Ébauche du sommaire

1. **Introduction**
   - ***1.1*** **Contexte générale –** *Evolution des masses de données de textuels, différentes approches de traitements, cas spécifiques des communications médiés par les ordinateurs et réseaux, contenus très peu structurés contenant beaucoup d'informations et comment extraire ces dernières*
   - **1.2 Objectifs des travaux** – *Allusion ici à ceux de la thèse*
   - **1.3 État de l'art**
   - **1.4 Organisation du document**

2. **Communication assistée par ordinateur (Computer Mediated Communicaiton)**
   - **2.1 Introduction**
     - **2.1.1 Avantages et inconvénients**
     - **2.1.2 Communication orale, communication écrite**
     - **2.1.3 Communication synchrones et asynchrones**
   - **2.2 Caractéristiques des canaux étudiés**
     - ***2.2.1*** **Courriels -** *définition, caractéristiques et actions, structuration et principaux travaux, métadonnées*
     - **2.2.2 Forums -** *--//--*
     - **2.2.3 Chat -** *--//--*
     - **2.2.4 Tableau comparatif**
   - **2.3 Analyse discursive de courriels –** *Identification et représentation des éléments de discours et de pragmatique dans les emails*
   - **2.4 Graphes et fil de discussion**
   - **2.5 Fil de discussion et analyse discursive**

     Qualifier les threads : de qui parle-t-on, de quoi parle-t-on… étudier les caractéristiques, les types… Explorer les relations discursives entre les threads constitués. Identifier les aspects de caractérisation, qualification et théorie de structure autour des threads.

3. **Analyse discursive**
   - **3.1 Introduction**
   - **3.2 Actes de discours et de dialogues**
     - **3.2.1 théorie des actes de Langage** – *Premières taxonomies (Searle et Austin)*
     - **3.2.2 Actes de discours et de dialogues**
       - **a) Analyse des conversations fonctionnelles -** *De la limitation des énoncés isolés à l'atteinte des objectifs communs entre participants d'une conversation*
       - ***b)*** **Prise en compte de contexte** – *Informations utiles à l'interprétation d'énoncés de conversation*
       - ***c)*** **Marqueurs pragmatiques multimodaux,** *L'étude de la façon dont le contexte du texte affecte le sens d'une expression, et quelles informations sont nécessaires pour déduire un sens caché ou présupposé*
     - **3.2.3 Schémas d'annotation –** *DAMSL, DIT++ et le standard ISO*

4. **Corpus et exploitation**
   - **4.1 Introduction**
   - **4.2 Corpus Orange**
     - **4.2.1 Contraintes juridiques liées à l'utilisation des données**
       - **a) Demande de consentement –** *description et procédure*
       - **b) Anonymisation -** *principe, techniques, avantages et inconvénients*
       - ***c)*** **Pseudonymisation -** *principe, --//--*
     - **4.2.2 Exploitation**
     - **4.2.3 Nettoyage** - Contenu superflus ou inintéressant et leur suppression
     - ***4.2.4*** **Encodage -** *Format de structuration de corpus pour traitement (TEI, Json)*
   - ***4.3*** *Autre Corpus*
     - ***4.3.1*** *Enron*
     - ***4.3.2*** *WikiDisc*
     - ***4.3.3*** *BC3*

5. **Formalisation de la problématique, approches proposées**
   - **5.1 Introduction**
   - **5.2 Formalisation**
   - **5.3 Approches de résolution du problème**
     - **5.3.1 Reconnaissance d'actes de dialogues**
     - **5.3.2 Prédiction d'email suivant**
     - **5.3.3 Prediction relation transverse**
     - **5.3.4 Combinaison des deux précédentes approches**
     - **5.3.5 Segmentation sémantique**
   - **5.4 Analyses**
   - **5.5 Évaluation**
   - **5.6**

6. **Implémentation des approches**
   - **6.1 Introduction**
   - **6.2 Approches basées sur les règles**
   - **6.3 Approches statistiques –** *objectif, principe de fonctionnement et limitation, exemple*
     Clustering , Arbres de décision, Classifieur bayesien naïf, Entropie Maximale, Induction de modèle structuré (HMMs, CRFs)

7. **Conclusion**

Rappel des notions et concepts présentés, orientation de ces derniers pour les objectifs de la thèse en prenant en compte les contraintes comme celle liée à une solution fonctionnant sur PC utilisateur. Ebauche des prochaines pistes de travail et d'avancement de la thèse.

Chap 1. Introduction

Dans le monde de la technologie et plus précisément dans le domaine de communication, on assiste de plus en plus à une évolution marquée tant sur le plan matériel que logiciel. On peut ainsi noter la sortie ces dernières années de téléphones très puissants avec des capacités pouvant supporter de multiples applications, telles que les natives (appels, SMS, calendrier, etc.) améliorées avec des services intelligents comme l'auto-complétion des textes lors de l'envoi d'un SMS, des assistants vocaux pour appeler des contacts. Les jeux vidéo et les applications de communications ont tout aussi connu des avancées considérables avec le plus qu'a apporté l'intelligence artificielle dans tous les domaines. Concernant les applications de communications, on a assisté à un boom d'applications de ***chat*** telles que WhatsApp, Messenger, Viber, WeChat, Line, etc. ; des plates-formes de communication collaborative telles que Slack, Discord, Fleep, Workplace, etc. On ne peut parler d'outil de communication sans mentionner le principal d'entre-eux utilisé en entreprise à savoir l'email au travers des applications et services comme Outlook, Gmail, Yahoo, etc.

Selon les estimations de The Radicati Group concernant la messagerie professionnelle, le nombre d'e-mails échangés quotidiennement au niveau mondial était estimé à plus de 128 milliards en 2019. Au niveau individuel, ces estimations se déclinent de la manière suivante : en 2019 chaque collaborateur a envoyé 30 e-mails par jour en moyenne et en a reçu 96, dont 19 étaient du spam.

Cette explosion d'outils numériques de communication et de collaboration a conduit à la génération d'énormes quantités de données notamment textuelles qui sont en général sont stockées sur des serveurs ou des postes personnels sous formes d'archives et sont au fur et à mesure supprimées pour des raisons de limitation d'espace de stockage ou de non exploitation. En général ces données ne sont pas exploitées du fait qu'elles sont peu ou pas structurées.

Comme exploitation primaire, on a la recherche d'informations faite par des cadres en entreprise. Ces derniers passent beaucoup de temps de travail sur cette tâche de recherche d'informations, dans leur messagerie par exemple. Des études estiment ente 20 et 30% ce temps passé à chercher de l'information enfouie dans les emails.

Concernant les conversations d'emails, elles regorgent d'informations très importantes surtout dans le contexte professionnel où plusieurs collaborateurs échangent entre eux via courriels dans le but d'atteindre un objectif commun comme, par exemple l'aboutissement d'un projet. Ces conversations échangées se présentent en général sous plusieurs formes (réponses, transferts, mailings List, …) avec ou sans pièces jointes. Les courriels vont de simples envois à des transferts, en passant par des réponses. Ces dernières sont parfois faites de façon imbriquée, c'est-à-dire que les phrases sont insérées directement au niveau de celles du courriel auquel on répond. On retrouve aussi des mails transférés qui donnent lieu à de nouvelles conversations. La non structuration des courriels est en partie due à ces formats d'édition de mails non standardisés, entrainant ainsi un accès difficile aux informations véhiculées dans ces courriels. Cette accès difficile ou incompréhension de l'information n'est pas seulement une conséquence de l'absence ou d'une mauvaise structuration des courriels, mais elle est aussi fonction des aspects linguistiques très variés utilisés par les participants dans une conversation par mails. Ces incompréhensions sont très souvent dues aussi à l'absence de connaissance d'un contexte commun par les participants.

Beaucoup de recherches ont émergé ces dernières décennies afin d'exploiter ces grandes quantités d'information issues notamment des courriels, mais aussi de tenter de résoudre ces problèmes de structuration logique et d'incompréhension soulevés précédemment. Ceci se fait via des techniques à

base de règles ou de traitement automatique du langage naturel. Dès les années 1990 des travaux de recherche portent sur l'apprentissage automatique (*Machine Learning* en anglais) appliqué à la classification de documents selon leurs thématiques [CITATION Lew92 \l 1036]. Quelques années plus tard l'apprentissage automatique est également utilisé pour classer des mails [CITATION Coh92 \l 1036] et pour détecter des sentiments [CITATION Pan02 \l 1036]. En s'appuyant sur les travaux de Searle [CITATION Sea75 \l 1036] sur les actes de langage (*Speech Acts* en anglais) et sur ceux de Finke et al. visant à détecter automatiquement des actes de langage dans des conversations téléphoniques [CITATION Fin98 \l 1036]. Cohen et al. [CITATION Coh04 \n \t \l 1036] proposent une approche visant à détecter l'intention des auteurs des mails grâce à une ontologie d'actes de langage. Leur conclusion était, d'une part, qu'il faudrait tenir compte du contexte d'un mail afin de pouvoir détecter des actes de langage implicites et, d'autre part, qu'il est fréquent qu'un message porte sur plusieurs sujets de discussion en même temps. Ceci soulève deux problèmes difficiles : la **segmentation** des messages et le **démêlage des discussions imbriquées**. Au-delà de ces deux problématiques identifiées, il en existe bien d'autres qui nécessitent qu'on s'y intéresse particulièrement parce qu'elles sont soit *connexes* ou *imbriquées* aux précédentes. Il s'agit de l'***identification d'actes de langage*** et de ***fils de discussions*** ciblés sur une tâche simple et précise dans des ***conversations asynchrones***. Nous les aborderons après les deux prochains paragraphes liés au démêlage et la segmentation.

Le problème des discussions imbriquées a fait l'objet de nombreux travaux récents. Dulceanu [CITATION Dul16 \l 1036] propose une méthode pour le démêlage (*Disentanglement*) des conversations de type *tchat* de manière à organiser toutes les « paroles » (*Utterances*) en fils de discussion logiquement ordonnés selon leurs contenus et les thématiques abordées. D'autres travaux plus récents [CITATION Mic10 \l 1036], [CITATION Mic11 \n \t \l 1036] et [CITATION Jyu19 \l 1036] ont utilisé les concepts de cohérence, de contexte avec des architectures récentes de réseaux de neurones profonds afin de résoudre le problème de **démêlage** de discussion.

La segmentation de texte a pour but de déterminer les frontières entre les différents sujets abordés dans des textes longs ou dans des flux de texte afin de diviser ces textes en un ensemble de segments, chacun consistant en une séquence consécutive de phrases et de paragraphes partageant un sujet cohérent [CITATION XJi03 \l 1036]. Elle est utilisée, par exemple, dans certaines approches de génération de résumé de texte. Il existe principalement deux catégories de tâches de segmentation de texte. L'une vise à identifier, dans des flux de texte, les endroits correspondant aux changements de thématiques. L'autre vise à identifier et isoler des sous-thématiques via le découpage de documents d'une longueur substantielle. La segmentation est plus difficile dans le cas des documents longs car, d'une part, ceux-ci abordent fréquemment des sous-thématiques proches les unes des autres ou des variantes de la même thématique et, d'autre part, les transitions sont souvent plus subtiles. La segmentation de texte semble être un prérequis pour traiter le problème des discussions imbriquées dans les cas où celles-ci contiennent de longs messages. Plusieurs techniques ont été développées au fil du temps afin de résoudre le problème de segmentation. Xiang & Hongyuan [CITATION Xia03 \l 1036] ont utilisés la technique de diffusion anisotropique des images couplée à des adaptations de méthodes de programmation dynamique pour identifier des frontières de texte indépendamment des domaines. Plus récemment Goran G. et Swapna S. [CITATION Gla20 \l 1036] se sont appuyés sur les nouveaux modèles de **NLP** en l'occurrence les **Transformers** pour résoudre la problématique de segmentation de texte avec des résultats qui sont meilleurs que ceux de l'état de l'art.

