Enchaînements verbaux – étude sur le temps et l'aspect utilisant des techniques d'apprentissage non supervisé

Catherine RECANATI, Nicoleta ROGOVSCHI
LIPN – UMR 7030 du CNRS, Institut Galilée, Université Paris 13
99, avenue J-B. Clément, 93430 Villetaneuse, France
{Catherine.Recanati,Nicoleta.Rogovschi}@lipn.univ-paris13.fr

Résumé. L'apprentissage non supervisé permet la découverte de catégories initialement inconnues. Les techniques actuelles permettent d'explorer des séquences de phénomènes alors qu'on a tendance à se focaliser sur l'analyse de phénomènes isolés ou sur la relation entre deux phénomènes. Elles offrent ainsi de précieux outils pour l'analyse de données organisées en séquences, et en particulier, pour la découverte de structures textuelles. Nous présentons ici les résultats d'une première tentative de les utiliser pour inspecter les suites de verbes provenant de phrases de récits d'accident de la route. Les verbes étaient encodés comme paires (*cat*, *temps*), où *cat* représente la catégorie aspectuelle d'un verbe, et *temps* son temps grammatical. L'analyse, basée sur une approche originale, a fourni une classification des enchaînements de deux verbes successifs en quatre groupes permettant de segmenter les textes. Nous donnons ici une interprétation de ces groupes à partir de statistiques sur des annotations sémantiques indépendantes.

Abstract. Unsupervised learning allows the discovery of initially unknown categories. Current techniques make it possible to explore sequences of phenomena whereas one tends to focus on the analysis of isolated phenomena or on the relation between two phenomena. They offer thus invaluable tools for the analysis of sequential data, and in particular, for the discovery of textual structures. We report here the results of a first attempt at using them for inspecting sequences of verbs coming from sentences of French accounts of road accidents. Verbs were encoded as pairs (*cat*, *tense*) – where *cat* is the aspectual category of a verb, and *tense* its grammatical tense. The analysis, based on an original approach, provided a classification of the links between two successive verbs into four distinct groups (clusters) allowing texts segmentation. We give here an interpretation of these clusters by using statistics on semantic annotations independent of the training process.

Mots-clés : temps, aspect, sémantique, apprentissage non supervisé, fouille de données.

Keywords: time, tense, aspect, semantics, unsupervised learning, data mining.

1 Introduction

L'intérêt de l'apprentissage *non supervisé* est qu'il permet la découverte de catégories initialement inconnues. Les techniques actuelles permettent d'explorer des séquences de

phénomènes alors qu'on a tendance à se focaliser sur l'analyse de phénomènes isolés ou sur la relation entre deux phénomènes. Elles offrent ainsi de précieux outils pour l'analyse de données organisées en séquences, et en particulier, pour la découverte de structures textuelles. Notre objectif est de les utiliser à cette fin et nous présentons ici l'interprétation des résultats d'une première tentative.

De nombreuses études ont montré l'importance des temps dans la structure narrative d'un récit (Vuillaume, 1990). L'opposition entre le passé simple et l'imparfait a en particulier fait couler beaucoup d'encre. Mais de nombreux liens unissant le temps et l'aspect, l'idée d'un couplage temps et catégorie aspectuelle nous a paru naturelle et intéressante. Il a en effet été démontré qu'on ne peut effectuer l'analyse temporelle des suites d'événements à partir du seul temps grammatical sans faire intervenir l'aspect (voir Kamp H., Vet C. ou Vlach F. dans (Martin et Nef, 1981), (Vet, 1994), (Gosselin, 1996), etc.). On trouve aussi dans la littérature des liens entre l'aspect et d'autres phénomènes sémantiques, comme l'intentionnalité ou la causalité. Dans cette première étude, nous avons cherché à voir si l'on pouvait détecter une certaine régularité dans les enchaînements de verbes au sein des phrases en couplant temps et catégorie aspectuelle des verbes, et s'il était possible, dans un cadre restreint, de leur attribuer un « sens ».

