Représentation sémantique

Sommaire

Résumé

Ce chapitre présente les quelques formalismes existants pour la représentations sémantiques, développés dans le cadre de la modélisation de la compréhension.

3.1 Introduction

Les théories et formalismes sémantiques présentés dans ce chapitre s'inscrivent dans un cadre applicatif contemporain. Le sens d'un message est envisagé d'un point de vue formel et non fondamental. Il s'agit en effet de représenter les informations présentes dans ce message sous une forme cohérente et apte à renseigner un système sur les attentes de son interlocuteur.

Le cadre théorique des représentations sémantiques présentées est donc celui de la sémantique procédurale (Woods, 1981) : le sens d'un symbole est portée par une procédure abstraite liant l'expression symbolique au monde réel par l'intermédiaire d'opérations calculatoires et inférentielles réalisées par un interpréteur. Le sens d'un message est alors approché d'un point de vue procédural : le monde réel est représenté par l'état du système, le module de compréhension est l'interpréteur reliant le message reçu aux fonctionnalités du système.

Ce module doit donc s'appuyer sur une représentation structurée pour extraire le sens formel d'un message. Cette représentation est un langage possédant syntaxe et sémantique, basé sur le principe de compositionnalité présenté en I. Les connaissances sémantiques d'une application forment une *base de connaissances* qu'il est possible de représenter à l'aide d'un ensemble de formules logiques, généralement du premier ordre. Ces formules contiennent des variables potentiellement typées et instanciées par des constantes liées au domaine de l'application. Dans ce contexte, les objets sémantiques sont définis par l'instanciation de toutes les variables d'une formule ou par la composition d'objets existants. La génération de ces objets sémantiques est donc le fait d'un processus inférentiel porté par le module de compréhension.

La section 3.2 présente les notions clés de la représentation des relations sémantiques. La théorie et le formalisme des cadres sémantiques, fondés sur cette représentation, sont explicités en 3.3. Le projet FrameNet, dont les principes ont été utilisés dans ce travail, est détaillé dans la section 3.4.

3.2 Réseaux sémantiques

La représentation des relations sémantiques par des liens entre classes et objets sémantiques est discutée dans (Woods, 1975). Les formules de la base de connaissances d'une application décrivent des concepts et leurs relations qui peuvent être représentés par un *réseau sémantique*. Ce réseau est composé de nœuds matérialisant les concepts et d'arcs correspondant aux relations inter-conceptuelles (Brachman, 1979). Cette structure permet de modéliser à la fois connaissances factuelles et relationnelles, telles les relations de composition étudiées par (Jackendoff, 1990).

Le plus célèbre langage développé pour représenter des structures de type réseaux sémantiques est le langage KL-ONE (Brachman et Schmolze, 1985) dont l'élément central est le concept. Une base de connaissances KL-ONE est un réseau sémantique dans

lequel les concepts génériques sont fortement hiérarchisés. Les concepts sont les composants à partir desquels l'interprétation du message est réalisée. Ils sont définis par un ensemble d'attributs descriptifs et relationnels, les *rôles sémantiques*.

3.3 Cadres sémantiques

Les concepts et relations d'un réseau sémantique peuvent être implémentés en utilisant le formalisme des *cadres sémantiques* présenté dans (Fillmore, 1985) dans la logique des *cadres de cas* (Fillmore, 1968). Un *cadre* définit tout système relationnel de concepts au sein duquel la compréhension d'un concept fait appel à la compréhension du système complet.

Dans ce contexte, Fillmore définit les cadres sémantiques comme des structures cognitives empiriques associées au processus de compréhension (Fillmore, 1982, 1985). Les cadres sémantiques sont rassemblés dans une grammaire de cadres. Une telle grammaire génére des cadres décrivant des concepts généraux et leurs instances spécifiques. Un cadre sémantique est une structure de données représentant un concept en associant à son nom un ensemble d'éléments décrivant ses rôles situationnels (attributs) ou relationnels (rôles sémantiques).