Dans les prochains paragraphes, nous allons brièvement présenter les problématiques adjacentes aux précédentes.

L'analyse des emails fait ressortir des structures logiques qu'ils intègrent et qui peuvent facilement être identifiées suivant les relations ***parent-child*** (réponse à un mail) et ***sibling*** (transfert d'emails). Cependant

couramment dans des échanges d'emails, les participants créent des conversations dans lesquelles plusieurs sujets sont abordés contenant des demandes et partages d'informations, des engagements, etc. Ces différentes actions, dans le cadre de l'analyse discursive ou de contenus écrits sont appelées aux ***actes de langages***. Ils sont généralement en relation (l'un étant une action conséquente à une précédente ou entrainant autre) ; par exemple un partage d'information serait dû à une demande préalablement effectuée. De cette façon, on peut identifier un ensemble d'actes de langages étant en relation les uns aux autres et dont l'extraction constitue un ***fil de discussion*** simple et orienté vers une cible spécifique considérée comme une ***sous-conversation.*** Cette extraction de fils de discussion serait très simple et même se ferait facilement/manuellement si elle est faite sur des petites conversations avec un nombre très réduit de participants. Cependant en considérant les grandes quantités d'informations stockées dont on a parlé plus haut, il serait difficile, voire même impossible d'en extraire des fils de discussion. Mais avec l'avènement des méthodes statistiques et de l'intelligence artificielle, plusieurs travaux ont été menés sur ces problématiques d'identification d'actes de langage et ainsi de fils de discussions.

La construction de structure d'emails en passant par l'***identification de fils de discussions*** a fait l'objet de plusieurs recherches. L'un des premiers travaux autour des emails est celui de Steve Wittaker et Candace Sidner [CITATION Whi96 \l 1036] qui ont exploré la gestion des données personnelles dans les emails en effectuant des comparaisons entre l'objectif premier des emails qui était juste un outil de communication et les multiples fonctions additionnelles dont il a fait l'objet par la suite, qu'ils ont qualifiées de surcharge d'emails (**Emails overload**). Dans les années 2000, Bernard Kerr à travers **Thread arcs** [CITATION Ker03 \l 1036] développe une nouvelle technique de visualisation des mails qui facilite tout aussi la recherche d'informations. Jen-Yuan Yeh [CITATION Jen06 \l 1036] a utilisé les appariements de similarité afin de reconstruire les fils de discussions d'emails. Les algorithmes de clustering couplés à des informations linguistiques ont permis à Dou Shen et al. [CITATION She06 \l 1036] d'améliorer les performances des méthodes qui existaient pour la détection de ***threads*** (fils de discussion) dans les contenus de messagerie instantanée et de forums. Un peu plus récemment, en intégrant des méthodes récentes d'apprentissage profond pour la classification et le clustering avec des données tridimensionnelles (contenus sémantiques, interactions sociales et temporalité) extraites des échanges, Giacomo Domeniconi et al [CITATION Dom17 \l 1036] ont d'avantage amélioré les performances de l'état de l'art sur la problématique de détection de fil de discussions. De ces différents travaux, on constate que plusieurs aspects des contenus issus des **communications médiées par ordinateurs (CMO)** ont été pleinement étudiés. Il ressort néanmoins de ces études surtout sur la détection de fils de discussions, que cette dernière n'a pas été faite avec un niveau de granularité assez fin. Ce que nous essaierons d'approcher dans nos travaux. Mise à part les reconstructions de structures de conversations en s'appuyant sur des caractéristiques linguistiques (**features**) et temporelles pour beaucoup de ces travaux, on a pu constater une autre vague de recherches avec pour point central les actes de langages ou actes de dialogues (***dialog act***) mais toujours orientée sur l'analyse des textes des CMO.

Les pionniers des travaux sur les théories d'actes de langages sont Austin [CITATION Aus75 \l 1036] et Searle [CITATION Joh69 \l 1036] [CITATION Joh76 \t \l 1036] qui ont développé les premières taxonomies d'actes de langages. Du fait des limitations de ces taxonomies parce que leurs éléments était étudiés de façon isolée et parce que l'objectif des actes de langage avait évolué pour l'annotation des dialogues, il y a eu bien d'autres paramètres qui sont rentrés en jeu comme la prise en compte du ***contexte*** dans une conversation ainsi que la pluralité des intentions des participants ; donnant ainsi lieu à la création de schémas d'annotations :

- ***DAMSL*** [CITATION Cor97 \l 1036] dont la fonction première était d'annoter des actes de dialogue en tant qu'opérations de mise à jour du contexte,
- ***DIT++*** [CITATION Har \l 1036] avec comme axe principal la définition un acte de dialogue comme combinaison de son contenu sémantique et de sa fonction communicative.

En 2012, les principaux contributeurs travaillant sur les actes de langages ont mené des travaux qui ont donné naissance à la norme ***ISO 24617-2*** [CITATION Bun12 \l 1033] qui est un standard d'actes de dialogue basé sur la sémantique qui est devenu depuis lors une référence pour des recherches autour desdits actes de dialogues, aidant ainsi l'annotation de conversations.

Des premières taxonomies d'actes de langage jusqu'à la norme ***ISO 24617-2,*** ces actes de dialogue ont fortement impacté les études portant sur l'analyse des contenus de CMO à travers plusieurs problématiques dont la modélisation de langage, la prédiction de ces actes, la classification des mails et même la segmentation sémantique dans certains contextes. En 1996, Stefan Wermter et Matthias Lochel [CITATION Wer96 \l 1033] proposent une nouvelle approche d'apprentissage d'actes de dialogue intégrant un parseur pour la segmentation sémantique et symbolique couplé à un réseau d'actes de dialogues. Quelques années plus tard W. Cohen et al. présentent une méthode qui observe passivement les emails et les classe en actes de dialogues [CITATION Coh041 \l 1033], ceci via des algorithme de ***machine learning (VP - voted perceptron, SVM-Support Vector Machine, AB-AdaBoost et DT-Decision Tree).*** Dans le même ordre d'idée, Jeong et al. [CITATION Min09 \l 1033] adoptent une approche ***semi-supervisée*** pour la reconnaissance des actes de dialogues dans les ***emails*** et les ***forums***. Leur approche est basée sur de grandes quantités de données non labélisées (***unlabeled data***) dont les features telles que les phrases et arbres de dépendance ont été extraites de façon automatique. Les travaux récents sur la modélisation de ***contexte*** couplé aux ***mécanismes d'attention,*** ont amélioré les performances sur la problématique de ***classification d'actes de dialogues***. Vipul Raheja et Joel Tetreault [CITATION Vip19 \n \ l 1033] atteignent 82,9% de précision sur cette tâche. Par ces multiples travaux, il ressort très bien que la prise en compte des actes de dialogue est un aspect à considérer avec une très grande attention dans l'analyse des CMO.

Dans les études que nous avons présentées de façon très brèves dans les précédents paragraphes, la notion de ***corpus*** a été abordée de façon peu marquée. Et pourtant sans les données, aucun de ces travaux ne saurait exister. Il est donc important de s'y pencher avec quelques lignes dans cette introduction bien qu'un chapitre dans ce document lui sera complètement consacré. Nous allons juste mentionner ici quelques travaux de constitution de corpus, plus précisément de corpus d'emails. ***Enron*** [CITATION Bry04 \l 1036] contenant environ 500 000 emails pour 150 participants est le corpus le plus largement utilisé depuis sa sortie pour les études ou recherches sur les mails. En 2015, Oard D. et al. [CITATION Dou15 \l 1036] mettent sur pied le corpus ***Avocado*** qui est une collection d'emails et de pièces de jointes extraits de 279 comptes d'un ancienne société d'informatique dénommé Avocado. Plus récemment Janek Bevendorff et al [CITATION Bev20 \l 1036] ont publié cette année (2020) à ACL le corpus ***Webis Gmane Email*** qui est le plus grand corpus d'emails accessible au public et entièrement prétraité contenant plus 153 millions d'emails extraits de 14 669 listes de diffusions analysés et segmentés en composants sémantiquement cohérents à partir de nouveaux modèles neuronaux de segmentation. Les travaux de ce corpus se sont faits entre février et mai 2019 à partir de gmane.io qui couvrent plus de 20 ans de listes de diffusion publiques. Ce corpus contient 1,8 million de d'emails en Français. Le corpus CoMeRe [CITATION Ref14 \l 1036] qui est un corpus d'apprentissage d'Interactions Simuligne (Simulation en ligne en apprentissage des langues) quant 'à lui possède en 2033 emails en Français. Ces emails seront utilisés exploités lors de nos expérimentations.

En France la collection de ces différentes données fait face à de nombreuses contraintes liées à l'utilisation des données privées d'utilisateurs régies en 2016 par le ***Règlement Général sur la Protection des Données (RGPD)***. Dans un cadre un peu plus restreint comme celui d'une entreprise privée, il existe

des protocoles à suivre pour l'exploitation de telles données afin de respecter les données privées d'utilisateurs. Ces protocoles passent par des demandes de consentement, d'anonymisation et/ou de pseudoanonymisation que nous aborderons de façon plus détaillée dans le chapitre traitant des ***Corpus***.

Nous venons là de toucher aux différents aspects que nous allons approfondir dans les prochains chapitres. Le premier sera ainsi consacré à l'***analyse discursive*** dans lequel nous traiterons des éléments liés aux discours dans les ***conversations écrites***, surtout ***asynchrones.*** Le second quant à lui portera sur les Communications Médiées par Ordinateurs (CMO), leurs avantages et inconvénients, ainsi que les différentes formes qu'elles peuvent prendre : ***écrite, orale, synchrone*** et ***asynchrone***. Dans le troisième chapitre, les méthodologies de constitution et d'exploitation de corpus seront abordées avec les contraintes qui leurs sont associées. Par la suite nous aborderons des aspects plus techniques liés aux traitements des données de corpus. Ainsi le quatrième nous permettra de mettre en exergue les techniques de représentations des données textuelles par les méthodes ***occurentielles*** (de comptage) de mots jusqu'aux ***représentations contextuelles*** de contenus en passant par la prise en compte de la ***sémantique*** intégrée dans certaines de ces représentations. Les *algorithmes statistiques (Clustering* , *Arbres de décision*, *Classifieur bayesien naïf*, *Entropie Maximale*, *Induction de modèle structuré - HMMs CRFs*, *AdaBoost*, *Votes Perceptron)*, les modèles ***d'apprentissage automatique*** et ***profond*** (Réseaux de neurones **Feedforward – *FNN*,** Réseau de neurones récurrents – ***RNN*, *LSTM*, *GRU*, *Bi-LSTM***, réseau de neurones convolutif – ***CNN***, ***Encodeur-decodeur***, ***mécanismes d'attention*** et ***Transformer***) feront l'objet du chapitre cinq avec des détails sur les correspondances des données manipulées dans les différentes couches de ces modèles et les données brutes prises en entrée. Le chapitre six sera un complément au précédent parce qu'il proposera un aperçu détaillé des différentes méthodes d'évaluation de ces algorithmes. Le dernier chapitre, le septième, contiendra des descriptions de travaux pratiques ou d'expérimentation de certaines problématiques telles que la segmentation de texte et l'identification des intentions dans les emails. Et enfin nous terminerons sur une conclusion qui reprendra les contenus abordés dans ce document, fera une analyse objective et ciblée sur les besoins de la thèse en fonction des derniers travaux pilotés dans ce domaine d'analyse des conversations asynchrones notamment les emails, les forums et **chats**.