Les récits qui ont été analysés sont des récits d'accident de la route provenant de la partie « observations » d'un constat à l'amiable destiné aux assureurs. Ils nous ont gracieusement été fournis par la MAIF, que nous remercions. Assez courts, leur intérêt *a priori* est de montrer comment s'exprime un accident, ses causes et la responsabilité de ses auteurs, dans un espace limité. Ces textes ont déjà été utilisés au LIPN pour des travaux sur les inférences causales (Nouioua et Kayser, 2006). Notre approche est néanmoins différente puisqu'il s'agit d'une approche statistique visant à catégoriser des enchaînements de verbes – le pari étant que, si de telles classes d'enchaînements existent, elles aient globalement un sens, à tout le moins pour le type de récit considéré.

Nous avons pleinement conscience de la difficulté concernant l'évaluation de ce premier travail, ce dernier étant basé sur le postulat que de tels enchaînements (ici relativement pauvres du point de vue syntaxico-sémantique) puissent se voir attribuer un sens, et ne pas être le reflet de statistiques contingentes. Mais le peu de ressources utilisées est un précieux avantage pour de futures applications au TAL, et l'expérience méritait donc d'être menée. Ajoutons que les outils mathématiques dont nous disposons nous ont permis de tester la validité des catégories obtenues du point de vue statistique, et que notre analyse sémantique a été effectuée à partir d'annotations complètement indépendantes de l'apprentissage.

1.1 Intérêts de notre approche formelle

Les SOM (Self Organizing Map) ou cartes topologiques de Kohonen (Kohonen, 1995) permettent un apprentissage non supervisé (*clustering*) efficace avec visualisation simultanée des résultats de la classification. Cette visualisation se fait grâce à la carte topologique des données (deux données similaires sont proches sur la carte) qui fournit en même temps un codage "intelligent" des données sous forme de prototypes. Ces prototypes étant de même nature que les données, ils sont interprétables, et la carte fournit ainsi un résumé des données. A partir de ce codage, nous avons pris les HMM (Hidden Markov Models) pour modéliser la dynamique des séquences de données (ici, les suites de verbes). Les HMM **(**Rabiner et Juang, 1986) sont la meilleure approche pour traiter des séquences de longueur variable et capturer leur *dynamique*. C'est pourquoi ces modèles ont été largement utilisés dans le domaine de la

reconnaissance de la parole et sont tout particulièrement adaptés à notre objectif. Pour la validation de notre approche, nous avons utilisé à la fois des données génétiques et des données textuelles (celles dont nous présentons ici l'interprétation). Pour plus de détails techniques sur cette méthode, voir (Rogovschi, 2006) ou (Rogovschi et al., 2006).

1.2 Codage des phrases

L'analyse a été réalisée sur une centaine de textes correspondant à 700 occurrences de verbes. On a considéré dans ces récits toutes les séquences de verbes délimitées par des phrases d'au moins deux verbes. Pour palier au faible nombre de données, nous avons utilisé des techniques de ré-échantillonnage basées sur des fenêtres glissantes permettant d'augmenter la redondance (la redondance assure une meilleure classification des données). Pour le codage, les quatre catégories aspectuelles de verbes (état, activité, accomplissement, achèvement) originellement dues à Vendler et Kenny (Vendler, 1967) ont été couplées avec le temps grammatical. Le Tableau 1 résume sommairement les différences entre ces quatre catégories sémantiques. L'indexation a été réalisée à la main en s'appuyant sur notre conception de ces catégories (Recanati C., Recanati F., 1999).

ETAT homogène, duratif, habituel ou indiquant une disposition *être (à l'arrêt) / vouloir / pouvoir*	**ACTIVITE** processus relativement homogène, non borné *rouler / circuler / slalomer / suivre*
ACCOMPLISSEMENT processus dirigé et borné par une fin *traverser / faire un créneau / aller à*	**ACHEVEMENT** événement quasi ponctuel *franchir / heurter / percuter*

Tableau 1 : Les quatre catégories aspectuelles de verbes

Ce type de récit n'utilise globalement que l'imparfait (24%) et le passé composé (34%), avec de temps à autre quelques phrases au présent. On y trouve aussi quelques occurrences (rares) de passé simple et de plus-que-parfait. Il existe par contre un nombre important de participes présents (11%) et d'infinitifs (20%). Nous avons donc décidé de les retenir, bien qu'ils ne participent pas de la même manière à l'ossature grammaticale. Nous avons ainsi effectué l'apprentissage en gardant 9 codes[1] pour les temps apparus sur les verbes.