Les mots ou groupes de mots associés aux cadres sémantiques représentent une catégorie d'expériences (situationnelles, événementielles...) liées au monde réel. Ils évoquent les cadres auxquels ils appartiennent lorsqu'ils sont présents dans un message. L'interpréteur peut alors invoquer ces cadres pour attribuer une interprétation au message (Petruck, 1996).

Un exemple d'instance de cadre sémantique est donné dans le tableau 3.1.

```
{idt0001929
instance_de        identite
  nom                Levi-Strauss
  prenom             Claude
  sexe               masculin
  date_naissance     28.11.1908
  lieu_naissance     Bruxelles
}
```

TABLE 3.1 – *Exemple d'instance du cadre sémantique* `identite`.

Le cadre sémantique présenté dans cet exemple associe au concept `identité` les éléments `nom`, `prenom`, `sexe`, `date_naissance` et `lieu_naissance`. Certains de ces éléments sont à leur tour des instances d'autres cadres sémantiques (`date` ou `lieu` dans notre exemple). La sémantique procédurale de (Woods, 1981) peut définir le mode de génération d'instances de cadres sémantiques : des procédures conditionnelles sont alors associées aux éléments des cadres. Ces procédures peuvent générer, supprimer ou modifier des cadres par inférences sur les cadres existants.

Les cadres sémantiques représentant des catégories d'expériences, leur définition - concepts et éléments associés - est fondamentalement dépendante de la tâche à traiter. Un certain nombre de projets tentent cependant de rassembler des ressources génériques. Le *répertoire de cas* de Fillmore (Fillmore, 1968), *TreeBank* (Marcus et al., 1994), *PropBank* (Kingsbury et Palmer, 2003) et surtout le projet *FrameNet* en sont quelques exemples.

Le formalisme sémantique utilisé dans ce travail étant inspiré de FrameNet, ce projet est détaillé dans la section suivante 3.4.

3.4 FrameNet

Le projet FrameNet (`http://framenet.icsi.berkeley.edu/`) de l'Université de Berkeley fournit une base de données de *frames* [1], cadres sémantiques pour la langue anglaise (Baker et al., 1998; Fillmore et al., 2003). Dans l'esprit des cadres sémantiques de Fillmore (Fillmore, 1982), les frames sont des représentations schématiques de situations. L'objectif du projet est la définition de frames alliant généricité et spécificité pour permettre leur utilisation dans des applications variées.

A chaque frame est associé de manière unique des rôles appelés *frame elements* (FE). Certains FE sont indispensable à l'instanciation de la frame, d'autres sont optionnels. La base de données construite dans le cadre du projet met en relation les frames, leurs FE et les unités lexicales (mots ou groupe de mots) qui les évoquent. Actuellement, celle-ci contient 963 frames reliées hiérarchiquement et plus de 10.000 unités lexicales (LUs). Une ressource de 135.000 propositions annotées à l'aide de ces frames et de FE est également disponible dans le cadre du projet.

L'exemple de la frame **Cogitation** est donnée ci-dessous :

```
COGITATION

Definition:
A person, the Cognizer, thinks about a Topic over a period
of time. What is thought about may be a course of action
that the person might take, or something more general.

    ex: The men were silently MULLING OVER the proposition
        of committing an assassination

FEs:
Core:

Cognizer [Cog]  With a target verb, the Cognizer is usually
```

