Le temps dans le discours, expérimentation d'un protocole d'observation des caractéristiques temporelles d'un corpus d'avis de salariés

Daniel Pélissier[1], Jérôme Bousquié[2]

[1] Université Toulouse Capitole - IUT Rodez - IDETCOM - daniel.pelissier@iut-rodez.fr

[2] Université Toulouse Capitole - IUT Rodez - jerome.bousquie@iut-rodez.fr

Abstract

Time and language are closely linked. Lexicometry has developed methods for understanding speech over time. This article proposes an orthogonal problematic by studying time in language. The experiment concerns a structured method for analyzing time in discourse from a corpus of massive data of employee reviews. The protocol used makes it possible to approach meaning through temporal specificities. The observed limits of this exploratory approach justify further research on this original question.

Keywords : time, experimentation, review, protocol

Résumé

Temps et langage sont étroitement liés. La lexicométrie a développé des méthodes pour comprendre les discours dans le temps. Cet article propose une approche complémentaire en étudiant le temps dans le langage. L'expérimentation porte sur une méthode structurée d'analyse du temps dans le discours à partir d'un corpus de données massives d'avis de salariés. Le protocole utilisé permet d'aborder le sens par les spécificités temporelles. Les limites constatées de cette approche exploratoire justifient une poursuite des recherches sur ce questionnement original.

Mots clés : temps, expérimentation, avis, protocole.

Temps et langage sont étroitement liés (Brunet, 1993). La lexicométrie a contribué à mieux comprendre cette relation depuis les travaux d'A. Salem (1988) qui ont défini la notion de série chronologique. Ces recherches initiales se sont ensuite développées dans de nombreux domaines pour souligner l'existence d'un temps lexical. L'analyse de discours politiques (Mayaffre, 2000 ; de Sousa, 2012 ; Ratinaud et Marchand, 2015 ; Diwersy et Luxardo, 2016) montre une association entre le temps et le contenu du langage utilisé à des échelles historiques. Des études se sont aussi attachées à dégager des tendances de l'évolution du langage dans le temps (Wang et al., 2014 ; Dugué et al., 2016 ; Joselin et al., 2014). Au-delà des méthodes de traitement utilisées, la question de la visualisation a particulièrement mobilisé les chercheurs comme, par exemple, les diagrammes alluviaux (Rosvall et Bergstrom, 2010). Cette rapide synthèse montre un usage important de la lexicométrie dans des domaines différents pour comprendre des évolutions du langage dans le temps. Or, l'association entre le temps et le discours peut aussi concerner le temps dans le discours (Klein, 2008). Cette problématique cherchera alors à repérer les caractéristiques temporelles d'un corpus à travers ses spécificités comme le temps des verbes, certaines formes grammaticales, les adverbes de temps ou les principes de discours. Cet article propose d'aborder cette problématique du temps dans le discours pour compléter les analyses des caractéristiques d'un corpus massif de textes.

La littérature lexicométrique sur le temps dans le discours est moins prolifique que celle sur le discours dans le temps. L'analyse des temps verbaux, particulièrement en français, est complexe et difficilement automatisable. De même, les ensembles de formes associables à des questionnements temporels sont rares et peu établis scientifiquement. Quelques auteurs ont cependant proposé des méthodes et outils d'analyse. E. Brunet (1993) met en évidence une évolution des temps verbaux sur une période longue avec un usage simplifié des formes possibles. D. Longrée et al. (2004) argumentent que la répartition des temps verbaux caractérise des auteurs de textes en latin en proposant des méthodes adaptées à ce type de forme grammaticale. O. Kraif et J. Sorba (2018) puis S. Diwersy et al. (2021) étudient des données massives de romans contemporains pour dégager des styles par genre en utilisant le temps comme un des critères d'analyse. Ces recherches soulignent les enjeux d'une exploration du temps dans un discours. Elles permettent, notamment, d'analyser certaines caractéristiques d'un corpus. Dans cet article, nous proposerons un protocole d'observation des caractéristiques temporelles d'un discours. Il sera expérimenté sur un corpus de données massives d'avis de salariés sur leur employeur. Les résultats seront détaillés puis discutés.