Chap 2. Analyse discursive

Introduction

Dans ce chapitre, il est question d'aborder l'analyse discursive ou plutôt l'analyse de discours. Pour plus de précision nous nous intéressons ici aux actes de discours qui sont nécessaires à l'étude des phénomènes conversationnels, à l'annotation de dialogues et même à la conception d'agent conversationnels.

Tout d'abord nous allons expliciter la notion de théorie d'actes de discours, son évolution au travers des travaux de Searle et Austin. Par la suite nous montrerons comment ces actes de dialogue sont utilisés ou appliqués dans l'analyse des conversations fonctionnelles, leur rôle sur les contextes conversationnels et leur influence sur les marqueurs pragmatiques. Enfin cette section se terminera par la description des principaux schémas d'annotations dont DAMSL, DIT++ et le standard ISO24617-2.

2.1 Actes de discours et de dialogues

2.1.1 Théorie des actes de Langage – *Premières taxonomies (Searle et Austin)*

Les pratiques linguistiques aujourd'hui déployées dans plusieurs domaines trouvent leur origine de la théorie des **actes de langages** ou **acte de discours.** Cette dernière stipule que le langage n'a pas que pour fonction essentielle de décrire le monde, mais aussi d'accomplir des actions.

Cette théorie a été initiée par un philosophe britannique au nom de **Austin** dans son ouvrage [CITATION AUS75 \n \l 1036]. Pour Austin, un acte de dialogue est un énoncé porteur d'une fonction performative. Les énoncés devraient être interprétés comme des actions sociales faites par des locuteurs. Ces actions sont ainsi portées par des verbes dits **performatifs** (« Je vous porterai secours »). Cependant fort est de constater que les actes de discours ne sont pas uniquement constitués de verbes performatifs.

En 1975, Austin développe une nouvelle théorie des actes de discours selon laquelle tout énoncé peut être analysé sur trois niveaux :

- Niveau ***locutoire*** : production d'une suite de sons, évoquant et reliant syntaxiquement les notions représentées par les mots.
- Niveau ***illocutoire*** : production d'un énoncé porteur d'intention rhétorique du locuteur
- Niveau ***perlocutoire*** : l'énonciation vise des effets plus lointains ou bien s'intéresse aux conséquences qui seront produites ou même de son interprétation par les allocutaires.

Ces trois niveaux peuvent respectivement se traduire par les questions, « que dit-il ? », « que fait-il ?» et « pour quoi faire ? ». De ces questions, il se dégage le second niveau qui est une des plus importantes qui est le niveau illocutoire, permettant de décrire les énoncés comme des fonctions communicatives (questions, réponses, remerciement…). Austin dans ses travaux propose cinq classes d'actes de discours regroupées dans le tableau ci-dessous.

Ces actes de langages sont aussi étudiés en profondeur par John R. Searle dans deux de ses ouvrages [CITATION Sea69 \n \l 1036] et [CITATION Sea79 \n \l 1036] traduits respectivement en **«Les actes de langage» (1972)** et **« Sens et expression » (1982).** Mais avant en 1969 Searle défend sa théorie selon laquelle tout acte de discours est ***illocutoire***. Cette théorie rejoint celle d'Austin sur ce que ce dernier appelle les « **actes de dialogue** ». Cette notion a été explicitée par Bunt (1994-1996) qui trouve trois aspects à ces actes dont leur forme de l'énoncé, leur fonction communicative et leur contenu sémantique. Pour Searle la fonction communicative des énoncés est des plus essentielles. Il a ainsi tout comme Austin proposé cinq classes d'actes (cf. tableau 2.2) qui ont des similitudes avec celles d'Austin.

Cette théorie d'actes de langage s'est vue au fil des années utilisée dans plusieurs domaines tels que la philosophie, la littérature et dans son domaine de base la linguistique où elle s'est d'autant développée pour

donner place à la pragmatique cognitive issue de la théorie de la pertinence de Sperber et Wilson. En informatique, la modélisation des conversations, les applications de dialogue homme-machine trouvent leur essence dans l'utilisation des actes de discours. Ces derniers sont donc à considérer avec beaucoup d'importance dans le cadre des analyses de conversations entre participants. Ce qui constitue le contenu de la prochaine section.

Austin (1975)	Searle (1976)	Quelques verbes
Expositifs (qui exposent de l'information)	**Assertifs** (qui affirment un état de fait)	affirmer, nier, postuler, remarquer…
Exercitifs (qui exercent un pouvoir)	**Directifs** (qui poussent l'interlocuteur à agir)	commander, conseiller, ordonner, pardonner, léguer…
Promissifs (qui engagent le locuteur)	**Promissifs** (qui engagent le locuteur)	promettre, faire vœu de, garantir, jurer de…
Comportatifs (qui expriment l'attitude)	**Expressifs** (qui expriment un état psychologique	s'excuser, remercier, féliciter, déplorer, critiquer…
Verdictifs (qui donnent un verdict)	**Déclaratifs** (qui ont un impact réel)	acquitter, condamner, décréter, baptiser

Tab 2.2 Taxonomies primaires de la théorie d'actes de langage/discours

-

2.1.2 Actes de langage /discours et de dialogues

Dans cette partie nous aborderons le passage des actes de langages aux actes de dialogues et comment ces derniers contribuent à l'analyse des conversations. Le contexte conversationnel sera aussi présenté avec l'influence sémantique de surface et sous-jacente qu'il peut permettre de percevoir sur des énoncés.

a) Analyse des conversations fonctionnelles

L'analyse par Vanderveken 1992 sur la limite des actes de discours largement répandue dans les précédentes années a montré que les actes de discours permettaient seulement de faire des études sur des énoncés de façon isolée ou indépendante de l'existence des autres énoncés lors d'échanges entre différents locuteurs par exemple. Cette limitation est d'autant marquée par Vanderveken sur l'atteinte d'un objectif commun par des participants à une conversation qui produisent des énoncés illocutoires. Cet objectif commun pouvant être l'accomplissement d'une mission ou la résolution d'un problème.

Les conversations fonctionnelles sont un grand ensemble de ce type de conversations ayant un but commun. L'extension des actes de discours pour la prise en compte de dépendances entre les énoncés a été faite par Traum & Hinkelman [CITATION TRA \n \l 1036] donnant lieu aux ***actes de dialogue*** ou ***actes de conversation***. Ces derniers sont très utilises dans la littérature pour modéliser les interactions communicationnelles entre les humains mais aussi entre les humains et les machines. Dans ces modélisations, l'un des aspects importants est le contexte qui contribue à différencier les actes de discours des actes de dialogue. Dans la prochaine sous-section, le contexte sera décrit sous l'angle d'interprétation des conversations.

b) Prise en compte de contexte

La théorie du ***contexte*** autour des conversations s'est développée de par des observations de Poesio et Traum [CITATION POE97 \n \l 1036] pour qui les conversations fonctionnelles possèdent des façades qui vont au-delà de l'atteinte d'un objectif. Pour eux déterminer une action coordonnée enfouie dans les énoncés est non négligeable. D'où les théories de contexte principalement utilisées pour représenter les effets des actes de discours sur les participants d'une conversation. Entre autres de ces effets on a les besoins, les obligations et les croyances. Pour Poesio et Traum, le contexte représente l'information sur laquelle les participants s'appuient pour interpréter les énoncées d'une conversation. Vu la polysémie que peuvent porter des énoncés d'une conversation, seul le ***contexte*** est le vecteur principal qui permet ainsi aux participants de converger vers une compréhension monosémique d'énoncés. Il est ainsi considéré comme un ensemble de connaissances communes et partagé par les participants. Ainsi donc, pour bien modéliser une conversation, il faut à tout niveau d'évolution de la conversation mettre le contexte à jour. Et donc pendant que les actes de discours s'attardent à capturer l'intention communicative du locuteur, les actes de dialogue vont plus loin à travers la prise en compte du contexte.

c) Marqueurs pragmatiques multimodaux

La pragmatique est un domaine vaste des sciences du langage qui s'intéresse aux éléments langagiers dont la signification ne peut être comprise qu'en connaissance du contexte de leur emploi. Les marqueurs pragmatiques (**MP**) sont une généralisation des marqueurs discursifs (**MD**) qu'on retrouve dans les conversations orales ou écrites. Les MP se divisent en deux grands groupes, les marqueurs pragmatiques verbaux (MPV) et les marqueurs pragmatiques non-verbaux(MPNV). Les MP, tout comme les actes de dialogue incorporent le contexte, ce qui n'est pas le cas avec les actes de discours. Les propriétés pragmatiques d'un énoncé peuvent s'identifier sous deux aspects : les ***actes expressifs*** (satisfaction, félicitation, remerciement, mécontentement, consternation, indignation, surprise) et les ***actes illocutoires*** (admettre une adéquation ou inadéquation d'un propos, d'une action, d'un comportement, approuver ou désapprouver ces derniers) que nous avons abordé plus haut. Dans le cadre de l'analyse des CMO, la problématique d'identification de pragmatique sous l'angle d'actes expressifs pourrait permettre d'avoir des informations émotionnelles des participants à une conversation lors de leurs interventions. Partant ainsi d'un glossaire de MVP et MPN, cette identification pourrait se mener facilement, mais reste la contrainte selon laquelle les pragmatiques sont fonction de leur ***contexte*** d'utilisation. Il en découle une fois de plus que la modélisation de contexte reste un verrou comme dans bien d'autres aspects d'analyse de conversations.

2.1.3 Schémas d'annotation

Dans cette sous-section nous présentons les schémas d'annotation utilisés pour représenter ou modéliser les conversations en actes de dialogues. Ces schémas d'annotation sont ***DAMSL, DIT++ et la norme ISO 24647-2,*** ils ont tous été développés sur le même corpus TRAINS. Avant d'expliciter les caractéristiques de chaque schéma d'annotation, nous allons présenter les principes qu'ils ont en commun.

- La **mise à jour du contexte** : les trois schémas d'annotation prennent en compte l'évolution du contexte commun qu'ont les participants lors d'une conversation. C'est pour cette raison qu'ils proposent d'annoter les fonctions communicatives des actes de dialogues, ces fonctions qui selon Core & Allen [CITATION Cor97 \n \l 1036] devraient représenter de façon directe le contexte évolutif d'une conversation.

- La **multi-dimensionnalité** : la principale limite identifiée dans les taxonomies d'Austin et Searle est la prise en compte du fait qu'un même énoncé peut exprimer plus d'une intention d'un locuteur. Par exemple dans une conversation, un locuteur peut détailler son propos dans un énoncé en posant une question à un autre allocutaire. C'est la **muti-dimensionnalité** d'un énoncé. Core, Allen et Bunt ont donc pris en compte cette

problématique et permettent par leur schéma d'annoter des énoncés sur plusieurs couches (layers).