1. Par habitude, nous préférons parler de frame sémantique. Mais les deux termes, frame et cadre, sont strictement équivalents pour nous.

```
Semantic Type   expressed as an External Argument, with the
Sentient        Topic appearing as an Object NP,  a gerundive
                verbal Complement, or a PP.
                ex: ...

Topic [Top]     With a target verb,the Topic is usually expressed
                as an Object NP, a gerundive verbal complement,
                or a PP.
                ex: ...

Non-Core:

Degree [Degr]   The FE Degree indicates the degree to which the
Semantic Type   cognizing occurs.
Degree

Depictive [Dep] Depictive phrase describing the actor of an action


Manner [Manr]   The FE Manner indicates the way in which the
Semantic Type   cognizing is being done.
Manner

Means [Mns]     An intentional action performed by the Cognizer
Semantic Type   that makes them able to cogitate.
State_of_affairs

Medium [Medium] The physical or abstract setting in which the
                Cognizer considers the Topic.


Purpose [pur]  The state-of-affairs that the Cognizer is trying to
               bring about by thinking.

Result [Result] Result of an event

Time [tim]      The time at which the Cognizer considers the Topic.


Inherits From:
Is Inherited By: Assessing, Emotion_active, Memorization
Subframe of:
Has Subframes:
Precedes:
Is Preceded by:
```

```
Uses: Mental_activity
Is Used By: Remembering_experience, Research
Perspective on:
Is perspectivized in:
Is Causative of:
See Also:

Lexical Units

brood.v, consider.v, consideration.n, contemplate.v, contemplation.n,
deliberate.v, deliberation.n, dwell.v, give, thought.v, meditate.v,
meditation.n, mull_over.v, muse.v, ponder.v, reflect.v, reflection.n,
ruminate.v, think.v, thought.n, wonder.v
```

Cette frame modélise une situation où une personne (Cognizer) pense à un sujet (Topic) pendant une période de temps. Le sujet de réflexion peut être relatif au déroulement d'une action impliquant la personne ou plus général. Deux FE principaux sont associés à la frame `COGITATION` : `Cognizer`, la personne qui réfléchit et `Topic`, le sujet de réflexion. D'autres FE secondaires lui sont également rattachés (`Depictive`...), détaillant la situation selon divers points de vue.

Un extrait du contexte relationnel de cette frame au sein de la base FrameNet est illustré par la figure 3.1.

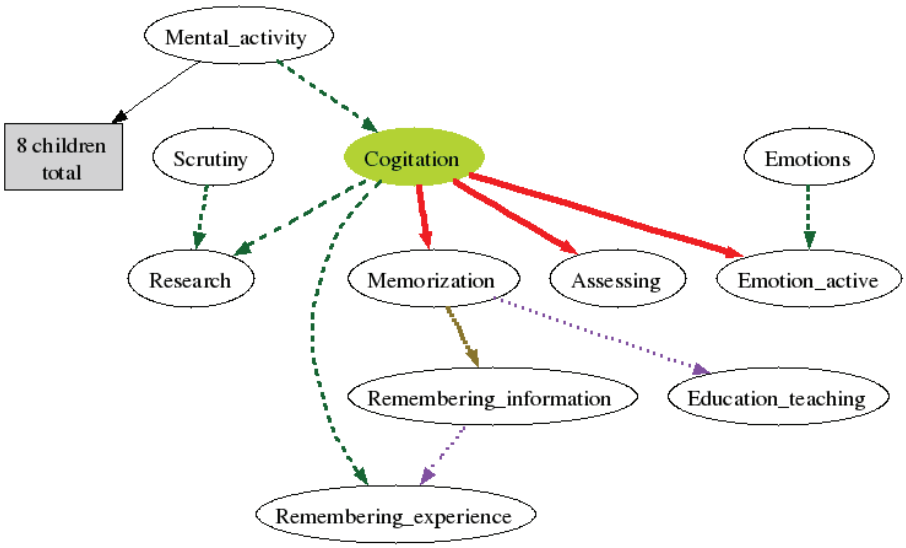

FIGURE 3.1 – *Extrait des relations liant la frame `COGITATION` à d'autres frames de FrameNet*

La liste de LUs associée à la frame `COGITATION` montre qu'une frame peut être évoquée par des LUs de différentes natures (ici, noms et verbes). Il en est de même pour les FE. Cette souplesse est un atout pour l'exploitation de ce formalisme dans les systèmes de compréhension automatique.

Par opposition aux représentations sémantiques “à plat”, la représentation en frames du sens d’un message est une représentation structurée. En effet, les représentations planes associent un concept de base à chaque segment du message mais ne composent pas les concepts ainsi obtenus. L’usage des frames définie dans FrameNet produit en revanche une représentation hiérarchiquement structurée et donc très adaptée à la tâche de composition sémantique.

