme de base nue.
$classe0	Les variations possibles[30] du formant verbal ($var) sont stockées dans la variable $classe0. Ce sont des désinences de genres, de nombres des formes finies.
$A	La variable $A regroupe la partie invariable des explicateurs verbaux[31], de la particule vālā, des morphèmes d'aspects secondaires et de modalités, de l'auxiliaire être.
$classe1	Cette variable contient les désinences possibles de la variable précédente.
$B	Elle contient la partie invariable de l'auxiliaire être, des déontiques, de l'actualisateur rahā.
$classe2	Cette classe regroupe les désinences de la variable $B.
$C	La variable $C contient la partie invariable de l'auxiliaire être.
$classe3	Les désinences possibles de l'auxiliaire être conjugués aux différents temps sont dans cette classe.
$particules	Des particules (to, hi, bhi, nahi, bhi nahi, hi nahi) peuvent s'intercaler entre les composants du PC. La variable est placée là où les particules sont susceptibles d'apparaître
$fr_droite	Des frontières à droite sont définies pour ne pas capturer les débuts des mots suivants identiques aux morphèmes des variables $A, $B, $C. Les frontières à droite sont : espace, fin de ligne, ponctuations.

Le contenu des variables est présenté dans l'Annexe 6.

6.6 Tests sur corpus

Nous avons appliqué notre méthode sur deux corpus différents : un corpus de presse et un corpus de textes littéraires. Le corpus de presse est composé des articles de l'année 2013 d'un site de presse[32], il comporte 117 083 mots. Les articles ont été aspirés automatiquement et le contenu textuel a été extrait à l'aide de quelques scripts Perl. Le corpus littéraire (116 659 mots) contient des extraits de romans, des nouvelles, ainsi que des pièces de théâtre d'auteurs indiens. Les textes ont été collectés manuellement à partir du site de « Mahatma Gandhi International Hindi University »[33] qui réunit les œuvres les plus connues d'écrivains renommés de langue hindi ainsi que des livres traduits en hindi.

[30] Les variantes orthographiques de tous les morphèmes variables sont prises en compte dans l'expression régulière.
[31] Les explicateurs verbaux considérés sont : jānā, ānā, lenā, denā, uṭhnā, paḍnā, ḍālnā, mārnā, baiṭhnā.
[32] http://khabar.ndtv.com/
[33] http://hindisamay.com/default.aspx

Sur le corpus de presse, un total de 6216 PC ont été détectés tandis que sur le corpus littéraire c'est un total de 5979 PC qui ont été détectés.

Afin d'évaluer plus précisément nos résultats, nous avons annoté manuellement une partie de chaque corpus. Nous avons décidé de faire deux évaluations différentes. Le corpus presse annoté a été utilisé pour évaluer la détection des prédicats complexes de type N/Adj/Adv +V, alors que le corpus littérature fut utilisé pour évaluer les verbes conjoints, ces derniers étant plus fréquemment utilisés en langage littéraire.

6.6.1 Locutions verbales

Le corpus de presse annoté (5354 mots) contient au total 250 prédicats complexes. Le programme en détecte 241 dont 9 incorrects.

Ceci donne une précision de 96,2%. Les erreurs sont dues à l'ambiguïté syntaxique de la forme *ki* qui peut être soit une postposition soit le participe accompli du verbe karnā « faire ».

सेवा **की** टीमों ने…
Sevā **kī** ṭīmoṃ ne…
Service GEN équipe.OBL.PL ERG

लड़की से बलात्कार **की** कोशिश का मामला…
Ladkī se blātkar **kī** kośiś kā māmlā…
Fille SOC viol GEN tentative GEN affaire

Le rappel est de 92,8%. Les PC non détectés peuvent être répartis en 3 catégories.