Exemple « Le véhicule B *circulait* sur la voie de gauche des véhicules *allant* à gauche (marquage au sol par des flèches). Celui-ci *s'est rabattu* sur mon véhicule, me *heurtant* à l'arrière. Il *a accroché* mon pare-choc et m*'a entraîné* vers le mur amovible du pont de Gennevilliers que j*'ai percuté* violemment. » sera réduit après codage aux suites de verbes apparus dans les phrases : (circulait, allant) / (s'est rabattu, heurtant)/ (a accroché, a entraîné, ai percuté) – lesquelles ont encore été encodées numériquement comme séquences de couples (temps, catégorie), soit ici : (act., IM) (acco., pp) / (acco., PC) (achè., pp) / (achè., PC) (acco., PC) (achè., PC).

[1] IM = imparfait, PR = présent, PC = passé composé, PS = passé simple, PQP = plus-que-parfait, inf = infinitif, ppr = participe présent, pp = participe passé et pps = participe passé surcomposé. En ce qui concerne les participes passés (peu nombreux ici), nous n'avons pas toujours compté les emplois adjectivaux contingents, et du point de vue du catégorisateur, ils constituent plutôt du bruit.

2 Premiers résultats

Les premiers résultats concernent les statistiques sur les verbes et les catégories, indépendamment de leurs enchaînements. Dans ce type de récit, les verbes d'état représentent 24% du corpus, ceux d'activité seulement 10%, et l'on trouve 34% de verbes d'accomplissement et 32% de verbes d'achèvement. La répartition non uniforme des temps sur les catégories confirme l'intérêt de notre couplage temps/catégorie aspectuelle (cf. Figure 1).

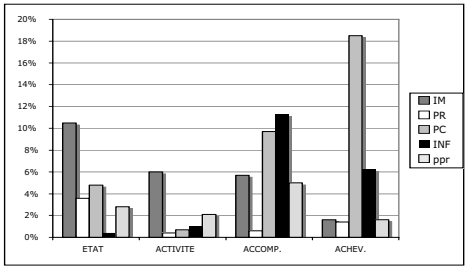

Figure 1 : Répartition des temps par catégorie

Ces pourcentages et leur répartition s'expliquent assez naturellement par la nature de chaque catégorie aspectuelle, la structure généralement typique de ces récits, et la spécialisation aspectuelle des temps grammaticaux (opposition perfectif/imperfectif).

Les verbes d'états (24%) sont répartis à plus de 70% sur l'imparfait, le présent et le participe présent. Cela n'est guère surprenant puisque les états sont homogènes, souvent duratifs ou caractérisant une aptitude (habituels, génériques). La proportion néanmoins non négligeable du passé composé s'explique sans doute par la fréquence de verbes comme « vouloir » ou « pouvoir », classés comme verbes d'états du fait de leur aspect dispositionnel (« j'ai voulu freiner », « je n'ai pu éviter »). La faible proportion de présent provient du fait que le récit est au passé et que le présent historique est trop littéraire pour le genre.

Les verbes d'activités (10%) dénotant des processus homogènes et non bornés, ils se répartissent tout naturellement à plus de 79% sur l'imparfait et le participe présent. La présence de 10% d'infinitif peut facilement s'expliquer par le fait qu'il s'agit de processus qui ont un début, et qui peuvent donc se trouver complément de verbe comme « commencer à », « vouloir », ou être introduits pour mentionner un but par la préposition « pour ».

Les accomplissements (34%) et les achèvements (32%) sont très fréquents au passé composé du fait de leur caractère télique (borné par une fin). Les achèvements sont présents de manière massive car, étant ponctuels ou de courte durée, ils supportent mal l'imparfait. A l'inverse, les accomplissements supportent bien l'imparfait et le participe présent, parce qu'ils ont une durée intrinsèque, et mettent l'accent sur le procès plutôt que sur sa fin – ce qui les rapproche finalement des activités. L'importance globale de ces deux catégories est sans doute liée à la typologie des textes analysés, un récit d'accident impliquant de décrire la séquence des événements successifs qui l'ont provoqué.