3.5 Conclusion

Les capacités d’un système de compréhension sont dépendantes de la représentation sémantique choisie. Initiés par la sémantique procédurale de Woods (Woods, 1981), les réseaux sémantiques figurent parmi les premiers modèles de représentation des relations inter-conceptuelles.

En introduisant la notion et le formalisme des cadres sémantiques, Fillmore (Fillmore, 1985) propose des objets sémantiques situationnels. Les cadres sémantiques peuvent être combinés au sein de structures relationnelles pour représenter une situation réelle.

Sous-tendu par ce formalisme, le projet FrameNet s’attache à définir des objets sémantiques plus génériques pour favoriser leur emploi dans des contextes applicatifs variés. Les frames sémantiques, leurs frame-éléments et les unités lexicales qui les évoquent sont rassemblés dans une base de données qui met en évidence les relations hiérarchiques entre les objets.

La définition par unités lexicales et la structure hiérarchique des frames en font des objets sémantiques particulièrement bien adaptés à la représentation du sens d’un message dans le contexte du dialogue. En effet, ces caractéristiques les rendent aptes à évoluer sans remise en cause fondamentale et à supporter la composition sémantique avec finesse.

Chapitre 4

Matériau expérimental : le corpus MEDIA

Sommaire

Résumé

Ce chapitre propose une présentation du corpus MEDIA *qui a servi de matériau d'expérimentation et d'évaluation à ce travail. Composé de dialogues en français issus de la simulation d'un serveur téléphonique d'informations touristiques et de réservation d'hôtel,* MEDIA *a été manuellement transcrit et annoté à l'aide de structure sémantiques de type attribut-valeur. La section 4.2 détaille le mode d'obtention et les caractéristiques des dialogues composant le corpus. Les différentes transcriptions et annotations du corpus sont ensuite présentées en 4.3. Enfin, la section 4.4 rapporte les résultats des mesures de qualité effectuées sur le corpus.*

4.1 Introduction

Le corpus ayant servi de materiau d'expérimentation et d'évaluation à ce travail est un corpus de dialogues en français, produit dans le cadre du projet MEDIA (Maynard et al., 2004). L'objectif de ce projet était de tester une méthodologie d'évaluation de la compréhension hors et en contexte des systèmes de dialogue basée sur le paradigme PEACE (*Paradigme d'Evaluation Automatique de la Compréhension hors et en contexte dialogique*) (Devillers et al., 2002), fondé sur la constitution de batteries de tests reproductibles issues de dialogues réels.

La section 4.2 détaille le mode d'obtention et les caractéristiques des dialogues composant le corpus MEDIA. Les différentes transcriptions et annotations du corpus sont ensuite présentées en 4.3. Enfin, la section 4.4 rapporte les résultats des mesures de qualité effectuées sur le corpus.

4.2 Collecte du corpus

Le corpus MEDIA est dédié à l'étude des applications de demande de renseignements accédant à des bases de données. Il est composé de dialogues en français issus de la simulation d'un serveur téléphonique d'informations touristiques et de réservation d'hôtels.

Ces dialogues ont été collectés en utilisant le protocole du *Magicien d'Oz* (*Wizard of Oz*, WoZ). Lors de l'échange, les utilisateurs croient converser avec une machine alors que le dialogue est en fait pris en charge par un opérateur humain qui simule les réponses d'un serveur d'information et de réservation. L'opérateur est assisté par l'outil WoZ dans la génération des réponses à fournir à l'utilisateur. Les informations relatives à la tâche sont issues de la consultation par l'opérateur du site de réservation d'hôtels de la chaîne ACCOR [1] et du site d'informations touristiques "Tourisme en France" [2].

Le protocole de collecte des dialogues est illustré par le schéma 4.1, emprunté au "Manuel d'utilisation de l'outil WoZ. Projet MEDIA".