- Prédicats complexes absents de la liste dont une sous-catégorie importante composent les prédicats complexes dont le formant non verbal est un emprunt à l'anglais
  अपसेट हो गए थे
  Apseṭ ho gae the

  रजिस्टर्ड कराया गया था
  Rajisṭard krāyā gayā thā
- faute d'orthographe dans le texte
  गिफ्तार कर लिया गया है
  Giftār kar liyā gayā hai (*giftār* au lieu de *giraftār*)
- La grande distance entre les composants du prédicat complexe (ils sont ici représentés en gras)
  निर्णय अकादमी की 21 दिसम्बर को हुई बैठक में **लिया गया**
  Nirṇay akādamī kī 21 disambar ko huī baiṭhak men **liyā gayā**

  फैसला जल्द से जल्द **किया जाएगा**

Phaislā jald se jald **kiyā jāegā**

6.6.2 Verbes conjoints

Le corpus littéraire annoté (5399 mots) comporte 135 verbes conjoints. 116 verbes conjoints ont été détectés dont 1 incorrecte. L'erreur est due au morphème *kar* qui a été identifié comme début d'un verbe conjoint mais qui en réalité était l'absolutif :

बुला	**कर**	**ले**	**गए**
Bulā	kar	le	gae
Appeler ABS		prendre	aller.AOR.PL

Ceci donne une précision de 99,1 % et un rappel de 85,9%. Le rappel relativement faible était attendu compte tenu l'usage relativement imprévisible des explicateurs verbaux.

Quelques exemples détectés :

कर	दी	होती		छोड़	दूँगा		
kar	dī	hotī		chod	duṃgā		
faire	donner.P2.F	être.IRR.F		laisser	donner.FUT.1SM		
टूट	गई	है		बचा	ले	जाती	है
ṭuṭ	gaī	hai		bacā	le	jātī	hai
se casser	aller.P2	être.PRES		sauver	prendre	aller	être.PRES
रो	पड़े			बता	देना	चाहिए	था
ro	padi			batā	denā	cāhie	thā
pleurer	tomber.P2.P			raconter	donner.INF	déontique	être.IMP

Voici quelques exemples de verbes conjoints qui n'ont pas été détectés car étaient absents de notre liste :

उठ	आई	थीं		समझ	बैठा	था	
uṭh	āī	thīṃ		samajh	baiṭhā	thā	
se lever	venir.P2	être.IMP.FPL		comprendre	s'asseoir.P2	être.IMP.S	
लड़	बैठे			हो	उठा		
lad	baiṭhe			ho	uṭhā		
se disputer	s'asseoir.AOR.P			être	se lever.AOR.S		

6.7 Discussion

Nous avons vu que la détection des locutions verbales donne de meilleurs résultats que celle des verbes conjoints. Nous avons pu observer que même en ayant une liste, il y a des limites en matière de détection des prédicats complexes. En dehors du fait que certains prédicats complexes soient simplement absents de notre liste, les limites sont dues à l'ambiguïté morphosyntaxique de certaines formes et à la grande distance entre les composants du prédicat complexe. L'ambiguïté morphosyntaxique peut être enlevée à l'aide de l'étiqueteur morphosyntaxique auquel cas la précision du résultat dépendra de celle de l'étiqueteur. La distance entre les composants ne peut pas être traitée avec notre méthode, sauf si on autorise un certain nombre de mots entre les composants, mais cela engendre inévitablement un bruit important ce qui nous a conduits à éliminer cette option. Parmi les méthodes présentées précédemment, celle qui pourrait traiter le mieux les prédicats complexes discontinus serait selon nous celle de (Mukerjee et Raina, 2006), méthode basée sur des corpus parallèles.
L'enrichissement de la liste aussi bien avec des locutions verbales qu'avec des verbes conjoints doit toujours être envisagé pour améliorer les résultats.

7 Conclusion

Ce travail a permis d'intégrer la détection des prédicats complexes dans DEJALU.

Il a été montré clairement que de solides connaissances linguistiques étaient un préalable indispensable à la détection et annotation des unités complexes telles les prédicats complexes. En effet, notre méthode qui consiste à utiliser une liste de prédicats complexes nécessite de prendre en compte toute la complexité du syntagme verbal hindi et donc une description détaillée de celui-ci.

Concernant le degré d'efficacité de la méthode employée, les tests d'évaluation effectués sur deux petits corpus révèlent une précision élevée et un rappel moins élevé ce qui n'est surprenant au regard de la nature de la méthode mais ce qui importe réellement pour l'ALAO c'est la précision. Nous sommes conscients qu'il faudrait réaliser des tests sur un plus grand corpus annoté afin de mieux évaluer la méthode.