1. Protocole d'observation des caractéristiques temporelles d'un discours

Pour contribuer à cette problématique qui relève de la littérature lexicométrique sur l'étude du temps, nous expérimentons un protocole d'observation du temps dans un discours appliqué à un corpus d'avis de salariés sur leur employeur publiés sur l'internet à partir duquel nous proposons une méthode d'analyse du temps dans le discours.

1.1 Protocole d'analyse du temps dans le discours

Le protocole suit les étapes suivantes (fig. 1) en s'inspirant des travaux de J.M. Leblanc (2015) : analyse de l'index hiérarchique à partir du corpus lemmatisé puis non lemmatisé (connecteurs et modalisateurs temporels, champ lexical du temps, temps verbaux) ; analyses contrastives (TGEN sur C.D.H. pour les formes temporelles), relations entre fréquences des formes temporelles et analyse contrastive ; étude des spécificités selon les variables du corpus ; synthèse des caractéristiques temporelles du corpus. Cette expérimentation utilise plusieurs logiciels : Tropes (Ghiglione et al., 1998), Iramuteq (Ratinaud, 2009) et TreeTagger (Schmid, 1995) et les méthodes statistiques suivantes : fréquences des formes, types généralisés (Lamalle et Salem, 2002), classification descendante hiérarchique (Reinert, 1983). Chaque étape sera détaillée dans la section suivante présentant les résultats.

1.2. Présentation du corpus

Le corpus étudié contient des avis publics et anonymes de salariés sur leur employeur collectés sur une plateforme internet. Chaque avis est associé à une note de 1 à 5 comme sur *Tripadvisor*. Exemple d'avis : « *La politique d'évolution interne est très intéressante, c'est un lieu où l'on est en perpétuel mouvement* ». Nous avons retenu les entreprises pour lesquelles plus de 50 avis avaient été publiés. Le corpus final est de 118 602 avis publiés pour 429 entreprises après nettoyage (avis en langue étrangère, doublons, etc.) contenant 4 451 793 occurrences pour 48024 formes et 38 formes par avis. Ce processus de nettoyage et d'adaptation du corpus à notre problématique s'est étalé de février à mai 2021.