Distinction perfectif/imperfectif. Il y a trois points de vue dans le système aspectuel du Français (Smith, 1991). Un point de vue perfectif présente les situations comme fermées, y

compris les états (le point final est alors un changement d'état). Le perfectif s'exprime par le passé composé et le passé simple. Les points de vue imperfectif et neutre présentent à l'inverse des situations ouvertes. Le point de vue neutre s'exprime par le présent. Le point de vue imperfectif s'exprime par l'imparfait, ou par la forme *en train de.* Paul J. Hopper (Hopper, 1979) a fort bien décrit les caractéristiques de ces deux modes vis-à-vis de la structure narrative, du focus et de l'aspect (cf. Tableau 2). Ici, l'opposition perfectif/imperfectif sera réalisée par l'opposition passé composé/imparfait. Cette opposition est néanmoins globale et l'on aurait tort d'attribuer de manière systématique un aspect imperfectif à tous les imparfaits.

Perfectif	**Imperfectif**
Chronologie stricte	Simultanéité ou recouvrement
Vue d'un événement comme un tout, dont la fin est une condition préalable pour l'événement suivant	Vue de la situation ou de ce qui arrive sans que la fin soit nécessaire pour ce qui va arriver ensuite
Identité du sujet à l'intérieur de chaque épisode discret	Changement fréquent de sujet
Topicalisation humaine	Variété de topiques, y compris des phénomènes naturels
Focus non marqué dans les clauses, avec présupposition du sujet et assertion dans le verbe et ses compléments immédiats	Distribution marquée du focus, sur le sujet, l'instrument, ou un adverbe
Evénements dynamiques et cinématiques	Situation descriptive, statique
Avant-plan. Evénement indispensable à la structure narrative	Arrière-plan. Etat ou situation nécessaire pour comprendre les motifs, les attitudes, etc.

Tableau 2 : Opposition perfectif/imperfectif

Structure typique. Un récit d'accident commence généralement par quelques phrases décrivant les circonstances et les états de choses précédant l'accident. Cette première partie est alors à l'imparfait, et contient de nombreux participes présents. On y trouve aussi quelques présents et de nombreux infinitifs introduits par « pour », ou compléments de verbes (« je m'apprêtais à tourner », « le feu venait de passer au rouge »). Essentiellement circonstancielle, cette partie contient une majorité de verbes d'états, et quelques activités et accomplissements. Elle est globalement caractérisée par un point de vue imperfectif, et le récit est en arrière-plan. Vient ensuite la description de l'accident proprement dit, qui mentionne la suite des événements ayant conduit à l'accident pour finir par le choc. Cette partie utilise massivement des verbes d'accomplissement et d'achèvement, le plus souvent au passé composé. Elle est caractérisée par un mode perfectif, mais le but du jeu étant d'indiquer les responsabilités des auteurs, on trouve ici encore beaucoup de participes présents et de tournures infinitives enchaînant souvent trois verbes (« J'ai voulu freiner pour l'éviter », « voulant éviter la borne, je n'ai pu »). En fin de récit, on trouve parfois une troisième partie constituée de commentaires et inventoriant notamment les dégâts. Cette partie est relativement courte et plus difficile à caractériser stylistiquement.

3 Catégorisation des enchaînements verbaux

Notre approche non supervisée a fourni une classification des paires de deux verbes successifs (au sein d'une même phrase) en quatre groupes. Il faut souligner que ce nombre de

quatre, particulièrement petit, est intéressant car il atteste du bien fondé de notre couplage temps/aspect (pas d'explosion combinatoire). Rappelons qu'avec 9 temps et 4 catégories on obtient 36 sortes de verbes, soit 1296 couples virtuels. La répartition des temps sur les catégories a en effet restreint le nombre de couples. Mais c'est aussi la capacité de réduction de la méthode qui permet d'obtenir ce résultat. Une première réduction de dimension a été effectuée par les cartes SOM, qui représentent ici les classes avec 36 paires, puis un élagage de la carte, effectué à partir des matrices de probabilités de transitions provenant des HMM, a réduit encore ces classes à un plus faible nombre de paires typiques.