Les résultats observés laissent apparaître les limites de la méthode. Ainsi, nous pouvons souligner en premier point que la liste utilisée n'est pas exhaustive et ne peut pas en définitive l'être. On peut aussi relever le fait que certains prédicats complexes passent aux travers les mailles du filet quand bien même ces derniers sont inscrits dans liste. C'est ce qui se produit en l'occurrence pour les prédicats complexes discontinus, prédicats complexes dont les composants sont séparés les uns des autres par d'autres mots. Enfin les PC détectés par erreur en tant que tel autrement dit les faux positifs l'ont été du fait de l'ambiguïté morphosyntaxique de certains morphèmes impliqués.

Dans ce travail nous n'avons pas traité quelques constructions qui sont en soi des prédicats complexes : par exemple les verbes simples au mode permissif et à l'aspect inceptif. Ces catégories ne posent à priori pas de problème d'identification car on a des règles (l'infinitif à l'oblique + denā/lagnā conjugué aux différents temps, aspects et modes). Nous pourrions l'intégrer sans trop de difficultés au travail précédemment effectué.

Suite à nos recherches, nous pouvons à présent isoler les prédicats complexes en entier, le champ des applications de cette détection ne se limite pas à ce qui avait été notre motivation de départ, en effet elle a toute son utilité pour diversifier les fonctionnalités de DEJALU, notamment pour mettre en place par exemple un concordancier des PC ou pour évaluer la lisibilité des textes (le nombre de prédicats complexes dans un texte pouvant être un des critères de lisibilité).

8 Bibliographie

[1] Begum, R., Jindal, K., Jain, A., Husain, S., Sharma, D.M., 2011. Identification of Conjunct Verbs in Hindi and Its Effect on Parsing Accuracy, in: Gelbukh, A.F. (Ed.), Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 29–40.

[2] Bharati, A., D. M. Sharma, L. Bai and R. Sangal. 2006. AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages. LTRC-TR31

[3] Bhattacharyya, P., Chakrabarti, D., Sarma, V.M., 2006. Complex predicates in Indian languages and wordnets. Lang Resources & Evaluation 40, 331–355. doi:10.1007/s10579-007-9032-x

[4] Butt, M., 1995. The structure of complex predicates in Urdu, Dissertations in linguisti. CSLI Publications, Stanford, Calif.

[5] Chakrabarti,D., Mandalia, H., Priya, R., Sarma, V., Bhattacharyya, P.: Hindi Compound Verbs and their Automatic Extraction.In: International Conference on Computational Linguistics, pp. 27-30 (2008).

[6] Lundquist, Lita; Minel, Jean-Luc; Couto, Javier. (2006). NaviLire, Teaching French by Navigatingin Texts. 2006. Conference: The 11th International Conference. IMPU 2006. Information Processing and Management of Uncertainty in Knowledge-based Systems, No. 11, Paris, Les Cordeliers, France, July 2, 2006 - July 7, 2006.

[7] L'Haire, S., 2011. Traitement automatique des langues et apprentissage des langues assisté par ordinateur : bilan, résultats et perspectives. University of Geneva.

[8] Montaut, A., 2012. Le hindi, Les langues du monde. Peeters, Leuven Paris.

[9] Montaut, A., 1991. Aspects, voix et diathèses en hindi moderne: syntaxe, sémantique, énonciation. Peeters Publishers.

[10] Montaut, A., Joshi, S., n.d. Parlons Hindi, Collection Parlon. L'Harmattan, Paris Montréal.

[11] Mukerjee, A., Soni, A., Raina, A.M., 2006. Detecting Complex Predicates in Hindi Using POS Projection Across Parallel Corpora, in: Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, MWE '06. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 28–35.

[12] Nerbonne, J., Paskaleva, E., Karttunen, L., Proszeky, G., Roosmaa, T., 1997. Reading more into foreign languages, in: Proceedings of the Fifth Conference on Applied Natural Language Processing. Association for Computational Linguistics, pp. 135–138.