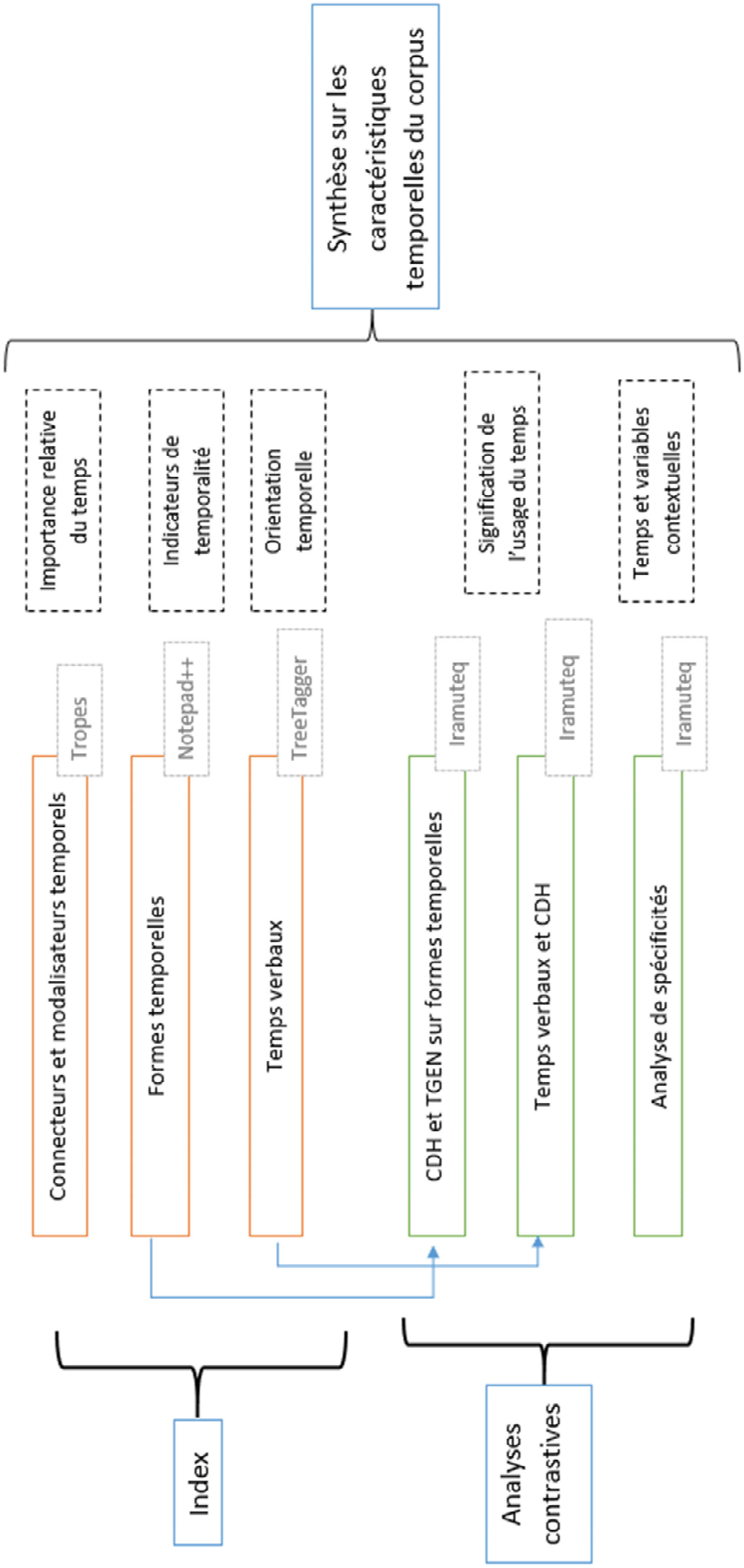

Figure 1. Synthèse du protocole d'observation des caractéristiques temporelles d'un discours

2. Résultats de l'expérimentation du protocole

Cette section présentera les principaux résultats du protocole suivi et appliqué au corpus utilisé pour cette expérimentation.

2.1 Analyse de l'index hiérarchique temporel

Cette première phase de la méthode consiste à analyser les grandes tendances temporelles présentes dans le corpus. Les approches abordent successivement (fig. 1) : le temps en référence à d'autres catégories d'analyse (le corpus est-il plus ou moins temporel ?), le vocabulaire du temps (quel type de vocabulaire temporel est présent dans le corpus ?), les temps verbaux (quel type de temps domine le discours (passé/présent/futur) ?).

2.1.1 La catégorie temporelle dans le discours

Cette phase a pour finalité de fournir des caractéristiques temporelles du discours. Pour cela, un comptage des formes préalablement classées par catégories est réalisé à l'aide du logiciel Tropes. Les discours de ce corpus utilisent peu les connecteurs temporels au sens de ce logiciel (connecteurs temporels : 2,99% des connecteurs). L'orientation argumentative explique la domination des connecteurs d'addition, d'opposition ou de cause. Les formes temporelles servent plus à nuancer, modaliser les avis (modalisateurs temporels : 10,96% des modalisateurs) loin cependant des modalisateurs d'intensité qui sont majoritaires. Ce corpus est ainsi faiblement temporel selon cette première approche globale.