Annotations sémantiques. Pour faciliter l'interprétation des classes obtenues, nous avions effectué par avance un certain nombre d'annotations. Ces annotations n'ont pas été utilisées pour l'apprentissage mais elles vont nous permettre de caractériser sémantiquement les classes de transitions verbales de façon globale. Ainsi, pour rendre compte de la structure typique de ces récits, nous avons indexé tous les verbes d'un numéro indiquant la « partie » thématique (1-*circonstance*, 2-*accident* ou 3-*commentaire*). Nous avons également marqué certains verbes des attributs *foreground* ou *background* pour indiquer que le récit est en avant-plan ou en arrière-plan. Pour déceler d'éventuelles chaînes causales conduisant à l'accident, nous avons marqué les verbes des attributs *causal* ou *choc* quand le verbe indiquait une cause directe de l'accident, ou le choc lui-même. Nous avons également marqué les verbes d'action en fonction de l'agent (*A* pour le conducteur auteur du récit, *B* pour « l'adversaire », et *C* pour un tiers). Nous avons aussi noté la présence de négation, et l'évocation plus générale de buts poursuivis ou de mondes possibles voisins qui ne se sont pas produits (attributs *négation* et *inertia*). Le Tableau 3 résume symboliquement les résultats que nous avons obtenus en faisant des statistiques sur nos marques sémantiques dans ces quatre classes. Le marquage de la négation s'est avéré peu discriminant, et celui des agents relativement peu informatif.

Groupe C (circonstances)	**Groupe CI (circonstances ou incident)**
Pas de causalité, arrière-plan, pas de choc, nombreux buts et alternatives *(étais, tournant) (reculais, repartir)*	Peu de causalité, arrière-plan, peu de choc, ni but ou alternative *(démarrais, ai entendu) (avait, trouvais)*
Groupe AA (actions menant à l'accident)	**Groupe CC (choc ou commentaires)**
Causalité forte, relief neutre, quelques chocs, nombreux buts et alternatives *(ai voulu, engager) (a percuté, abîmant)*	Causalité très forte, en avant-plan, choc fréquent, buts et alternatives *(a accroché, a entraîné) (n'a pu, stopper)*

Tableau 3 : Synthèse sur les annotations sémantiques

3.1 Groupe C des circonstances

Ce groupe se distingue par une nette différenciation du premier verbe et du second. Le premier verbe est à 93% à l'imparfait, pour seulement 7% de présent, tandis que le second est à 63% à l'infinitif et à 30% au participe présent. Du point de vue des catégories aspectuelles, le premier verbe est à 56% un verbe d'état, et le second à 63% un verbe d'accomplissement (les autres catégories se trouvant distribuées de manière régulière entre 12% et 16%). On pourrait résumer globalement les transitions de ce groupe comme présentant un verbe d'état (ou d'activité) à l'imparfait, suivi d'un verbe d'accomplissement à l'infinitif ou au participe

présent. Le Tableau 4 (voir plus loin) nous donne une synthèse plus fine. Ce groupe privilégie les états et les activités au détriment des achèvements – et les accomplissements sont massivement représentés en seconde occurrence verbale. Ce groupe est celui où l'attribut *inertia* (indiquant un but ou un monde possible proche) est le plus important. Cela s'explique par les nombreux accomplissements introduits par la préposition « pour » (« je *reculais* pour *repartir* », « je *sortais* du parking pour me *diriger* ») ou les auxiliaires à l'imparfait introduisant un infinitif et indiquant des intentions du conducteur (« je m*'apprêtais* à *tourner* à gauche »). C'est une des raisons pour laquelle nous l'avons baptisé groupe C des circonstances. L'autre raison est que ce groupe contient une majorité de verbes appartenant à la première partie du récit (63%), et peu de la seconde et la troisième. On constate en outre que ce groupe ne contient pratiquement aucun verbe indiquant les causes de l'accident ou le choc. L'acteur A y est le plus présent, et le récit est en arrière-plan.