- La **généricité** : elle se décèle des schémas d'annotation notamment DAMSL et DIT++ et même la norme ISO 24617-2 (largement basé sur DIT++). Ces schémas permettent en fait d'annoter les conversations en actes de dialogue suivant deux approches dont une ontologique (spécifique au domaine) et une autre pour une couverture plus large : d'où la notion de généricité.

a) DAMSL

Dialogue Act Markup in Several Layers (DAMSL) en français balisage d'actes de dialogues en plusieurs couches a été le premier schéma d'annotations permettant d'assigner de multiples labels à des énoncés de conversations. Développé par Core & Allen [CITATION Cor97 \n \t \l 1036], DAMSL est constitué de quatre principales couches et de plusieurs dimensions comme le montre la figure (FIG 1) ci-dessous. Les super-couches de DAMSL sont les fonctions prospectives, les fonctions rétrospectives, le niveau d'information et le statut communicatif. Les deux premières permettent d'annoter les énoncés en fonction de leur intention communicative. Les deux autres indiquent sur quoi portent les énoncés, ce sont des catégories du niveau informationnel. Les dimensions sur la figure ci-dessous correspondent aux premiers niveaux en dessous des super-couches, elles sont indépendantes les unes des autres et optionnelles.

Fonctions rétrospectives :
- *Agreement*
  - *Accept*
  - *Accept-Part*
  - *Maybe*
  - *Reject-Part*
  - *Reject*
  - *Hold*
- *Understanding*
  - *Signal-Non-Understanding*
  - *Signal-Understanding*
    - *Acknowledge*
    - *Repeat-Rephrase*
    - *Completion*
  - *Correct-Misspeaking*
- *Answer*
- *Information-Relation*

Fonctions prospectives :
- *Statement*
  - *Assert*
  - *Reassert*
  - *Other-Statement*
- *Influencing Addressee Future Action*
  - *Open-Option*
  - *Directive*
    - *Info-Request*
    - *Action-Directive*
- *Committing Speaker Future Action*
  - *Offer*
  - *Commit*
- *Performative*
- *Other Forward Function*

Niveau d'information :
- *Task*
- *Task Management*
- *Communication Management*
- *Other*

Statut communicatif :
- *Abandoned*
- *Uninterpretable*
- *Self-talk*

Fig 1. Taxonomies DAMSL

Plusieurs travaux sur l'analyse des conversations avec annotations en actes de dialogue s'appuient sur la taxonomie DAMSL d'où son importance. Bunt entre 1989 et 1994 porte des critiques sur les éléments de DAMSL qui selon lui manque de signification conceptuelle et ne s'appuient pas sur des fondements théoriques. C'est ainsi qu'il propose le schéma DIT++ qui possèdent des bases théorique assez solides.

b) DIT^{++}

Cette taxonomie est une extension de la taxonomie DIT(Dynamic Interpretation Theory) qui avait été développée par Bunt(1989, 1994) avec plusieurs types d'actes de dialogue de DAMSL et d'autres études sur les actes de dialogue. DIT++ est un Framework sémantique développé pour l'analyse des conversations entre humains mais aussi entre hommes et machines. Il permet aussi d'annoter en actes de dialogues des segments de texte contenant des fonctions communicatives. Cette taxonomie est constituée de :

1. d'une taxonomie compréhensible multidimensionnelle de fonctions communicatives qui sont sémantiquement définies sur plusieurs états de changements d'informations
2. d'une définition de 10 dimensions orthogonales aux quelles devrait appartenir un acte de dialogue; ces dimensions offrent une base de compréhension de la multifonctionnalité des énoncés de dialogue.
3. d'une définition de plusieurs types de relations sémantiques et pragmatiques entre les actes de dialogue
4. d'un petit ensemble de qualificatifs permettant d'indiquer l'incertitude, la réserve ou les sentiments d'un interlocuteur.

La taxonomie DIT++ a beaucoup évolué jusqu'en 2009 via les travaux de Bunt. Cette évolution a donné lieu a des caractéristiques plus solides tant sur ses dimensions que sur le plan de ses différentes fonctions communicatives. Ces caractéristiques sont présentées dans les trois images ci-dessous :

1. *Task/Activity*, pour tout ce qui se rapporte à la tâche qui est l'objet de la conversation ;
2. *Auto-Feedback*, pour les actes signifiant le niveau de compréhension et d'interprétation du locuteur ;
3. *Allo-Feedback*, *idem* pour l'allocutaire ;
4. *Turn Management*, pour les actes portant sur la gestion du tour de parole ;
5. *Time Management*, pour les situations où il est nécessaire de signifier que le locuteur a besoin de plus de temps pour contribuer ou qu'il faut faire une pause ;
6. *Contact Management*, pour les actes qui servent à établir et maintenir la communication ;
7. *Own Communication Management*, pour les actes servant à indiquer que le locuteur prépare ou modifie sa contribution au dialogue ;
8. *Partner Communication Management*, pour les actes effectués par un participant endossant le rôle d'allocutaire, servant à assister son partenaire dans la formulation de sa contribution ;
9. *Discourse Structure Management*, pour les actes servant à structurer thématiquement la conversation ;
10. *Social Obligations Management*, pour les actes de gestion sociale du dialogue.

Fig 2. Dimensions retenues par [CITATION Har09 \l 1036]

Dimension	Exemples de fonction
Task / Activity	*Open Meeting, Appoint, Hire*
Auto-Feedback	*Perception Negative, Evaluation Positive*
Allo-Feedback	*Interpretation Negative, Evaluation Elicitation*
Turn Management	*Turn Grab, Turn Take, Turn Keep*
Time Management	*Stalling, Pausing*
Contact Management	*Contact Check, Contact Indication*
Own Communication Management	*Self-Correction*
Partner Communication Management	*Completion, Correct Misspeaking*
Discourse Structure Management	*Opening, Topic Introduction*
Social Obligations Management	*Return Greeting, Apology, Thanking*

Fig 3. Exemples de fonctions communicatives spécifiques de DIT++

En 2012, la taxonomie DIT^{++} a été utilisé comme principal socle d'un standard international pour l'annotation dialogique : ISO 24617-2 [CITATION Bun12 \l 1033]

c) Norme ISO 24617-2

En 2012, Bunt et al. [CITATION Bun12 \n \y \l 1033] mettent sur pied la norme ***ISO 24617-2***. Cette norme est en fait un framework d'annotations sémantiques, qui n'est rien d'autre qu'une mise à jour d'une version antérieure ***ISO DIS 24617-2:2010*** développée par Bunt et al. en 2010 [CITATION Bun10 \n \y \l 1036]. Pour passer de cette dernière à au nouveau standard, des concepts d'annotation de ***relations rhétoriques***, de ***dépendances fonctionnelles*** et ***de feedback*** entre des unités de dialogue y ont été ajoutés. Le langage basé sur XML, ***DiAML*** (Dialog Act Markup Language) a ainsi subit une restructuration ou utilisation différentes des certains de ses éléments et attributs dans la norme ***ISO 24617-2*** pour la prise en compte de ces nouveaux concepts.

Avant d'aborder de façon un peu plus consistante les différents ajouts qui ont contribués à la nouvelle norme, nous allons ci-dessous présenter les principaux caractéristiques de la version ***ISO DIS 24617-2:2010 :***

1. Les aspects de dimensions sémantiques dans l'analyse d'actes de dialogue y sont incorporés et ces dimensions sont au nombre de 9 (cf. fig.2). Ce sont en fait celles que Bunt a défini dans la taxonomie DIT++. La seule différence est que la dimension *«Contact Management»* qui permettait d'annoter les actes d'établissement et de maintient de communication, a été supprimée et sa fonction a été incorporée dans la dimension *«Discourse Structuring»* qui elle permet d'annoter les actes de gestion de sujet, ouverture et fermeture des dialogues (sous-dialogues aussi) et à les structurer. Cette fonction de structuration était la seule de la dimension *«Discourse Structuring»* dans DIT++. Ces dimensions se distinguent les unes des autres sur les fondements empiriques et théoriques et permettent ainsi d'annoter de façon de multidimensionnelle, c'est-à-dire qu'un segment de dialogue peut être annoté par plus d'une de ces dimensions sans ambiguïté.

2. Deux classes de fonctions communicatives dont une spécifique à une dimension et l'autre d'un usage plus générale constituent aussi l'une des caractéristiques de cette norme. Ces fonctions peuvent ainsi être combinées avec n'importe quel contenu sémantique et donc former un acte de dialogue dans la dimension correspondante.
3. Des ***fonctions qualificatives*** définies pour spécifier si un énoncé (acte de dialogue) a été effectué de façon conditionnel ou non, avec certitude ou pas ou bien avec un sentiment particulier.
4. Les relations de dépendance fonctionnelle ou de feedback sont définies pour mettre en relation des actes de dialogue à des segments préalablement identifiés. Rendant ainsi plus explicite l'association d'une réponse à une question en amont ou bien d'un feedback à l'énoncé d'un interlocuteur.
5. La notion de ***segment fonctionnel*** est utilisée comme unité d'annotation d'acte de dialogue. C'est l'entité minimale comportementale possédant une ou plusieurs fonctions communicatives.
6. La segmentation multifonctionnelle est appliquée, montrant ainsi la distinction que peut avoir chaque dimension sur un segment fonctionnel. Un segment portant une fonction de feedback peut se chevaucher avec un segment qui zest une fonction liée à une tâche.
7. DiAML est tout aussi représenté sous trois angles (i) une syntaxe abstraite qui spécifie les annotations possibles avec un ensemble de termes théoriques,(ii) une sémantique qui spécifie les interprétations des structures définies par la syntaxe abstraite, (iii) une syntaxe concrète fixant une représentation XML des structures d' annotations.

Tous ces aspects sont représentés dans le méta-model d'annotation d'actes de dialogue ci-dessous.

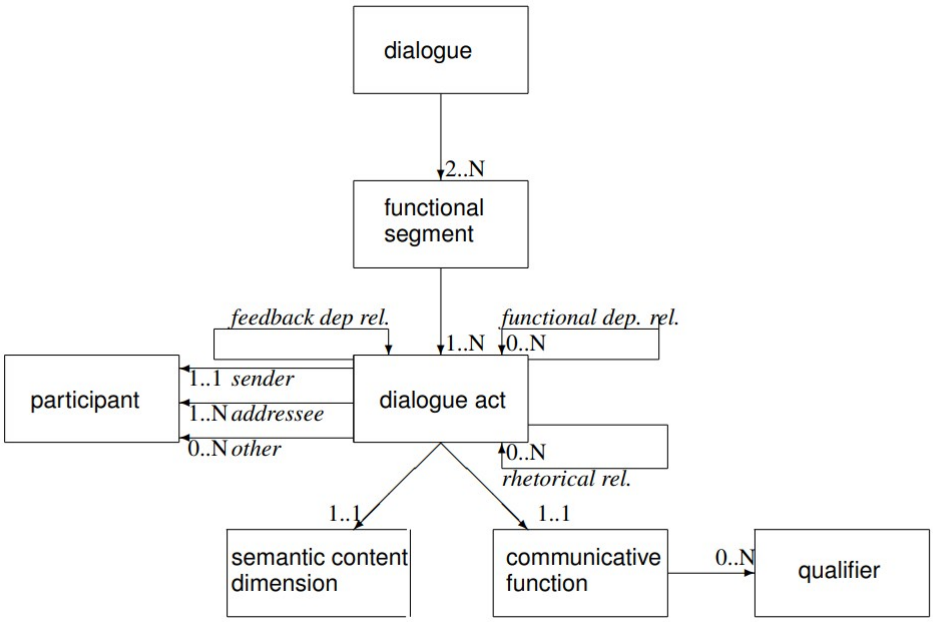

Fig 4. Métamodel d'annotation d'actes de dialogue de [CITATION Bun12 \t \l 1033]