2.1.2 Les formes temporelles du discours

Bien que peu temporels, ces avis de salariés utilisent du vocabulaire associé au temps. Les listes de formes temporelles sont difficiles à établir, sans doute car la catégorisation de mots aux sens multiples est toujours sujette à caution. Nous avons ainsi établi deux listes avec des sources différentes. La première liste comporte 52 connecteurs temporels indiquant la simultanéité, la fréquence, la postériorité et l'antériorité[1]. La deuxième liste est issue des travaux de O. Kraif et J. Sorba (2018)[2] et comprend 42 formes associées au temps que nous avons catégorisées en période, moment, fréquence, antériorité, événement, situation, découpage et saison. Ces listes sont complémentaires car elles abordent des thématiques temporelles différentes, certaines formes sont cependant communes aux deux listes. Les occurrences de chaque forme ont été comptabilisées, les formes retenues représentant environ 90% du total de chaque liste.

La liste 1 met en évidence principalement des connecteurs de simultanéité (42,22%) comme 'quand' ou 'pendant' et des connecteurs de fréquence (33,45%) comme 'toujours' ou 'jamais'. Ces types de formes sont aussi très présents dans la liste 2 (27,92%). Ce deuxième ensemble met surtout en avant les formes de découpage du temps (39,43%) avec 'journée' principalement ou encore 'mois' et 'semaine'.

Ce relevé montre ainsi un usage des formes temporelles favorisant un découpage en journée de travail, en description d'actions simultanées et en soulignant une fréquence très variable, de la rareté à la profusion.

---

[1] à peine, à présent, après, au fur et à mesure, au moment, aujourd'hui, auparavant, aussitôt, autrefois, avant, bientôt, cependant, déjà, demain, depuis, derechef, dès que, désormais, en attendant, enfin, ensuite, hier, jadis, jamais, jusqu'à, le temps que, longtemps, lors, lorsque, maintenant, naguère, parfois, pendant, postérieurement, puis, quand, quelquefois, sitôt, soudain, souvent, sur ces entrefaites, sur-le-champ, tandis, tant, tantôt, tard, tôt, toujours, tout à coup, tout à l'heure, tout de suite, ultérieurement. Liste établie à partir du dictionnaire Le Robert et de la publication de J.-M. Kalmbach (2020).

[2] an, année, après-midi, aujourd'hui, automne, autrefois, brièvement, décennie, demain, fois, heure, hier, hiver, immédiatement, instant, jadis, jamais, jour, journée, lendemain, longtemps, maintenant, matin, millénaire, minute, mois, moment, nuit, parfois, printemps, seconde, semaine, siècle, soir, soirée, souvent, subitement, tard, temps, tôt, toujours, tout à coup, tout de suite. La forme « été » n'a pas été retenue en raison des confusions entre le verbe être et la saison.

2.1.3 Les temps verbaux

Le temps des verbes est un marqueur temporel important dans un discours mais difficile à analyser sur des données massives. Les confusions sont possibles, par exemple : '*j'ai adoré allez travailler*' (erreur comprise) sera difficilement interprété par un algorithme. Afin d'obtenir des résultats assez satisfaisants et automatiques, nous avons utilisé TreeTagger[3]. Pour éviter des interprétations peu précises, nous avons retenu seulement les analyses ayant une probabilité supérieure à 80% ce qui a permis d'interpréter 82% des verbes de ce corpus. Le résultat montre une orientation très majoritaire vers le présent avec quelques rares discours au passé. Malgré la précision des mesures et les précautions prises, ces données doivent être interprétées avec prudence. Des erreurs d'interprétation demeurent ('missions' est analysé comme un subjonctif imparfait et pas un nom) mais la tendance est cependant suffisamment claire pour permettre des analyses.

Le croisement de ces orientations générales avec des sous-catégories du discours permettra de faire une analyse contrastive, mieux située dans le contexte du corpus.

2.2 Analyses contrastives du discours

L'analyse contrastive a croisé les deux listes de formes temporelles (2.1.2) avec une classification descendante hiérarchique grâce aux types généralisés et à des variables associées aux discours (analyse de spécificités).