[13] Maamouri, M., Zaghouani, W., Cavalli-sforza, V., Graff, D., Ciul, M., n.d. Developing ARET: An NLP-based Educational Tool Set for Arabic Reading Enhancement.

[14] Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., Ott, N., Tübingen, U., n.d. 2010. Enhancing Authentic Web Pages for Language Learners.

[15] Sag, I. A. and Baldwin, Timothy and Bond, Francis and Copestake, Ann and Flickinger, Dan. . Multi-word expressions: a pain in the neck for nlp . Proceedings of CICLing , 2002 .

[16] Sanjay Kumar Dwivedi & Pramod Premdas Sukhadeve, (2010) "Machine Translation System in Indian Perspectives", Journal of Computer Science6 (10): 1082-1087, ISSN 1549-3636, © 2010 Science.

[17] Sinha, R.M.K., 2009. Mining Complex Predicates in Hindi Using a Parallel Hindi-English Corpus, in: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, MWE '09. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 40–46.

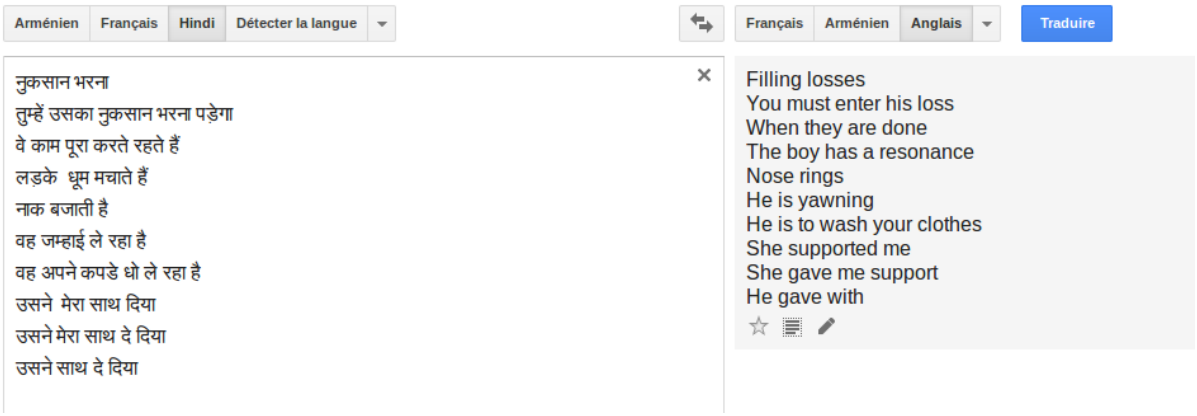
[18] Sitender & Seema Bawa, (2012) "Survey of Indian Machine Translation Systems", International Journal Computer Science and Technolgy, Vol. 3, Issue 1, pp. 286-290.

[19] Yasuda, S., 2010. Learning Phrasal Verbs Through Conceptual Metaphors: A Case of Japanese EFL Learners. TESOL Quarterly 44, 250–273. doi:10.5054/tq.2010.219945

9 Annexes

Annexe 1

La traduction de certains prédicats complexes par Google Traduction



Annexe 2

La translittération de la devanagari selon IAST

Voyelles		Consonnes					Autres consonnes
अ a	आ ā	क ka	ख kha	ग ga	घ gha	ङ ṅa	क़ qa
उ u	ऊ ū	च ca	छ cha	ज ja	झ jha	ञ ña	ख़ kha
इ i	ई ī	ट ṭa	ठ ṭha	ड ḍa	ढ ḍha	ण ṇa	ग़ ġa
ऋ ṛ		त ta	थ tha	द da	ध dha	न na	फ़ fa
ए e	ऐ ai	प pa	फ pha	ब ba	भ bha	म ma	ज़ za
ओ o	औ au	य ya	र ra	ल la	व va		ड़ ṛa
अं ṃ	अः ḥ	ष ṣa	श śa	स sa	ह ha		ढ़ ṛha

Annexe 3

Les étiquettes définies par A. Bharati :

NN	Nom
NNP	Nom propre
PRP	Pronom
DEM	Démonstratif
VM	Verbe principal
VAUX	Verbe auxiliaire
JJ	Adjectif
RB	Adverbe
PSP	Postposition
RP	Particule
CC	Conjonction
WQ	Interrogatif
QF	Quantifieur
QC	Cardinal
QO	Ordinal
INJ	Interjection
NEG	Négation
SYM	Ponctuation

Annexe 4

Le tableau ci-dessous présente la formation des temps/aspects/modes qui sont identiques pour les verbes réguliers et irréguliers (B désigne la base nue).