3.2 Groupe CI des circonstances ou de la survenue d'un incident

Le groupe CI est celui des circonstances ou de la survenue d'un incident. On note ici un grand nombre de verbes d'états (37,5%) et d'activités (17%), encore plus important que dans le groupe précédent et très supérieur à la moyenne. On a par contre un nombre moyen d'achèvements (29%), absents du premier verbe mais massivement représentés en second. Cela distingue ce groupe du précédent, où les accomplissements jouaient ce rôle. Ici, à l'inverse, les accomplissements sont exclus de la seconde place et nettement sous représentés (16,5%). On peut synthétiser les enchaînements de ce groupe en disant qu'on a généralement affaire à un état ou une activité à l'imparfait, suivi d'un état ou d'un achèvement, à l'imparfait ou au passé composé. On enchaîne donc deux points de vue imperfectifs, et parfois, un point de vue imperfectif et un point de vue perfectif. On a en effet 36% des verbes qui proviennent de la partie circonstancielle (« Je *circulais* à environ 45 Km/h dans une petite rue à sens unique où *stationnaient* des voitures de chaque côté »), mais également des séquences finissant par un achèvement au passé composé, provenant de la seconde (34%, « Je *roulais* dans la rue Pasteur quand une voiture *a surgi* de ma droite »). Ce groupe contient en outre 25% de séquences verbales situées à cheval entre les deux parties, soit environ la moitié de ces dernières. C'est la raison pour laquelle nous l'avons baptisé « groupe des circonstances ou de la survenue d'un incident ». Le récit est principalement en arrière-plan. L'acteur A (ou un tiers C) se trouvent fortement représentés au détriment de l'acteur B. Il y a peu d'allusions aux causes de l'accident et au choc. L'évocation de buts ou d'alternatives y est insignifiante.

3.3 Groupe AA des actions menant à l'accident

Les verbes du groupe AA proviennent essentiellement du récit de l'accident proprement dit. Ce groupe est caractérisé par l'abondance des accomplissements, au détriment des états et des activités, et c'est pourquoi nous l'avons baptisé « groupe des actions menant à l'accident ». Le mode est généralement perfectif mais on y trouve aussi beaucoup d'infinitifs. Les éléments les plus typiques sont listés sur le Tableau 4. Ces enchaînements se prêtent à des constructions de trois verbes comme « j'*ai voulu m'engager* pour *laisser* », ou « n'*ayant* pas la possibilité de *changer* de voie et la route *étant* mouillée ». 56% des séquences proviennent de la seconde partie, mais les participes présents et les infinitifs permettant l'expression de buts et de mondes possibles (« *désirant* me *rendre* à », « *commençant* à *tourner* »), 26% proviennent de la première partie. On constate ici une assez forte proportion d'acteurs A et B, peu de tiers C, et très peu de marques de relief – le récit n'étant ni spécialement en avant-plan, ni

spécialement en arrière-plan. On trouve beaucoup de verbes participant à la chaîne causale de l'accident, mais relativement peu mentionnant le choc.

3.4 Groupe CC du choc ou des commentaires

Les verbes d'achèvements figurent ici (45%) en plus grand nombre que partout ailleurs, au détriment des activités (seulement 6,5%) et des états (seulement 14,5%). Cela explique que ce groupe favorise globalement la partie descriptive de l'accident (57%). On observe aussi une augmentation des infinitifs et des participes en premier verbe au détriment de l'imparfait et du présent, et une augmentation massive du passé composé sur le second verbe au détriment de toutes les catégories – sauf le présent (8%, légèrement plus que la moyenne). Cette apparition du présent explique peut-être la présence de la partie 3-commentaire (29% au lieu de18% en moyenne). La mention de but ou d'alternative est moyenne. C'est ici par contre que l'avant-plan est le plus marqué. Il y a un nombre important d'acteur B (le conducteur adverse) et c'est là que l'on trouve le plus de verbes relatifs aux causes de l'accident et au choc lui-même. Le Tableau 4 indique seulement deux éléments typiques dans ce groupe qui, bien qu'assez volumineux, est plus difficile à caractériser. L'analyse en termes de point de vue montre néanmoins que la séquence finit généralement par un point de vue perfectif.