2.2.1 Classification descendante hiérarchique (CDH) et type généralisé

Les deux listes de formes temporelles, limitées aux formes les plus fréquentes ont été croisées avec une classification descendante hiérarchique[4] (Reinert, 1983) par un type généralisé (Lamalle et Salem, 2002). Les résultats permettent de connaitre le degré d'association des formes temporelles avec chaque classe. Cette analyse distingue plusieurs types de classe au filtre de cette analyse temporelle. Certaines comme la classe 6 utilisent un vocabulaire non temporel. La classe 6 correspond à des avis de stagiaires très courts qui énoncent rapidement leur satisfaction. Au contraire, d'autres classes comme les 7, 8 et 9 contiennent des formes temporelles. Globalement, les résultats des deux TGEN présentent des résultats similaires en mettant en évidence ces trois dernières classes. Il existe cependant des nuances. Par exemple, la classe 8 qui critique les salaires utilise des formes temporelles de découpage, le mois en particulier, ou la classe 9 avec la journée. Les raisons de ces différences peuvent être liées à la longueur des avis regroupés dans les classes. Un discours long aura plus de chance de contenir des formes temporelles ce qui pourrait correspondre aux classes 7, 8 et 9. La classe 4, très longue aussi, est cependant un contre-exemple. L'hypothèse, pour ce corpus, est que les classes contenant des discours critiques et/ou décrivant une action sont des récits, des narrations intégrant des logiques temporelles.

2.2.2 Temps verbaux et CDH

Le croisement des temps verbaux avec les différentes classes souligne les caractéristiques de quelques regroupements. Pour montrer des différences, nous avons centré les analyses sur les deux temps principaux, présent et passé. L'usage du temps passé correspond à des classes présentant le contenu d'emploi, la description d'une journée type pour les classes 2, 9 et 11. L'usage du passé pour la classe 6 s'explique plutôt par le type de narrateur, des stagiaires. Les autres classes sont dans la moyenne générale. On peut cependant noter que l'usage de

[3] Tags utilisés : VER:pres (présent), VER:ppre (participe présent), VER:subp (subjonctif présent), VER:futu (futur), VER:impf (imparfait), VER:simp (passé simple), VER:subi (subjonctif imparfait), VER:cond (conditionnel), VER:impe (impératif)

[4] Dans ce cas, nous avons obtenu 11 classes regroupant 92,45% des segments du corpus.

l'impératif est beaucoup plus fréquent dans les classes 4, 7 et 8 qui contiennent des discours critiques avec, notamment, l'expression '*Fuyez* !'.

2.2.3 Spécificités du discours et formes temporelles

Cette dernière étape consiste à croiser les listes de formes temporelles (listes 1 et 2) avec les variables associées aux discours étudiés. Dans ce cas, nous avons choisi la variable des notes et deux formes associées à la fréquence, la seule catégorie commune aux deux listes. Ces deux formes ont des fréquences opposées qui sont associées à des notes également différentes. La forme '*toujours*' est associée à des notes élevées comme la note 5 alors que '*jamais*' est plus associée à la note 2. Les avis associés à la note 5 tendent à généraliser une interprétation positive de leur employeur alors que les notes faibles soulignent des absences. Ce croisement aurait pu aussi se faire avec des variables temporelles ce qui permettrait une analyse du temps dans le discours… dans le temps.

3. Discussions et conclusion

Le protocole expérimenté dans cet article a permis de mieux comprendre une partie du sens des discours. L'utilisation massive du présent montre que les auteurs sont, *a priori*, des employés actuels. L'approche montre aussi que certains discours intègrent peu les formes temporelles. Enfin, il existe un lien entre le contenu du discours et ses caractéristiques temporelles comme l'ont montré O. Kraif et J. Sorba (2018) puis S. Diwersy et al. (2021) pour les romans. La critique des employés pour ce corpus est structurée temporellement et propose une forme plus narrative que d'autres avis enthousiastes associés aux logiques de marque employeur. L'apport principal est cependant de proposer un travail d'élaboration d'une méthode d'observation du temps dans le discours. A ce titre, il a de nombreuses limites. D'abord, le corpus choisi est faiblement temporalisé ce qui n'a pas facilité le travail sur ce questionnement. D'autres corpus plus variés pourront répliquer, compléter et affiner cette première approche. Ensuite, le choix des mots temporels est une étape importante mais délicate et orientant les explications suivantes. De plus, la détermination du temps des verbes, comme souligné précédemment, est améliorable avec de nouveaux outils comme des algorithmes intégrant de l'apprentissage supervisé. De même, nous n'avons pas étudié les liens entre temps verbaux, comme variable explicative, et classification par des tests du Chi². Ces trois limites importantes (choix des formes, temps verbaux, variable) ouvrent de nombreuses pistes de recherche futures.