A la différence du standard ***ISO DIS 24617-2:2010,*** le méta-modèle de la figure ci-dessus fait ressortir de nouveaux concepts tels que les ***relations rhétoriques*** (qui n'existaient pas dans ***ISO DIS 24617-2:2010***) entre actes de dialogues. Ces relations sont plus marquées dans les textes écrites, puisqu'elles ont été

intensivement étudiées dans ce contexte avant de se rapporter au dialogue. Dans ces études passées [CITATION Hob85 \l 1036], [CITATION Man88 \l 1036] et [CITATION Pra08 \l 1036], ces relations rhétoriques ont des appellations différentes : ***relations de cohérence*** ou ***relations de discours.***

Dans un dialogue (entre deux personnes) qui se rapproche facilement à une conversation écrite (tchats ou échange d'emails), on peut distinguer deux types de relations rhétoriques :

i. Relations rhétoriques de dépendance fonctionnelle : comme exemple une réponse d'un interlocuteur à une question rhétorique dans un énoncé antérieur à cette réponse. Ceci pouvant s'effectuer de façon bidirectionnelle entre deux intervenants ou dans le cadre d'une conversation multipartite (forum).
ii. Relations rhétoriques de dépendance de feedback : Celles –ci se matérialisent par une relation entre un énoncé du type ***«sûr»*** et un ***hochement de tête*** d'un autre interlocuteur. Ce type de relation est plus visible en communication parlée, mais avec l'avènement des émoticons/émojis, on les décèle désormais en conversations écrites. La relation de dépendance de feedback peut aussi se manifester par une explicitation d'un concept ou enrichissement d'une information suite à un énoncé du genre ***«suis pas sûr d'avoir bien compris».***

Au travers du schéma de description XML de **DiAML**, ces relations rhétoriques sont représentées avec l'élément ***« rethoricaLink »*** qui possède trois attributs dont les deux premiers font référence à des actes de dialogue identifiés et la dernière ***@rethoRel*** ayant pour valeur le type de relation (dépendance fonctionnelle ou de feedback) entre les deux actes de dialogues référencés. Le standard ISO 24617-2 n'inclut pas un ensemble de type de relation rhétorique, il laisse le libre choix suivant le contexte de la conversation et l'objectif de l'annotation.

À contrario des deux types de relations rhétoriques mentionnés ci-dessus, les relations de dépendance fonctionnelle et de feedback existaient déjà avec le standard ***ISO DIS 24617-2:2010.***

Dans cet ancien paradigme, les relations de dépendance fonctionnelle **(DF)** surviennent avec des actes de dialogue de nature réactive tels que des réponses, confirmations, accords, acceptations, excuses, refus ; les contenus sémantiques de ces types d'acte de dialogue dépendent fortement des précédents actes auxquelles ils répondent. C'est pour cette raison qu'ils peuvent s'exprimer avec des énoncés (oui, non merci, ok) qui eux ne possèdent aucun contenu sémantique. Avec ce standard ***ISO DIS 24617-2:2010,*** ces relations de dépendance fonctionnelle étaient représentées avec ***DiAML*** par un élément ***functionalLink*** avec deux attributs : ***@dact*** dont la valeur dans le cadre d'une relation DF question/réponse est la référence de la réponse et ***@functionalAntecedant*** dont la valeur possède l'index de la question. La partie gauche de la figure 5 explicite cet encodage. Cependant dans le standard ***ISO 24617-2*** cette relation est directement intégrée dans l'acte de dialogue qui répond à la question. Ceci se fait grâce l'attribut ***@functionalDependence*** qui prend comme valeur l'index de l'acte de l'acte de dialogue de la question.

Si l'on considère la conversation ci-dessous :

1. C Y-a-t-il une voiture disponible?
2. A: Non, désolé, il n'y en a plus.

Les annotations ci-dessous des actes contenus dans cette conversation représentent à gauche celle avec standard ***ISO DIS 24617-2:2010 et*** à droite celle de ***ISO 24617-2.***

```
<diaml xmlns:
  "http://www.iso.org/diaml/"/>
<dialogueAct xml:id="e1"
  target="#fs1"
  sender="#c"
  addressee="#a"
  communicativeFunction=
    "propositionalQuestion"
  dimension="task"/>
<dialogueAct xml:id="e2"
  target="#fs2"
  sender="#a"
  addressee="#c"
  communicativeFunction="answer"
  dimension="task"/>
<functionalLink
  dact="#e2"
  functionalAntecedent="#e1"/>
<dialogueAct xml:id="e3"
  target="#fs3"
  sender="#a"
  addressee="#c"
  communicativeFunction="apology"
  dimension="social obligations"/>
</diaml>
```

ISO DIS 24617-2:2010

```
<diaml xmlns:
  "http://www.iso.org/diaml/"/>
<dialogueAct xml:id="e1" target="#fs1"
  sender="#c" addressee="#a"
  communicativeFunction="prop.Question"
  dimension="task"/>
<dialogueAct xml:id="e2" target="#fs2"
  sender="#a" addressee="#c"
  communicativeFunction="answer"
  dimension="task"
  functionalDependence="#e1"/>
<dialogueAct xml:id="e3" target="#fs3"
  sender="#a" addressee="#c"
  communicativeFunction="apology"
  dimension="social obligations"/>
</diaml>
```

ISO 24617-2

Fig. 5 Exemple d'annotation d'actes de dialogue

De façon identique aux relations de DF, les relations de dépendance de feedback prennent en compte le contexte sémantique des précédents énoncés auxquelles ils sont liés. Ils sont tout aussi représentées avec DiAML dans le standard ***ISO 24617-2*** avec l'attribut ***@feedbackDependence*** avec comme valeur l'index de l'acte de dialogue auquel fait un retour l'acte content cet attribut.

Un autre concept intégré dans le standard ***ISO 24617-2*** est celui des ***qualificatifs*** qui sont applicables dans le cadre des fonctions communicatives à objectifs généraux (General Purpose communicative Functions – GPFs). On distingue différents types de qualificatifs qui sont applicables à différents groupes de fonctions communicatives.

i. Les qualificatifs sentimentaux qui s'appliquent à tous les GPFs.
ii. Les qualificatifs conditionnels s'appliquant à des GPFs d'actions-discussions telles que les promesses, offres, suggestions, acceptations, déclinaisons, acceptations de suggestions.
iii. Les qualificatifs de in/certitudes qui s'établissent sur des fonctions de partage d'informations (informer, accords, désaccords, corrections, réponses, non/confirmations

Les fonctions communicatives spécifiques à des dimensions s'expriment par des formules (***salut, bonjour, okay, oui, je te présente mes excuses, au revoir, …***) et n'admettent pas de qualificatifs.

Les qualificatifs que nous venons de présenter sont bien pris en compte lors de l'annotation des actes de dialogue par des attributs comme : uncertainty, certainty, conditionaly… pour ne citer que ceux-ci.

Tout récemment, une mise à jour du standard ***ISO 24617-2*** baptisée ***ISO 24617-4*** [CITATION Har20 \l 1036] a vu le jour avec des améliorations notamment au niveau de la précision des annotations des relations de dépendances rhétoriques dans les dialogues. Ce nouveau standard intègre des emotionML(Emotion Markup Language) qui comme son nom l'indique est un langage de balisage d'émotions permettant d'annoter et de représenter des émotions dans des contextes écrits ou parlés. ***ISO 24617-4*** inclut aussi un plugin à trois couches permettant l'enrichissement de la description des actes de dialogue avec des contenus sémantiques couplés à des émotions et bien d'autres informations. Tout ceci contribuant à améliorer les annotations de corpus et la conception de composants pour les systèmes de dialogue parlé et multimodal.

Conclusion

Dans ce chapitre, nous avons parcouru l'analyse discursive, partant des premières taxonomies d'actes de langages jusqu'au standard ***ISO 24617-4*** qui permettent d'annoter des contenus écrits ou de paroles transcrites. Ces taxonomies et standard ont aussi permis de mieux cerner et capturer les différentes caractéristiques linguistiques dans des conversations. Ceci au travers des différentes couches, dimensions et relations de dépendances entre elles, identifiées et conciliées dans ces taxonomies et standard. Tous ces travaux donnant ainsi la possibilité d'implémenter des systèmes automatiques de dialogues ou de conversations (chabots). Mais aussi les standards les plus récents ***ISO 24617-2*** et ***ISO 24617-4*** poussent la connaissance sur ces caractéristiques jusqu'à la prise en compte des émotions des participants à une conversion en fonction de leurs interventions. De la même façon les relations rhétoriques entre les interventions des interlocuteurs sont aussi décelées. Force est de constater que la majorité des briques autour de l'analyse des discours a été abordé depuis quelques décennies jusqu'à nos jours. Et depuis leur prémices, ces éléments de l'analyse discursive sont utilisées dans différents projets de recherches, d'annotation des corpus à la conception d'applications de communication intelligente et autonome (Djingo, Siri, Alexa, …). Plusieurs autres aspects restent néanmoins pas totalement explorés ou bien n'atteignent pas encore des performances souhaitées. C'est par exemple le cas pour notre problématique de décomposition de conversation d'emails en sous-conversations. Ainsi malgré les nombreux travaux faits sur l'analyse de conversations, il ne reste tout de même pas évident même avec les évolutions des nouveaux modèles de traitement automatiques de répondre de façon précise à une problématique telle que la nôtre. Cependant, afin de mieux approcher notre problématique, les différents aspects de l'analyse discursive abordés dans ce chapitre nous seront très utiles en première ligne notamment pour annoter des corpus ou bien pour catégoriser des segments de textes d'emails en actes de dialogues comme dans certains travaux [CITATION Mic111 \l 1036], [CITATION Jon19 \l 1036], [CITATION Wil04 \l 1036].

Mais avant de pouvoir analyser ou classer quoi que ce soit, nous allons dans le chapitre à venir nous attarder sur les différents types de communications qui ont émergés ces dernières années qu'on qualifie de communications médiées par les ordinateurs.

Chap3. Communication médiée par ordinateur -CMO (Computer Mediated Communicaiton)

1 Introduction

Dans le début des années de 1980, le concept de communication médiée par ordinateur (CMO) a vu jour dans le milieu universitaire, désignant l'ensemble des modalités de communication qui s'effectue via des machine. L'évolution de cette notion a poussé à reconsidérer l'ordinateur comme un médium de communication plutôt que comme un outil.

De nos jours, faire allusion à ce concept revient à s'intéresser un domaine très vaste et complexe, parce qu'il inclut non seulement les interactions homme-machine dotées de nouvelles technologie, mais aussi des multiples aspects langagiers utilisés lors des échanges. Que ces derniers soient oraux ou écrits, synchrones ou asynchrones, ils présentent tous des avantages et des inconvénients.

1.1 Avantages et inconvénients

Il existe de plus en plus de multiples applications et outils autour des CMO que nous aborderons de façon plus explicite dans les deux prochaines parties. Nous notons toutefois des avantages accrues que nous procurent ces outils de communications. On peut citer entre autres leur utilisation pour rester en contact facile avec nos proches dans le monde entier, le partage d'informations, la croissance économique de multiples entreprises passent aujourd'hui par des canaux de CMO, ces mêmes canaux ont montrés leur grande puissance dans le maintien des communications suite à la crise sanitaire récente. Sur le plan éducationnel et d'apprentissage, les CMO au travers des plateformes d'e-learning, des moocs permettent de perpétuer les enseignements dans multiples domaines scientifiques ou littéraires. Un autre avantage est la production des données utiles à l'amélioration de l'application de CMO via les technologies d'intelligence artificielle.