Groupe	Type	verbe 1	verbe 2
C	1	Etat ou act., IM	Etat ou act., ppr
	2	Etat, IM (ou PR)	Acc., INF
	3	Act. ou ach., IM	Acc. (ou ach.) INF
CI	4	Etat ou act., IM	Etat (ou ach.), IM
	5	Etat ou act., IM	Etat (ou ach.), PC
AA	6	Acc.(ou ach.) INF	Acc.(ou ach.) INF (ou ppr)
	7	Ach. (ou acc.), PC	Acc.(ou ach.) INF
	8	Etats, PC	Ach. INF
CC	9	Ach. (ou acc.)., INF	Ach., PC
	10	Ach.ou état, PC	Ach. (ou acc) PC

Tableau 4 : Éléments les plus typiques des quatre groupes

3.5 Bilan et commentaires

Cette catégorisation a bien distingué les états et activités (groupes C, CI) des événements (groupes AA, CC). De manière plus intéressante, les accomplissements sont aussi distingués des achèvements, justifiant la distinction accomplissement/achèvement (par opposition à la notion plus générale d'événements). On a pu également mettre en évidence que l'expression de buts ou d'alternatives passe souvent par l'utilisation de verbes au participe présent ou à l'infinitif – ce qui explique les taux réalisés par les groupes C et AA. Mais la catégorie utilisée influence aussi cette expression, car le second verbe dans ces deux groupes est généralement un accomplissement. En outre, les groupes C et CI (non marqué pour cet indice) se distinguent justement sur le type d'événement qui apparaît en second. De même les éléments différenciant les groupes AA et CC (lesquels ont cette fois une majorité d'événements) montrent que le groupe AA (qui favorise les accomplissements), bien que véhiculant un mode perfectif, est peu marqué sur le plan du relief narratif. Ce groupe est également moins concerné par les causes de l'accident que le groupe CC, et il fait peu allusion au choc. Les

buts et intentions s'exprimeraient donc plus facilement par des accomplissements que par des achèvements – lesquels seraient porteurs de plus de causalité. En effet, le groupe CC, qui favorise les achèvements, est plus fortement marqué pour l'avant plan, le choc lui-même et la chaîne causale des événements l'ayant directement provoqué. Ajoutons à ce propos qu'il semble que pour les achèvements et les activités, le sujet ait une relation de pouvoir sur l'objet direct (ou sur l'objet oblique). On peut tester son existence en utilisant des adverbes de manière (doucement, précautionneusement, etc.). Cela explique peut-être aussi la plus forte responsabilité du sujet avec des verbes d'achèvements.

Mais quoi qu'ayant bien repéré l'opposition perfectif/imperfectif (groupes AA et CC vs C et CI), cette classification a mis dans le même groupe CI des séquences à l'imparfait et les ruptures imparfait/passé composé. Une des explications est que notre algorithme d'apprentissage n'a pas tenu compte de l'ordre des phrases (ni de la distinction entre les textes), de sorte que la succession de plusieurs phrases à l'imparfait, et la structure typique de ces récits (telle que nous la percevons), n'a pas pu être bien repérée. On a ainsi manqué une part importante de notre objectif. Mais les résultats obtenus sont déjà prometteurs, puisque les trois parties ont tout de même été distribuées de manière non uniforme sur les quatre groupes. On notera également que cette classification a mis en évidence l'importance des tournures infinitives et des participes présents, et la subtilité de leurs enchaînements (cf. Tableau 4).

Améliorations techniques possibles. On a construit ici les HMM en déplaçant une fenêtre de taille 2 : un verbe est analysé au regard du verbe qui le précède et de celui qui le suit, mais pas au regard des *n* précédents ou des *n* suivants. Cela n'est pas très gênant si l'on se situe comme ici au niveau de la phrase, (dans ces récits, les phrases comportent rarement plus de trois verbes) mais pour une analyse globale prenant en compte tout le texte, on aura certainement besoin de cette amélioration. D'autre part, nous aurions aimé produire des séquences typiques de longueur variables. Ainsi, les groupes AA et CC auraient fourni des séquences de plusieurs verbes. (Cela se voit sur les éléments du Tableau 4, et nous l'avons par ailleurs constaté sur la segmentation des textes). Ce résultat pourrait être obtenu automatiquement à partir des HMM, mais nous n'avons pas eu encore le temps d'implanter cette méthode.