Bibliographie

Brunet, E. (1993). Quand le temps change avec le temps. *Texto ! Textes et Cultures*, XXI(1), 1-27.

Diwersy, S., Gonon, L., Goossens, V., Kraif, O., Novakova, I., Sorba, J., & Vidotto, I. (2021). La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase. *Corpus*, 22, 1-23.

Diwersy, S., & Luxardo, G. (2016). Mettre en évidence le temps lexical dans un corpus de grandes dimensions : L'exemple des débats du Parlement européen. *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 1-12.

Dugué, N., Lamirel, J.-C., & Cuxac, P. (2016). Visualisation pour la détection d'évolutions dans des corpus de publications scientifiques Indexation, classification et analyse diachronique pour la visualisation. *Les Cahiers du Numérique*, *4*, 157-183.

Ghiglione, R., Landré, A., & Molette, P. (1998). *L'Analyse automatique des contenus*. Dunod.

Joselin, L., Eliot, E., Jeanne, P., Lepastourel, N., Gasquet, C., & Amalric, M. (2014). Dynamiques temporelles de la pandémie de grippe A/H1N1 dans la presse écrite francophone. *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 1-12.

Kalmbach, J.-M. (2020). *Guide de grammaire française pour étudiants finnophones*. Institut des Sciences du langage et de la communication Université de Jyväskylä Finlande.

Klein, W. (2008). Time in Language, Language in Time. *Language Learning*, *58*(Supplément 1), 1-12.

Kraif, O., & Sorba, J. (2018). Spécificités des expressions spatiales et temporelles dans quatre sous-genres romanesques (policier, science-fiction, historique et littérature générale). *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 392-399.

Lamalle, C., & Salem, A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. *Journées internationales d'analyse statistique des données textuelles (JADT*), 403‑411.

Leblanc, J.-M. (2015). Proposition de protocole pour l'analyse des données textuelles : Pour une démarche expérimentale en lexicométrie. *Nouvelles perspectives en sciences sociales*, *11*(1), 25-63.

Longrée, D., Luong, X., & Mellet, S. (2004). Temps verbaux, axe syntagmatique, topologie textuelle : Analyses d'un corpus lemmatisé. *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 743-752.

Mayaffre, D. (2000). Temps lexical ou temps politique ? *Lexicometrica*, 1-10.

Ratinaud, P. (2009). *Iramuteq* (Version 0.7 Alpha 2) [Linux, Windows, MacOS].

Ratinaud, P., & Marchand, P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). *Mots. Les langages du politique*, *108*(2), 57-77.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, *8*(2), 187-198.

Rosvall, M., & Bergstrom, C. (2010). Mapping Change in Large Networks. *PLoS ONE*, *5*(1), 1-7.

Salem, A. (1988). Approches du temps lexical Statistique textuelle et séries chronologiques. *Mots*, *17*, 105-143.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*, 1‑9.

Sousa (de), S. (2012). À l'épreuve des temps… Temps lexical et temps politique dans le discours de Fidel Castro (1959-2008). *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 337-349.

Wang, X., Cheng, Q., & Wei, L. (2014). Analyzing evolution of research topics with NEViewer : A new method based on dynamic co-word networks. *Scientometrics*, *101*, 1253–1271.