TAM	**Formation**	**Négation**
Présent général	Participe 1 + être.PRES	nahīṃ
Imparfait général	Participe 1 + être.IMPRF	nahīṃ
Irréel	Base nue+ t+désinences adjectivales	nahīṃ
Présent actualisé	Base nue+ rahā+être.PRES	nahīṃ
Imparfait actualisé	Base nue+ rahā+être.IMPRF	nahīṃ
Duratif	Participe 1+rahnā	nahīṃ
Duratif-progressif	Participe 1+ānā, calā jānā, calnā, ho jānā	nahīṃ
Inceptif	Infinitif+lagnā	nahīṃ, mat, na
Terminatif	Infinitif+cuknā	nahīṃ
Déontique	Infinitif+cāhie, padnā	nahīṃ, na
Permissif	Infinitif+denā	nahīṃ, mat, na
Acquisitif	Infinitif+pānā	nahīṃ, mat, na
Potentiel	Base nue+saknā, pānā	nahīṃ, na
Absolutif	Base nue+kar/ke	

Le tableau ci-dessous présente les temps dont la formation est différente pour certains verbes.

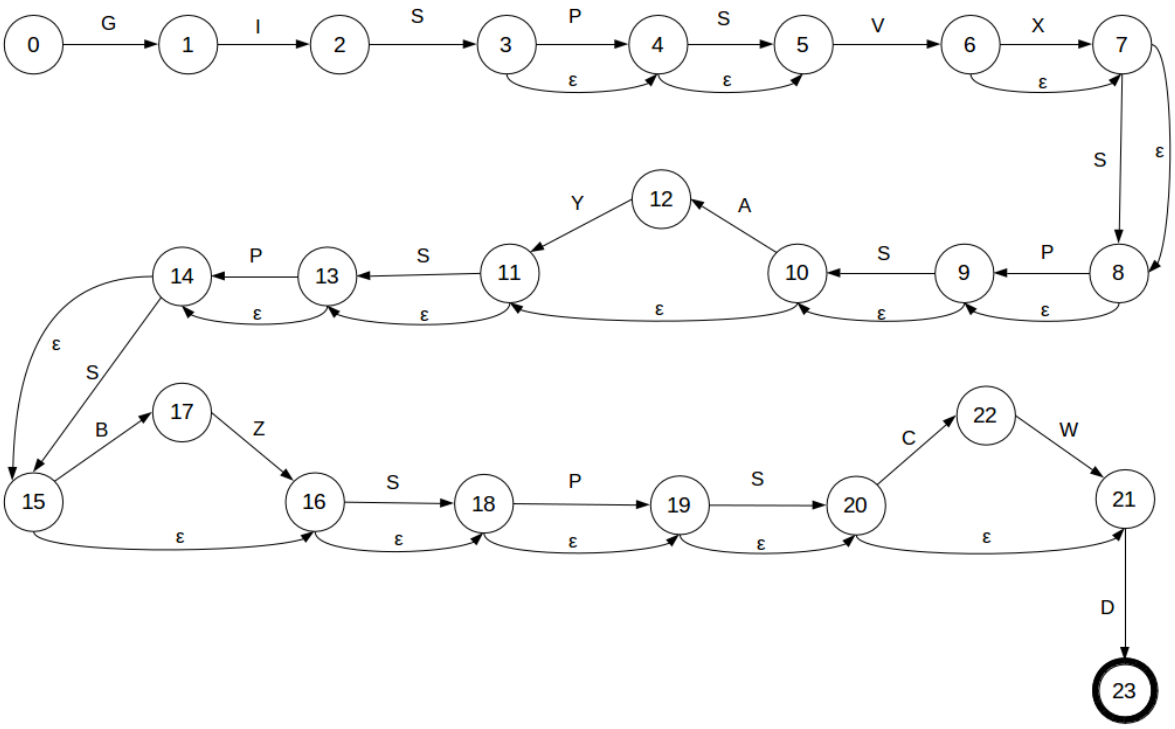
TAM	**Formation**	**Negation**
Impératif	Base nue + marques de personne	mat, na
Subjonctif	Base nue + marques de personne	na
Futur	Base du subjonctif + g+désinences adjectivales	nahīṃ
Aoriste	Base nue+ dés. adj	nahīṃ
Perfectif composé 1	Participe 2 + être.PRES	nahīṃ
Perfectif composé 2	Participe 2 + être.IMPRF	nahīṃ
Subjonctif passé	Participe 2+ être.SUBJ	nahīṃ
Irréel passé	Participe 2+être. IRR	nahīṃ
Fréquentatif	Participe 2+karnā	nahīṃ
Passif	Participe 2+jānā	nahīṃ