Après chaque phrase de l'utilisateur, l'opérateur consulte l'outil WoZ qui lui propose la réponse à fournir en fonction du nouvel état du dialogue. Pour diversifier les réponses de l'opérateur, l'outil WoZ est paramétré au niveau des messages, des consignes et des scenarii. Un ensemble de messages est associé à l'application pour varier les formulations des réponses. A chaque appel, l'opérateur doit respecter une série de consignes (par exemple, faire semblant de ne pas avoir compris l'utilisateur pour simuler les erreurs que ferait un système réel). Ces consignes doivent être fournies à l'outil WoZ et dépendent du scenario choisi pour le dialogue à enregistrer.

1. http ://www.accorhotels.com
2. http ://www.tourisme.fr

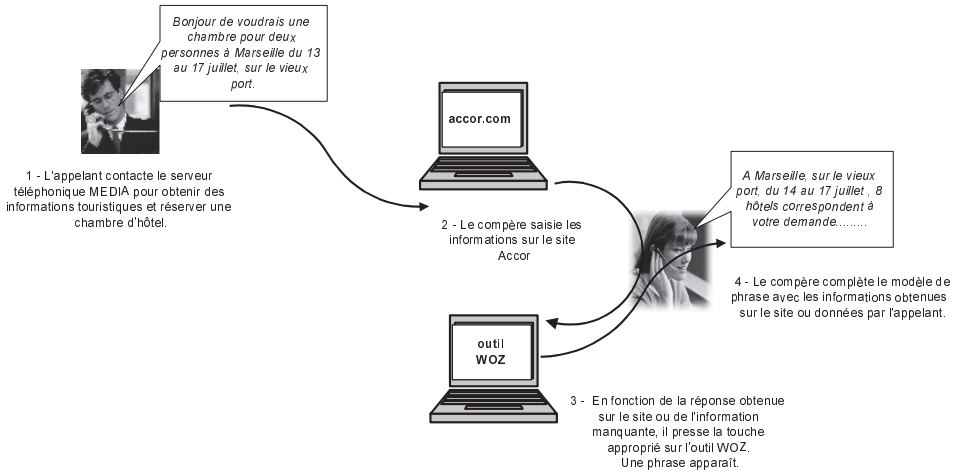

FIGURE *4.1* – *Protocole du Magicien d'Oz*

L'outil WoZ ne permet pas la gestion de l'ensemble du dialogue. L'opérateur doit gérer l'intégration des informations pratiques aux schémas de réponses proposés par le WoZ. Il doit également prendre en charge toutes les opérations de saisie, de correction, de négation d'items ainsi que l'énonciation orale de la réponse à l'utilisateur.

Les dialogues ont pour but la réservation d'une ou plusieurs chambres dans un ou plusieurs hôtels. Les réservations s'effectuent dans le cadre de l'organisation d'un week-end, de vacances ou d'un séjour professionnel. Ainsi, les dialogues peuvent débuter à partir de thèmes variés : choix d'une ville de séjour, recherche d'un itinéraire ou d'un événement touristique, satifaction d'une contrainte de prix ou de date (Bonneau-Maynard et al., 2005). Huit catégories de scenarii ont été définies avec différents niveaux de complexité, chaque locuteur ayant enregistré cinq scenarii différents. Un extrait de dialogue est présenté dans le tableau 4.1.

Le corpus MEDIA est composé de 1257 dialogues produits par 250 locuteurs, pour une durée totale d'environ 70 heures d'enregistrement audio. L'ensemble de ces dialogues contient un total de 15.000 requêtes réparties sur les portions d'entraînement, de développement et de test du corpus comme indiqué dans le tableau 4.2.

Une description détaillée des caractéristiques techniques du corpus est donnée dans le tableau 4.3.

La collecte du corpus et la gestion technique du projet MEDIA ont été prises en

WoZ	Bienvenue sur le serveur MEDIA système d'informations touristiques et de réservations d'hôtel. Quelle information désirez-vous ?
Locuteur	Je souhaite réserver un hôtel à Nancy
WoZ	A Nancy, quelles sont vos dates de séjour ?
Locuteur	Une nuit, le quinze juillet deux mille quatre
WoZ	Dans quelle partie de la ville souhaitez vous que votre hôtel se situe ?
Locuteur	Je n'ai pas d'exigence particulière.
WoZ	Deux hôtels correspondent à votre demande...