4 Conclusion

Notre projet général est d'appliquer les techniques de la fouille de données à la découverte de structures textuelles. Nous avons développé à cette fin une technique d'apprentissage non supervisé permettant de détecter des structures séquentielles. Elle a permis d'analyser les séquences de verbes constituant une phrase, et de proposer une classification des apparitions de deux verbes successifs en quatre groupes. Nous avons réussi à valider sémantiquement ces groupes de manière satisfaisante, en nous basant sur des annotations et des statistiques. Cela confirme à la fois le bien-fondé de la technique employée, et celui de notre couplage des temps grammaticaux avec la catégorie aspectuelle d'un verbe.

Mais ce travail n'en est encore qu'à ses débuts, et il nous reste de nombreux points à élucider. Nous regrettons tout d'abord de ne pas avoir pu comparer nos statistiques globales sur les emplois temps/catégorie à celles d'autres types de récits (et en particulier à celle de récits simples d'incident au passé). Il nous reste en effet à déterminer quelle est la part "typologique" des groupes d'enchaînements que nous avons isolés. Nous n'avons pas non plus eu le temps d'exploiter les automates probabilistes obtenus à partir de notre méthode, et ces derniers pourraient se révéler intéressants pour des applications (en particulier en génération). Il reste enfin des améliorations à apporter à la méthode générale pour prendre en compte la

structure globale des textes (non prise en compte ici), et la modélisation reste à poursuivre pour la recherche de séquences de longueur supérieure à 2.

Remerciements

Nous remercions vivement Y. Bennani pour l'aide précieuse qu'il a apportée pour le développement de la technique utilisée, et sans qui ce travail n'aurait pu être mené à bien. Nous remercions également A. Nazarenko et D. Kayser pour leurs aimables relectures.

Références

GOSSELIN L. (1996), Sémantique de la temporalité en français, Louvain-la-Neuve : Duculot.

HOPPER J. (1979). Some observations on the typology of focus and aspect in narrative language. *Studies in Language* 3.1, 37-64, Amsterdam : J. Benjamins.

KOHONEN T., (1995). *Self-Organizing Map*. Springer.

MARTIN R, NEF F. eds (1981). Le temps grammatical. *Langage* 64, KAMP H. 39-64, VLACH F. 65-79, VET C. 109-124, Paris : Larousse.

NOUIOUA F., KAYSER D. (2006). Une expérience de sémantique inférentielle. *Actes de TALN 2006,* 246-255.

RABINER L.R., JUANG B.H. (1986). An Introduction to Hidden Markov models. *IEEE ASSP Magazine*, jan. 86, 4-16.

RECANATI C., RECANATI F. (1999). La classification de Vendler revue et corrigée. *La modalité sous tous ses aspects, Cahiers Chronos 4*, 167-184. Amsterdam/Atlanta, GA.

ROGOVSCHI N. (2006). *Systèmes d'apprentissage non supervisé connexionnistes et stochastiques pour la fouille de données structurées en séquences*. Rapport de stage de Master Recherche, LIPN, Université Paris 13.

ROGOVSCHI N., BENNANI Y., RECANATI C. (2007). Apprentissage neuro-markovien pour la classification non supervisée de données structurées en séquences. Actes des *7èmes journées francophones Extraction et Gestion des Connaissances*. Namur, Belgique.

SMITH C. S. (1991). *The parameter of aspect, Studies in Linguistics and Philosophy*, Kluwer Academic publishers.

VENDLER Z. (1967). Verbs and Times. *Linguistics in Philosophy*, 97-121. Ithaca, New-York: Cornell University Press.

VUILLAUME M. (1990). *Grammaire temporelle des récits*. Paris : Minuit.

VET C. (1994). Relations temporelles et progression thématique. *Études Cognitives 1, Sémantique des Catégories de l'aspect et du Temps*, 131-149. Warszawa : académie des Sciences de Pologne.