Le paradigme de conjugaison des verbes irréguliers

	honā	karnā	lenā	denā	jānā
Impératif	B(o|ūjie)	k(r|ro|īje|ījiegā|ījiye|ījiyegā)	l(e|o|īje|ījiegā|ījiye|ījiyegā)	d(e|o|īje|ījiegā|ījiye|ījiyegā)	B(o|ie|iegā)?
Subjonctif	B(ūṃ|ṃ) ?	B(ūṃ|o|e|eṃ)	l(ūṃ|o|e|eṃ)	d(ūṃ|o|e|eṃ)	B(ūṃ|o|e|eṃ)
Futur	B(ūṃ|o|oṃ]g(ā|ī|e)	B(ūṃ|o|e|eṃ)g(ā|ī|e)	l(ūṃ|o|e|eṃ)g(ā|ī|e)	d(ūṃ|o|e|eṃ)g(ā|ī|e)	B(ūṃ|o|e|eṃ)g(ā|ī|e)
Aoriste	hu(ā|ī|yī|e|ye|īṃ|yīṃ)	k(iyā|ī|ie|iye|īṃ)	l(iyā|ī|ie|iye|īṃ)	d(iyā|ī|ie|iye|īṃ)	g(yā|ī|yī|e|ye|īṃ|yīṃ)
Perfectif composé1	hu(ā|ī|yī|e|ye)	k(iyā|ī|ie|iye|īṃ)	l(iyā|ī|ie|iye)	d(iyā|ī|ie|iye|īṃ)	g(yā|ī|yī |e|ye)
Perfectif composé 2	hu(ā|ī|yī|e|ye|īṃ|yīṃ)	k(iyā|ī|ie|iye|īṃ)	l(iyā|ī|ie|iye)	d(iyā|ī|ie|iye|īṃ)	g(yā|ī|yī |e|ye)
Passif		k(iyā|ī|ie|iye)	l(iyā|ī|ie|iye)	d(iyā|ī|ie|iye)	ja(yā|ī|yī|e|ye)
Fréquentatif	huā	kiyā	liyā	diyā	jayā

Annexe 5

L'automate associé à l'expression régulière :



où G=$fr_gauche, I=$inv, S=espace, P=$particules, V=$var, X=$classe0, A=$A, Y=$classe1, B=$B, Z=$classe2, C=$C, W=$classe3, D=$fr_droite

Annexe 6

Le contenu des variables $A, $B et $C :

$A	$B	$C
वाल	थ	ह
रह	हो	हो
थ	ह	थ
ह	रह	
हो	चाहि	
पड	पड़	
चाहिये		
चाहिए		
चुक		
कर		
लग		
पा		
सक		

आ चल आ चले जा चली जा चला जा चले ग चली ग चला ग चल जा ग द उठ पड़ डाल ल बैठ मार		

Le contenu des variables $classe0, $classe1, $classe 2 et $classe3 :