Cependant les inconvénients existent tout aussi avec les CMO. On peut noter dans le cadre du partage d'informations la propagation des fake news, le piratage d'informations et demande de rançon par des techniques allant du simple hameçonnage au cyber attaques. On a aussi enregistré sur les réseaux sociaux ces dernières des harcèlements de tout genre et même des groupes de radicalisation. Un autre aspect négatif est l'addiction distractive à certains de ces médiums par des personnes impactant ainsi leur productivité pour des travailleurs par exemple.

Sur ces désavantages de plus en plus marquées et en constance croissance, des mesures de sécurité et même des cellules de lutte contre les cybers attaques ont été mises sur pied au niveau des Etats et même sur le plan international. Des nouvelles technologies de sécurité sont de plus en plus développées dans le cadre de la protection de ces données privées d'utilisateurs ou d'organismes qui peuvent se présenter soit sous forme orale ou écrite, types que nous allons étudier par la suite.

1.2 Communication orale, communication écrite

En termes de communication, on dénombre aujourd'hui trois principales façons via la parole, les écrits ou les signes. Nous nous intéressons ici aux deux premières parce ce qu'elles sont facilement utilisées via des outils de CMO. L'oral est un discours en interaction exprimé et transmis de vive voix et véhiculé par la parole. Par contre l'écrit est consigné à travers des graphies présentes dans le temps et dans l'espace. Ces deux moyens de communication présentent des caractéristiques langagières et linguistiques qui les distinguent fortement. Toutefois, la présence physique, la distance et le temps sont les principales contraintes expliquant la différence linguistique entre l'oral et l'écrit.

L'oral possède des spécificités qui le différencient clairement de l'écrit. On distingue entre autres la spontanéité, la prosodie (pauses, accents d'instance, intonation, débit, etc…), les liaisons, les enchaînements vocaliques, les fréquences de signaux de régularisation, les accents et la disfluence (hésitations, amorces, constructions interrompus, etc.). Ces différents traits permettent de doter les discours oraux de fonctions précises, certains (disfluence) impactent fortement la syntaxe de l'oral lui donnant notamment un aspect disloqué.

Au-delà de ces aspects que nous venons d'évoquer, notons que des linguistes émérites ont travaillés sur la question de différences et ou rapprochement entre les communications écrites et parlées. C'est ainsi que Halliday, M. [CITATION Hal85 \n \l 1036], Michael McCarthy [CITATION Mic91 \n \l 1036], Vivian Cook [CITATION Viv04 \n \l 1036] et bien d'autres ont tenté de mettre en place des caractéristiques du ***langage écrit*** comme une thématique séparée du ***langage parlé.*** Cependant, ils sont tous arrivés à ola conclusion selon laquelle ces deux types de langage sont interdépendants du fait que le langage parlé nécessite l'écrit et vice-versa. Par le passé certains linguistes ont suggéré que l'écrit était juste une façon de d'enregistrer les paroles [CITATION Leo33 \l 1036], aussi De Saussure selon définissais l'objectif de l'écrit comme une représentation du langage oral.

Des opinions récentes stipulent que les communications écrites n'ont jamais été et ne seront jamais une façon de poser ou de transcrire de la parole. Ceci peut très bien se voir sur des annotations de conversations parlés qui sont très distantes d'un texte écrits (ouvrages, rapport, romans, etc…), de par l'ensemble des éléments linguistiques utilisés pour ces annotations. De même l'invention du magnétophone a permis de mieux cerner ce qu'était finalement un enregistrement de paroles. Et donc retranscrire de la parole serait à des fins d'analyses de discours ou conversations parlés.

En complément des paragraphes précédent qui présentent les différences entre la parole et l'écrit, il existe bien d'autre aspect de différenciation de ces modalités de communication. Ce sont ceux de leur typage fonctionnel, de leurs caractéristiques linguistiques.

Sur l'aspect fonctionnel, la parole est très fréquemment utilisée tous les jours pour communiquer avec notre entourage, ce qui n'est pas le cas avec l'écrit. Cette assertion reste tout de même mitigée depuis l'année 2020 marquée par la crise sanitaire et les différents confinement et couvre-feux instaurés dans plusieurs pays, limitant ainsi les interactions physiques entre des personnes et réduisant très fortement par la même occasion la fréquence d'utilisation de la parole. Il ressort clairement de cette petite analyse que la fonction principale de la parole est le maintien des relations sociales, mais aussi des interactions entre individus. Ce type de fonction de fonction est qualifié de fonction ***phatique*** prédominante pour la parole. Lors d'un échange entre des personnes via la parole en présentiel, certains signaux non-verbaux ou caractéristiques paralinguistiques sont perceptibles. Ce sont par exemple l'intonation, le rythme de la parole (pitch, vitesse, silence, etc.), le contact visuel, la gestuelle, les expressions faciales d'où la multi-modalité de la parole. Ces éléments non-verbaux laissent transparaitre les attitudes et émotions des interlocuteurs mettant ainsi en avant la fonction ***émotive*** et/ou ***expressive*** de la parole. A propos de l'écrit, il est majoritairement utilisé pour la transmission des connaissances et des informations, raison pour laquelle on lui attribue la fonction ***référentielle.*** La fonction ***expressive*** fait aussi partir des fonctions de l'écrit de par la prosodie qui y est incluse.

Concernant les caractéristiques linguistiques, la différence entre la parole et l'écrit peuvent se faire sur plusieurs pans ci-dessous :

- Prosodie (ou traits suprasegmentaux) et la ponctuation : la segmentation de la parole s'identifie premièrement par les critères paralinguistiques tels que les pauses, l'intonation et le rythme de la parole. L'écrit quant 'à lui utilise les ponctuations et les paragraphes.

- Compréhension et incompréhension : En général des interlocuteurs impliqués dans un échange partagent un grand champ contextuel et ceci parce qu'ils ont l'opportunité d'avoir des retours immédiats et de cette façon des incompréhensions peuvent être évitées. Cependant le langage conversationnel est souvent peu explicite et vague, caractérisé par une haute fréquence d'expressions déictiques et de pronons. Mais aussi d'une basse fréquence d'utilisation des noms. Par contre l'écrit ne peut s'appuyer sur une contextualisation parce qu'il n'y pas de possibilités d'avoir des feedbacks intermédiaires. Ainsi l'écrit tend à être précis et explicite et est caractérisé par une moindre utilisation d'expressions déictiques.

- Densité lexicale: Sur cet aspect, l'écrit est largement plus dense contrairement à la parole.

- Interactivité : De par la fonction prédominante phatique de la parole et peu représenté dans l'écrit, il ressort très facilement que l'interactivité est beaucoup plus marqué lors de l'utilisation de la parole.

- In/Formel : En tenant compte des traits externes de la parole et de l'écrit, il n'y a point de doute que l'écrit reste très formel au niveau de l'élaboration des structures syntaxiques afin de produire des phrases complexes. Ce qui n'est pas le cas de la parole qui est utilisée par des interlocuteurs au travers des phrases courtes et simples et des fois pas très structurées syntaxiquement. L'élaboration de structure syntaxique lorsqu'on parle aboutit des fois à des incompréhensions. La parole est fortement caractérisée de répétitions prépondérantes, des mots d'argots et contractions, tout ceci montre bien le caractère informel de la parole.

Ces précédents points présentent sur plusieurs aspects linguistiques et paralinguistiques les différences entre les médiums de communication ***écrit*** et ***parole***. Ces caractéristiques ici présentées ont pour source des travaux sur les emails [1].

1.3 Communication synchrones et asynchrones

Au-delà des communications écrites et orales, il existe d'autres façons de communiques via des Emails, SMS, Whatsapp, Messenger, Slack, etc … toutes des applications nous permettant d'échanger de l'information soit de façon synchrone ou asynchrone. Le choix de synchronicité ou non est importante parce qu'entrainant le gain ou une perte de temps.

La communication synchrone est une communication temps réel entre deux ou plusieurs interlocuteurs, les échanges y sont directs et instantanées. La conversation en face à face, les réunions, appels, visioconférences ou messageries instantanées sont des communications synchrones. Ce type de communications a l'avantage d'être riche (dans une conversation en présentielle, plusieurs informations non-verbales sont échangées), rapide et plus humain du fait de la transmission direct de ses émotions qui peuvent être perceptible même sur une conversation téléphonique. L'abondance d'interruptions et la perte de liberté sont les principaux inconvénients de la communication synchrone.

La communication asynchrone quant 'à elle est une communication qui se déroule en différé, les contraintes spatiales ou temporelles sont inexistantes. Les locuteurs décident eux-mêmes quand et où échangé. La majorité des outils de communications professionnelles, emails, Skype, Slack, Workplace permettent la communication asynchrone caractérisée par la présence d'un délai plus ou moins long entre le moment où l'information est été émise et le moment où elle est reçue. Toutefois il existe certains canaux de

[1] Language of the Emails: Forms of Spoken and Written Language by Kateřina Pardubová

communications qui sont considérés pour des communications synchrones et asynchrones. Prenons par exemple l'exemple des SMS ou messagerie instantané, lors de la réception d'un message via ces canaux, on peut répondre sur le coup suivant qu'on soit disponible pour le faire ou bien plus sinon. Les avantages liés à la communication asynchrones sont nombreux, on distingue entre autres une plus grande liberté, un meilleur contrôle de votre temps, le contrôle de l'information reçue, le respect du temps des autres. Cependant les communications asynchrones sont moins riches et lentes.

En général dans les entreprise ces deux modes de communications sont très utilisées, mais pour notre étude nous allons nous focalisé sur les communications asynchrones. Les courriels, forums et tchats vont être présentés dans les prochains paragraphes.

2 Caractéristiques des canaux étudiés

2.1 ***Courriels*** **-** *définition, caractéristiques et actions, structuration et principaux travaux, métadonnées*

D'origine québécoise dans les années 90, le mot « courriel » est une contraction de « courrier électronique » et a été accepté par l'Académie française et redu obligatoire en France pour les textes officiels depuis le 20 juin 2003. Le courriel est donc un courrier électronique destiné à un tiers ou plusieurs pouvant contenir des messages de différentes natures (travail, publicité, loisirs…) qui transite par le biais d'une connexion à un réseau informatique. C'est une traduction du terme anglais « Email » qui est souvent très utilisé à sa place.

D'après Radicati Group, février 2019, on dénombre plus de la moitié de la population mondiale (3.9 milliards de personne) comme utilisateurs de courriels et la prévision pour 2023 situe ce nombre à environ 4.3 milliards. D'après Médiamétrie, janvier 2019, en France on a dénombré 42.2 millions en 2019 Français se connectant par mois à un webmail et 22.7 millions se connectant à au moins un compte mail chaque jour. Pour Map global Provider Report octobre 2019, les principales messageries utilisées en France par ordre décroissant sont : Gmail (27 %), Outlook.com / Hotmail (26%), Orange (18%), Yahoo mail (12%) et SFR (6%). Dans le monde de l'entreprise d'après une étude d'Adobe en août 2015, des cadres estiment passer plus de 5 heures par jour en moyenne à consulter leur messagerie, en France leur estimation était autour de 5,6 heures et 5.4 en Europe. Aux USA ce chiffre monte à 6,3 heures. Au vue de ces chiffres, on peut en effet s'imaginer la grande quantité d'informations produite par les courriels et aussi l'urgence d'optimiser ces temps de consultation de messagerie par des cadres en facilitant la recherche d'informations par exemple. Ceci étant un des de nos travaux. Cependant pour atteindre un tel objectif et bien d'autres, il faut s'intéresser aux différents caractéristiques des courriels, tant sur les aspects linguistiques, paralinguistiques, lexicaux et syntaxiques que structurels.

a) Structure d'un courriel

Les courriels sont apparus il y a quelques décennies et ont progressivement remplacé depuis lors les méthodes traditionnelles de communication qui consistaient à l'envoi de lettres et mémos. Et donc ils ont gardé la structure globale de leurs prédécesseurs, qui était constituait d'un en-tête et d'un corps.