ANNEXE 1. Fréquences des connecteurs et des modalisateurs dont temporels (Tropes)

Catégories de connecteurs

Types	Nombre	%
Addition	108 356	59,08%
Opposition	26 182	14,27%
Cause	13 587	7,41%
Disjonction	9 944	5,42%
Comparaison	9 822	5,36%
Condition	7 373	4,02%
Temps	**5 481**	**2,99%**
But	2 669	1,46%
Total	*183 414*	

Catégories de modalisateurs

Types	Nombre	%
Intensité	155 227	52,26%
Négation	49 021	16,50%
Temps	32 546	10,96%
Lieu	29 585	9,96%
Manière	21 580	7,27%
Affirmation	8 642	2,91%
Doute	432	0,15%
Total	297 033	100%

ANNEXE 2. LISTE 1, fréquences des formes (89,5% des connecteurs pour ce corpus)

Forme	Nombre
toujours	5926
tant	3594
quand	3429
après	3020
jamais	2456
pendant	2164
avant	2152
depuis	2000
parfois	1844
puis	1839
souvent	1763
lors	1580
cependant	1195
lorsque	1169
ensuite	1035
déjà	675

Type	Nombre	%
Antériorité	2827	7,89%
Fréquence	11989	**33,45%**
Postériorité	5894	16,44%
Simultanéïté	15131	**42,22%**
Total	*35841*	*100%*

ANNEXE 3. LISTE 2, fréquences des formes (88,37% des connecteurs pour ce corpus) liste 2

Forme	Nombre
journée	8118
temps	6303
toujours	5926
mois	4564
jour	2758
jamais	2456
fois	2090
parfois	1844
matin	1822
semaine	1795
souvent	1763
heure	1357
année	1096
moment	1050

Types	Nombre	%
Découpage	16930	39,43%
Moment	9443	21,99%
Fréquence	11989	27,92%
Période	4580	10,67%
Total	*42942*	*100,00%*

ANNEXE 4. Temps verbaux (TreeTagger, probabilité > 80%)

Orientation générale	Nombre (p>0,8)	%
Présent	244 069	**86,98%**
Futur	3 906	1,39%
Passé	29 821	**10,63%**
Conditionnel	2 666	0,95%
Impératif	150	0,05%
Total	*280 612*	*100,00%*

ANNEXE 5. CDH et TGEN pour les deux listes de formes temporelles (Iramuteq)

Classe	TGEN liste 1	TGEN liste 2	Contenu des classes	Longueur (Chi²)	Présent	Passé
1	-97,47	-24,7	Métiers de préparateur	Très court (1382,22)	85,41%	13,34%
2	-44,76	-13,95	Hôtesses de caisse	Très court (1821,47)	83,93%	14,86%
3	-72,62	-719,46	Culture d'un grand groupe	NS	90,51%	8,55%
4	-1,71	-465,57	Analyse et critique du management, commercial	Très long (879,73)	90,93%	8,08%
5	-71,44	-721,05	Expérience d'apprentissage	NS	87,40%	11,77%
6	-437,64	-936,55	Avis positifs de stagiaires	Très court (2031,67)	85,46%	14,06%
7	1757,18	1572,73	Alerte sur les conditions de travail	Très long (6300,78)	87,03%	11,68%
8	974,09	7163,81	Alerte sur les salaires	Très long (1953,42)	87,22%	11,65%
9	298,96	1485,48	Gestion de rayon	Très long (536,92)	82,45%	17,07%
10	-264,68	-517,85	Evolution de carrière	NS	91,88%	7,26%
11	-335,39	-434,58	Gestion de la maintenance	NS	84,91%	14,36%

ANNEXE 6. Temps verbaux et classes CDH

Classe	Présent	Passé
1	85,41%	13,34%
2	83,93%	14,86%
3	90,51%	8,55%
4	90,93%	8,08%
5	87,40%	11,77%
6	85,46%	14,06%
7	87,03%	11,68%
8	87,22%	11,65%
9	82,45%	17,07%
10	91,88%	7,26%
11	84,91%	14,36%

ANNEXE 7. Analyse de spécificités sur formes lemmatisées, Chi² (Iramuteq)

Forme	Note 1	Note 2	Note 3	Note 4	Note 5
jamais	0	20	-19	-63	**-29**
toujours	**-5**	-3	-7	1	20