$classe0	$classe1	$classe2	$classe3
ियेगा	ीजिएगा	ियेगा	ऊँगा
इयेगा	ीजियेगा	ूँगा	ऊँगी
येंगे	येंगे	ूँगी	ूंगा
येंगी	येंगी	ूंगा	ूंगी
ूँगा	ीजिये	ूंगी	ंगी
ूँगी	ियेगा	िएगा	ंगे
ूंगा	इयेगा	ेंगे	ुईं
ूंगी	ीजिए	ेंगी	ऊँ
योगे	िएगा	ऊँगा	ूं
योगी	इएगा	ऊँगी	गा
ेंगी	ूँगा	ूंगा	गी
ऊँगा	ूँगी	ूंगी	गे
ऊँगी	ूंगा	ोगे	ूँ
ूंगा	ूंगी	ोगी	ूं

ऊंगी	येगी	िये	ैं
येगी	ेंगे	ेगा	ुआ
ेंगे	ेंगी	ेगी	ुए
येगा	येगा	ंगी	ुई
िएगा	ऊँगा	ंगे	ीं
इएगा	ऊँगी	ुई	ै
ोगी	ऊंगा	तीं	ं
कर	ऊंगी	ता	ा
ओगी	योगे	ते	े
ोगे	योगी	ती	ी
ओगे	यीं	िए	
ेगा	ेगा	ूँ	
ेगी	ंगी	ूं	
ंगी	ंगे	ैं	
ंगे	िये	ुआ	
िये	इये	ुए	
इये	ेगा	ुई	
यीं	ेगी	ऊँ	
तीं	ोगे	ऊं	
यें	ओगे	ये	
ुईं	ोगी	गा	
ेंं	ओगी	गी	
ीं	िया	गे	
या	ुईं	तीं	
ये	यें	ता	
एं	िये	ते	
एँ	तीं	ती	
यी	ेना	ीं	
ूँ	ेनी	ेंं	
ूं	ेने	ं	
ैं	ता	ो	
ुआ	ते	ी	
ुए	ती	ा	
ुई	कर	ै	

ऊँ ऊं गा गी गे िए इए ता ते ती कर के ना ने नी ईं ा े ो ओ ै ं ई ी ए	ना ने नी िए ूँ ूं ऊँ ऊं गा गी गे ये एं एँ यी ैं ुआ ुए ेंं ीं ईं या िए इए ुई ो ओ ा े ं ई ए ी ै	ए े	

10 Liste des abréviations

Adj	adjectif
Aux	auxiliaire
D	déterminant
GEN	génitif
IMP	impératif
IMPRF	imparfait
Inf	infinitif
IRR	irréel
LV	locution verbale
N	nom
P1	participe inaccompli
P2	participe accompli
PC	prédicat complexe
SN	syntagme nominal
SOC	sociatif
SOV	sujet objet verbe
SV	syntagme verbal
V	verbe

11 Glossaire de termes grammaticaux français-hindi

Verbes conjoints	संयुक्त क्रियाएं
Absolutif	पूर्वकालिक कृदन्त
Adjectif	विषेशण
Adverbe	क्रिया विषेशण
Aoriste	सामान्य भूतकाल
Déontique	अनिवार्यता बोधक
Futur	सामान्य भविष्यत्
Imparfait actualisé	सातत्यता बोधक पूर्ण भूतकाल
Imparfait général	अपूर्ण भूतकाल
Impératif	विधिकाल
Infinitif	सामान्य क्रिया
Irréel	सामान्य संकेतार्थ
Locutions verbales	नाम धातु क्रियाएं
Participe accompli	पूर्ण क्रियाद्योतक कृदन्त
Participe inaccompli	अपूर्ण क्रियाद्योतक कृदन्त
Perfectif composé 1	पूर्ण वर्तमान काल
Perfectif composé 2	पूर्ण भूतकाल
Permissif	अनुमती बोधक
Présent actualisé	सातत्यता बोधक पूर्ण वर्तमान
Présent général	सामान्य वर्तमान
Subjonctif	संभाव्य भविष्यत्
Verbe	क्रिया
Verbe intransitif	अकर्मक क्रिया
Verbe transitif	सकर्मक क्रिया
Verbes causatifs	प्रेरणार्थक क्रियाएं
Voix passive	कर्मवाच्य
Présomptif	संभावनाबोधक
Inceptif	आरम्भ बोधक