En-tête : Cette partie peut être vue comme l'extérieur d'une enveloppe physique. La différence entre les deux se trouve au niveau des informations renseignées. Alors qu'une enveloppe physique ne possède que les adresses de l'expéditeur et du dentinaire avec souvent la date d'envoi et un petit message, l'en-tête d'un email aura en plus un sujet, une copie carbone (Cc. :), une copie carbone invisible (Cci. :), une date d'expédition et de réception. Ces parties d'un en-tête se retrouvent sous différents formes dépendant du logiciel client utilisé (Outlook, eMClient, etc.).

Corps : Le corps d'un email est généralement subdivisé en quatre sections : ***salutation***, ***contenu principal***, ***partie au revoir et signature.*** Ces parties ont été identifiés dans notre corpus d'emails actuellement en

constitution (février-mars 2021) pour nos travaux. [CITATION Dav01 \l 1033] dans ses travaux identifie tout aussi ces sections excepté la signature dans le corps d'un email. Mais de façon général et après multiples observations sur nos corpus et s'appuyant sur les travaux Crystal sur le langage des d'emails, on peut dire que la partie *au revoir* et *signature* ne forment en fait une seule et même partie. La signature est surtout identité pas le nom, prénom ou initiales que l'expéditeur renseignent après une formule d'au revoir et qui marquent la fin d'un email. Cependant certaines applications clients d'emails nous permettent de configurer une signature contenant des toutes ou parties des informations suivantes : nom, prénoms, numéro de téléphone, adresse mail, fonction, nom d'équipe de travail ou d'entreprise qui sera automatiquement ajoutés à la fin de tous les emails que l'expéditeur enverra de sa boîte mail. Ce type de signature est très souvent utilisé pour éviter de renseigner à chaque fois ses informations à la fin d'un email. Mais il se trouve que cette signature s'ajoute à la simple signature (content formule d'au revoir suivie du nom/prénom/initiales) et rend redondante un certain nombre d'informations pour l'expéditeur. Cette superposition de signatures a tout aussi été observée sur notre corpus.

D'autres chercheurs Lampert et al. en 2009 [CITATION And09 \n \y \l 1036], proposent que les courriels peuvent être découpés en trois zones :

- ✓ La zone de locution (sender zones) contenant le texte écrit par l'expéditeur
- ✓ La zone de contenu cité (quoted conversation zones) qui contiennent à la fois le contenu retransmis d'autres conversations et celui cité du message auquel l'auteur répond.
- ✓ La zone d'encadrement (boilerplate zones) incluant le contenu réutilisé sans modifications dans plusieurs messages comme la signature ou les coordonnées de l'auteur.

De ces analyses structurelles de courriels, certaines zones telles que la zone d'encadrement (d'après Lampert et al.) et l'en-tête servent à extraire des informations pour l'analyse peu profondes et moins détaillés des conversations et permettent de reconstruire des structures de conversations. Quand il s'agit par contre d'analyser des conversations pour par exemple identifier des actes de dialogues, des évènements ou les fils de discussions, il faudrait plutôt se tourner vers les zones de locutions[2].

b) <u>Analyse de courriels</u>

Les emails ou courriels de par leur rapprochement aux anciens systèmes de lettres et des différences de temps entre des emails et leur réponse respective, les courriels sont clairement des communications asynchrones sur l'échelle de synchronicité des communications médiées par ordinateur. Avant d'être envoyés à leurs dentinaires respectifs, les courriels peuvent être relus, réédités ; ce qui enrichit ainsi leur aspect syntaxique et linguistique. De la même façon des emails qu'on peut considérer de spontanés à cause de leur temps d'envois et de réponses très courts (quelques minutes) et en général très courts (longueur de texte) sont quelque peu dépourvus de richesses grammaticales, lexicales ou syntaxiques. Ils se rapprochent ainsi des messages échangés via chat ou sur des forums et aussi la parole dont les caractéristiques ont été décrites plus haut dans la section présentant les différences entre la parole et l'écrit. Ce type d'emails pourrait être traite d'***email informel***. Mais dans un cadre professionnel, ces emails dit informels retrouvent une certaine richesse sur plusieurs aspects, ceci à cause de la responsabilité des travailleurs de connaissances. Ceux-ci à cause du cadre professionnel vont fortement éviter des abréviations, chercheront à produire des phrases bien structurés afin de faciliter et fluidifier les échanges. Les sections suivantes vont rentrer de façon plus détaillé sur les caractéristiques paralinguistiques, lexicales et syntaxiques.

---

[2] Ce paragraphe est plus au moins repris des travaux d'État de l'art : analyse des conversations écrites en ligne porteuses de demandes d'assistance en termes d'actes de dialogue de Soufian Salim **[CITATION Sou15 \l 1036]**

i. *Caractéristiques paralinguistiques*
ii. *Caractéristiques lexicales*
iii. *Caractéristiques syntaxiques*

2.2 Forums - --//--
2.3 Tchat - --//--
2.4 Tableau comparatif

	Paralinguistiques	**Lexicales**	**Syntaxiques**
Courriel ou email			
Forum			
Chats			

3 Analyse discursive de courriels – *Identification et représentation des éléments de discours et de pragmatique dans les emails*

L'analyse de discours est souvent définie comme l'analyse du langage au-delà des phrases. D'après Larousse[3], et sur un plan logique le discours est un ***ensemble d'énoncés liés entre eux par une logique spécifique et consistante, faite de règles et de lois qui n'appartiennent pas nécessairement à un langage naturel, et qui apportent des informations sur des objets matériels ou idéels.*** La linguistique moderne étudie la grammaire, les petits morceaux de langage tels que les sons (phonétique et phonologie), les parties des mots (morphologie), le sens (sémantique) et l'ordre des mots dans les phrases (syntaxe). Cependant depuis quelques années les analystes de discours étudient de plus gros morceaux de langage co-localisés et cohérents, ceci afin de saisir le contexte des discours pour mieux les analyser. Dans l'analyse de discours conversationnel, les principaux éléments à prendre en compte sont le contexte de la conversation, le tour de parole, les marqueurs de cohésion, des règles de cohérence et actes de dialogues.

Cependant pour Reboul et Moeschler [CITATION Jac98 \l 1033] qui dans leur ouvrage, font une critique de l'analyse de discours. Pour eux cette analyse de discours ne permet pas d'expliquer la compréhension des énoncées de discours, cette compréhension étant juste réduiteà des informations encodées en linguistique. Ils proposent un nouveau programme de recherche qui est la ***pragmatique de discours*** visant à appliquer au discours la prgarmatique inférentielles et cognitive. Ce paradigme de pragmatique de discours est intervenu en linguistique essentiellement pour des concepts qui décrivent la « signification ». Les premiers travaux de pragmatique ont vu le jour avec des philosophes du langage dont nous avons fait mention (Austin**, Searle, Strawson**) dans le précédent chapitre. Ainsi la théorie des actes de langages (cf. Chap2.) a joué un rôle très important à la consolidation de la pragmatique linguistique. Cette pragmatique qui au-delà des implications conversationnelles et conventionnelles, s'est vu enrichie au fil des années, avec d'autres concepts et principes tels que :

- Principe de la coopération
- Théorie de la pertinence
- Principe cognitif

[3] D'après Larousse : https://www.larousse.fr/dictionnaires/francais/discours/25859

- Concept de présuppositions
- Etc.

La pragmatique de discours a ainsi évolué dans une direction différente des approches de cohérence. Reboul et Moeschler [CITATION Jac98 \l 1033] ont montré que la notion de cohérence est plus une conséquence de l'interprétation du discours qu'une propriété définitoire du discours. Pour eux, la pragmatique de discours est basée sur les trois axiomes suivants :

- Le discours est une suite non-arbitraires des énoncées
- L'interprétation du discours est fonction de l'accès à l'intention informative globale du locuteur
- L'accès à l'intention globale est dépendant de l'accès à un ensemble d'intentions locales, basée sur l'interprétation des énoncées

• Cependant ces éléments ne suffisent pas toujours à bien interpréter ou comprendre les intentions réelles globales des interlocuteurs dans des conversations. La pragmatique a ainsi vu le jour, tirant ses origines de la philosophie du langage, de la linguistique et des sciences cognitives. Elle consiste à rendre compte des phénomènes linguistiques qui ne sont pas explicables de manière purement « interne », c'est-à-dire qui ne semblent pas pouvoir se réduire à un pur fonctionnement linguistique, mais nécessitent le recours à une analyse de *l'usage* du langage (pour faire certaines choses). À ce titre, elle prend en compte différents paramètres externes au langage, que ce soit la situation de communication, les rapports d'interlocution, le statut des locuteurs ou les intentions et croyances de ces derniers – tous éléments qu'on regroupe souvent sous l'appellation « contexte ». Il s'agit ainsi pour elle d'expliquer le langage tel qu'il se déploie en usages, ou tel qu'on le pratique en situation[4].

Ainsi donc la pragmatique vise l'étude du sens en contexte : elle cherche la signification réelle des phrase liées aux conditions situationnelles et contextuelles. Et donc pour une meilleur analyse des conversations, la pragmatique va donc prendre en compte les intentions globales et aller même jusqu'à une étude cognitives des interlocuteurs au travers de leur énoncées. Mais dans un cadre pratique d'analyse d'emails l'identification des marqueurs ou connecteurs pragmatiques va se référer à l'identification de trois types de marqueurs : les marqueurs de ***fonction illocutoire***, les marqueurs de ***fonction interactive*** et les marqueurs de ***structuration de conversation***. Les marqueurs de la fonction illocutoire sont initiatifs et réactifs, par exemple : *J'aimerais te demander quelle action as-tu entrepris ?* La partie soulignée de la phrase est le marqueur initiatif qui se rapproche beaucoup d'une demande d'informations. Les marqueurs de structuration de conversation sont très rencontrés dans la parole (qui possède beaucoup de traits communs avec les emails) et ont la particularité d'organiser le discours en unités avec la cohésion entre ceux-ci. Ces marqueurs ouvrent ou clôturent tout aussi des échanges sur des sujets précis dans un discours. L'identification de tels marqueurs avec cette particularité d'ouverture et de clôture pourront permettre dans le cadre de notre problématique de délimiter des segments d'énoncés (a.k.a segmentation de texte) portant sur des thèmes bien spécifiques. Quant 'aux marqueurs de fonction interactive, ce sont les plus typiques et le plus utilisés, ils déterminent les rapports qui existent entre différents énoncées dans un discours. Ils possèdent des caractéristiques syntaxiques et pragmatiques qui permettent de les distinguer. Notamment sur les aspects pragmatiques, on distingue trois catégories de connecteurs interactif, ceci en fonction des relations entre des énoncées de la même intervention écrite ou orale :

[4] Extrait de : **Le tournant cognitif en pragmatique. Un aller-retour transatlantique et ses impacts philosophiques**

- ***les connecteurs argumentatifs*** : car, parce que, en effet, d'ailleurs, puisque, donc, alors, par conséquent, aussi, etc.
- ***les connecteurs contre-argumentatifs*** : mais, bien, quand même, etc.
- ***les connecteurs conclusifs*** : finalement, au fond, etc

Dans le cadre de l'analyse de discours, la théorie de la structure rhétorique en anglais Rhetorical Structure Theory ou RST [CITATION Man881 \l 1033] est une théorie de l'organisation de texte et qui décrit les relations qui existent entre les parties d'un texte. Indépendamment des formes lexicales et grammaticales des textes, cette théorie rend compte de la cohérence textuelle. Elle est la base fonctionnelle de l'étude des formes spécifiques de structuration discursive, des marqueurs discursifs et autres aspects de discours. Depuis les premières publications sur cette théorie, elle a eu une attention particulière dans les programmes de recherche ayant un intérêt pour l'analyse de texte et plus particulièrement l'analyse de discours. Elle a été implémentée, comparé à d'autres approches et critiqué dans plusieurs domaines notamment celui de l'analyse de discours, de la linguistique et psycholinguistique et linguistique computationnelle.