12 Liste de prédicats complexes

अंदाजा लगाना	estimer
अनसुना करना	ignorer
अपमान करना	injurier
अभ्यस्त होना	être habitué
अमल में लाना	appliquer
अरेस्ट करना	arrêter
अरेस्ट होना	être arrêté
आज्ञा पालन करना	obéir
उपभोग करना	consommer
उपयोग करना	utiliser
उपयोग होना	être utilisé
उपस्थित होना	être présent
उपेक्षित करना	rejeter
कंपित होना	trembler
कदम उठाना	prendre des mesures
कसम खाना	jurer
कसरत करना	s'exercer
क़ाबू में लाना	contrôler
किडनैप करना	kidnapper
क्षमा करना	pardonner
खत्म होना	se terminer
खयाल रखना	prendre soin
ख़ून करना	tuer
ख़ुश करना	rendre heureux
खाली करना	vider

गंदा करना	salir
गड़बड़ करना	perturber
गर्मी लगना	avoir chaud
गाली देना	insulter
गिरफ्त में लेना	mettre en garde à vue
गिरफ्तार करना	arrêter
गुस्सा करना	mettre en colère
गुस्सा दिलाना	mettre en colère
गोद लेना	adopter
चाँटा मारना	gifler
चाक करना	détruire
चाबी भरना	enfermer
चुसकी लेना	prendre une gorgée
चेक भुनाना	encaisser un chèque
छलांग मारना	sauter
छापा मारना	attaquer
छेड़छाड़ करना	harceler
जंग लगना	rouiller
जबरदस्ती करना	forcer
जमा कराना	ramasser, réunir
जमा होना	être réuni
जरूरत पड़ना	avoir besoin
जवाब देना	répondre
जान लेना	tuer
जुरमाना लगाना	verbaliser

झाड़ू लगाना	balayer
झूठ बोलना	mentir
टालमटोल करना	trafiquer
टेक ऑफ करना	décoller
ठीक करना	corriger
डींग हाँकना	se vanter
ढेर करना	empiler
तंग आना	s'ennuyer
तय करना	décider
तर्क वितर्क करना	discuter
तस्वीर खींचना	prendre une photo
तैयार करना	préparer
थाह लेना	mesurer
थ्रो करना	lancer
दफन करना	enterrer
दर्ज करना	enregistrer
दस्तखत करना	signer
धूम्रपान करना	fumer
नकल करना	imiter
नजर आना	sembler
नफ़रत होना	détester
निंदा करना	blâmer
नुक़सान भरना	dédommager
परामर्श देना	conseiller
परिश्रम करना	faire du zèle
पसन्द आना	plaire
फर्क करना	distinguer
फलाँग मारना	sauter

फ़ायदा उठाना	tirer un profit
फ़ैसला करना	prendre une décision
बंद करना	fermer
बयान करना	faire une déclaration
बरदाश्त करना	supporter
ब्याह करना	se marier
भरोसा दिलाना	assurer
भाग करना	diviser
भेद करना	différencier
मना करना	interdire
मनोरंजन करना	s'amuser
महसूस करना	sentir
माँग करना	demander
मौज मस्ती करना	s'amuser
यत्न करना	essayer
यात्रा करना	voyager
याद करना	se souvenir
रद्द करना	annuler
रजिस्टर कराना	s'inscrire
राज़ी होना	être d'accord
रुचि होना	être intéressé
लंबा करना	allonger
लज्जित करना	intimider
लड़ाई करना	se battre
लाभ उठाना	profiter
वचन देना	faire une promesse
वर्णन करना	décrire
वादा करना	promettre

विकास करना	développer
विश्लेषण करना	analyser
विश्वास करना	faire confiance
शक करना	douter
शांत करना	apaiser
शादी करना	se marier
शिरकत करना	participer
शुरू करना	commencer
शोर मचाना	faire du bruit
संतुष्ट करना	satisfaire
संकोच करना	hésiter
संबोधित करना	s'adresser

सचेत करना	informer
सज़ा देना	punir
सज्जित करना	équiper
सफल होना	réussir
समाप्त करना	finir
सलाह देना	conseiller
हल निकलना	résoudre
हस्तक्षेप करना	interrompre
हस्ताक्षर करना	signer
हासिल करना	obtenir
हैरान करना	étonner