- La constitution de corpus annotés basée sur la théorie de structure rhétorique [CITATION Car01 \l 1033] fait partie de travaux dans le cadre de l'analyse de discours. Les approches à base de règles syntactiques, sémantiques et lexicales ont permis de mettre en place de nouvelles approches pour l'analyse symbolique de discours [CITATION Pol04 \l 1033]. Ces recherches ont permis de mettre en place des algorithmes de compression de texte comme le système PALSUMM de la société Palo Alto.

-

- **3.3 Graphes et fil de discussion**

  1 Fil de discussion et analyse discursive

- **Conclusion**

Dans ce chapitre, les communications médiées par ordinateur (CMO) ont été présentées sur les plans de leurs avantages et inconvénients, de leur synchronisme ou non, aussi selon leur type de modalité. La parole et l'écrit ont tout aussi été analysés sur différents aspects linguistiques. Un peu plus loin dans les sections de ce chapitre, les caractéristiques linguistiques, syntaxiques et paralinguistiques des emails se sont vues exposées. Les emails sont en fait un type de communication asynchrone que nous allons étudiés et analysés prioritairement afin de proposer des approches pour la résolution de nos problématiques. L'analyse de discours bien que abordé dans le précédent chapitre a aussi eu un intérêt particulier dans l'une des sections de ce chapitre, notamment sur le prisme des principaux marqueurs de discours et pragmatique dans les emails. Quelques phrases ont été dédiées à la théorie de structure rhétorique qui dans la littérature est la base de plusieurs Frameworks d'analyse de discours.

Chap4. Méthodologie de constitution de corpus et exploitation

A. **Introduction**
B. **Contraintes juridiques liées à l'utilisation des données**
   1. **Demande de consentement –** *description et procédure*
   2. **Anonymisation -** *principe, techniques, avantages et inconvénients*
   3. **Pseudonymisation -** *principe, --//--*
C. **Exploitation**
   1. **Nettoyage** - Contenu superflus ou inintéressant et leur suppression
   2. **Encodage -** *Format de structuration de corpus pour traitement (TEI, Json)*
D. **Constitution de corpus synthétique**

Chap5. Représentation de données

Introduction

5.1 Représentation occurentielle – *Description, pourquoi et comment*
bag-of-words et bag-of-ngrams, fréquence tf-idf, Singular Value Decomposition, *SVD*

5.2 Représentation sémantique

5.2.1 Word embeddings – *description, technique d'implémentation, utilisation, avantages, limitations*

a) Word2Vec - CBOW, skipgram
b) gloVe -- > X
c) FastText → X

5.2.2 Word embeddings contextuel

BERT est un modèle d'apprentissage automatique largement utilisé dans le domaine du traitement automatique du langage naturel. C'est un grand mosdèle incorporant des Tranformers qui sont des modèles avancés de réseaux de neurones récurrents, avancé parcequ'ils sont capables de paralléliser des traitements et des entrainements d'inférences. Un des avatages des Transformers est qu'on peut les traiter comme des réseaux de bneurone convolutifs du fait que leur taille en entré sont fixes C'est là qu'on retrouve les avantage d'un modèle comme BERT parce qu'il permet d'effectuer d'entrainer des modele immense par des calculs parallèles. Ce qui est un grand progrès contrairement aux RNN. Par exemple avec ces derniers, le traitement d'une phrase passe par une suite de sous-traitement séquentiel de chaque mot de la phrase et pourtant avec des modèles basés sur les Tranformers, ces sous-traitements se font en parallèle.

2. Bibliographie

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Oxford University Press.

AUSTIN , J. (1975). *How to Do Things With Words.* Oxford University Press.

Austin, J. (1975). How to Do Things With Words. *Oxford University Press*,.

Bevendorff, J., Khatib, K. A., Potthast, M., & Stein, B. (July 2020). Crawling and Preprocessing Mailing Lists At Scale for Dialog Analysis. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 1551-1558). Online.

Bloomfield, L. (1933). *Language*. Chicago : University of Chicago Press.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. *InProceedings of the AAMAS 2009 Workshop"Towards a Standard Markup Language for Embodied Dialogue Acts" (EDAML 2009)*, (pp. 13-24). Budapest, Hungary.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. *In Proceedings of the AAMAS 2009 Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts" (EDAML 2009)*, (pp. 13-24). Budapest.

Bunt, H., Alexandersson, J., Carletta, J., Chae, J.-W., Fang, A., Hasida, K., . . . Traum, D. (2010). Towards an ISO standard for dialogue act annotation. *In Proceedings of the 7th International Conference on Language Resources and Systems (LREC 2010)*, (pp. 2548–2558). Malta.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., & Popescu-Belis, A. &. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA)*, 430-437.

Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., & Prévot, L. (2020). The ISO Standard for Dialogue Act Annotation, Second Edition. *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 549–558). Marseille, France: European Language Resources Association.

Cohen, W. (1992). *Learning Rules that Classify E-Mail.*

Cohen, W. W., Carvalho, V. R., & Mitchell, T. M. (july 2004). Learning to Classify Email into “Speech Acts”. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 309-316). Barcelona, Spain: Association for Computational Linguistics.

Cohen, W. W., Carvalho, V. R., & Mitchell, T. M. (July 2004). Learning to Classify Email into “Speech Acts”. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, (pp. 309-316). Barcelona, Spain.

Cohen, W., Carvalho, V., & Mitchell, T. (2004). *Learning to classify email into « Speech Acts »*.

Cook, V. (2004). *The English writing system.* London: Arnold.

Core, M. G., & Allen, J. (1997). Coding Dialogs with the DAMSL Annotation Scheme. *In Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, (pp. 28–35). Boston, MA, USA.

Crystal, D. (2001). *Language and the Internet.* Cambridge: Cambridge University Press.

Domeniconi, G., Semertzidis, K., Moro, G., Lopez, V., Kotoulas, S., & M. Daly, E. (Juillet 2017). Identifying Conversational Message Threads by Integrating Classification and Data Clustering. *Conference: International Conference on Data Management Technologies and Applications.*

Dulceanu, A. (2016). *Recovering implicit thread structure in chat conversations.*

Elsner, M., & Charniak, E. (2010). *Disentangling Chat.*

Elsner, M., & Charniak, E. (Juin 2011). Disentangling chat with local coherence models. *HLT '11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, (pp. 1179–1189).

Finke, M., Lapata, M., Lavie, A., Levin, L., Mayfield Tomokiyo, L., Polzin, T., . . . Zechner, K. (1998). *CLARITY: Inferring Discourse Structure from Speech.*

Glavas, G., & Somasundaran, S. (2020). *Two-Level Transformer and Auxiliary Coherence Modelingfor Improved Text Segmentation.*

Halliday, M. (1985). *Spoken and Written Language.* Oxford: Oxford University Press.

Hobbs, J. (1985). *On the Coherence and Structure of Discourse.* CSLI Research Report 85-37.

Jeong, M., Lin, C.-Y., & Lee, G. G. (2009). Semi-supervised Speech Act Recognition in Emails and Forums. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (p. August). Singapore.

Ji, X., & Zha, H. (2003). *Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming*.

Jyun-Yu , J., Chen, F., Chen, Y.-Y., & Wang, W. (2019). *Learning to Disentangle Interleaved Conversational Threads with a Siamese Hierarchical Network and Similarity Ranking*.

Kerr, B. (2003). *Thread arcs: An email thread visualization*.

Klimt, B., & Yang, Y. (July 30-31, 2004). Introducing the Enron corpus. *Conference: CEAS 2004 - First Conference on Email and Anti-Spam*. Mountain View, California, USA.

Kummerfeld, J. K., Gouravajhala, S. R., Peper, J., Athreya, V., Gunasekara, C., Ganhotra, J., . . . Lasecki, W. S. (2019). A Large-Scale Corpus for Conversation Disentanglement. *ACL*, 3846-3856.

Lampert, A., Dale, R., & Paris, C. (August 2009). Segmenting Email Message Text into Zones. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (pp. 919–928). Singapore.

Lewis, D. (1992). *Representation and Learning in Information Retrieval (PhD Thesis, University of Massachusetts at Amherst)*.

M., P., & R., T. D. (1997). Conversational Actions and Discourse Situations. *Computational Intelligence*, (pp. 309-347).

Mann, W., & Thompson, S. (1988). Rhetorical structure theory: toward a functional theory of text organisation. *MIT Press*.

McCarthy, M. (1991). *Discourse Analysis for Language*. Cambridge: Cambridge University Press.

Micha, E., & Eugene, C. (2011). *Disentangling Chat with Local Coherence Models*.

Oard, D., Webber, W., Kirsch, D., & Golitsynskiy, S. (2015). *Avocado Research Email Collection*. Philadelphia: Linguistic Data Consortium.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques*.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse Treebank 2.0. *In Proceedings of LREC*. In Proceedings of LREC.

Raheja, V., & Tetreault, J. (2019). Dialogue Act Classification with Context-Aware Self-Attention. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (pp. 3727-333). Minneapolis, Minnesota.

Reffay, Chanier, C., Lamy, T., M.-N., Betbeder, & (2014), M.-L. (2014). *Corpus d'apprentissage Interactions Simuligne (Simulation en ligne en apprentissage des langues)*. Récupéré sur https://repository.ortolang.fr/api/content/comere/v3.3/cmr-simuligne.html

Searle, J. (1969). Speech Acts: An Essay in the Philosophy of Language. *Cambridge University Press*.

Searle, J. (1975). *A taxonomy of illocutionary acts.*

Searle, J. (1976). A Taxonomy of Illocutionary Acts. *Linguistic Agency University of Trier*.

Searle, J. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts.* Cambridge University Press .

Shen, D., Yang, Q., Sun, J.-T., & Chen, Z. (2006). *Thread detection in dynamic text message streams.*

TRAUM, D., & HINKELMAN, E. (1992). Conversation Acts in Task-Oriented Spoken Dialogue. *Computational Intelligence*, (pp. 575-599).

Wermter, S., & Lochel, M. (1996). Learning dialog act processing. *Conference: Proceedings of the 16th conference on Computational linguistics - Volume 2.*

Whittaker, S., & Sidner, C. (1996). *E-mail overload: exploring personal information management of e-mail.*

Xiang, J., & Hongyuan, Z. (2003). *Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming*.

Yeh, J.-Y. (2006). *Email Thread Reassembly Using Similarity Matching*.