TABLE *4.1* – *Extrait de dialogue du corpus* MEDIA.

portion du corpus	nombre de requêtes
entraînement	11.010
développement	1.009
test	3.005

TABLE *4.2* – *Nombre de requêtes sur les différentes portions du corpus* MEDIA.

nombre de mots moyen par phrase utilisateur	8,3
nombre de mots moyen par phrase système	14,4
taille du vocabulaire utilisateur	2.715 mots
taille du vocabulaire système	1.932 mots
durée moyenne d'un dialogue	3 min et 30 s

TABLE *4.3* – *Caractéristiques du corpus* MEDIA.

charge par ELDA/ELRA [3]. La société VECSYS [4] a mis en place la plate-forme d'enregistrement du corpus (matériel et outil WoZ).

4.3 Transcription et annotation du corpus

L'annotation sémantique du corpus a nécessité la définition d'une représentation sémantique adaptée au domaine de la tâche MEDIA. Cette représentation est générique, assure une bonne couverture du domaine et est cependant suffisament simple pour permettre l'annotation d'un corpus de la taille de MEDIA. Elle est basée sur une structure de type attribut-valeur dans laquelle les relations conceptuelles sont représentées implicitement par le nom des attributs. Les attributs sont donc les concepts liés au domaine.

Le dictionnaire sémantique utilisé associe à un mot (ou un groupe de mots) une paire *concept-valeur* puis un spécifieur définissant des relations entre concepts et enfin un *mode* (affirmatif, négatif, interrogatif ou optionnel) attaché au concept. Avec 19 spé-

3. http ://www.elda.org
4. http ://www.vecsys.fr

cifieurs pouvant être associés aux 83 concepts de base, le schéma d'annotation MEDIA offre un mécanisme simple permettant de préserver certaines relations élémentaires entre les concepts au sein de la phrase.

Chaque tour de parole du locuteur est scindé en segments sémantiques correspondant à une unique paire.

Un exemple de message annoté du corpus MEDIA est donné dans le tableau (4.4). La première colonne contient les séquences de mots W^c supports de chaque concept, présenté dans la seconde colonne. La troisième colonne indique le mode et la quatrième colonne fournit les spécifieurs associés aux concepts. La dernière colonne présente les valeurs normalisées des concepts c associés aux séquences W^c.

W^c	*concept c*	*mode*	*spécifieur*	*valeur*
Je voudrais réserver	`commande`	+		réservation
une chambre	`chambre-quantité`	+	réservation	1
pour deux nuits	`séjour-nbNuit`	+	réservation	2
à Marseille	`localisation-ville`	+	hôtel	Marseille

TABLE 4.4 – *Exemple d'annotation sémantique du corpus* MEDIA.

Dans cet exemple, le spécifieur *réservation* est lié aux concepts `chambre-quantité` et `séjour-nbNuit`. On obtient ainsi une structure hiérarchique représentant une réservation associée au concept `commande` et développée grâce aux valeurs des concepts `chambre-quantité` et `séjour-nbNuit`. Le spécifieur *hôtel* adjoint au concept `localisation-ville` permet de relier le lieu évoqué dans le segment *"à Marseille"* à la partie précédente de l'énoncé.

La combinaison des spécifieurs et des concepts permet de recomposer un premier niveau de représentation hiérarchique de la requête du locuteur à partir de l'annotation *à plat*.

4.4 Qualité du corpus : l'accord inter-annotateur

Le corpus MEDIA est transcrit manuellement et enrichi par une annotation conceptuelle également manuelle réalisée par deux annotateurs de la société ELDA.

Pour déterminer la qualité des annotations, l'accord entre les différents annotateurs a été évalué. Cet accord inter annotateur (IAg pour Inter-annotator Agreement) est mesuré en utilisant la mesure de Kappa k telle que :

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

où $P(A)$ est le rapport du nombre de fois où les annotateurs sont d'accord sur le nombre total d'annotation et $P(E)$ la probabilité que l'annotation correcte ait été posée par hasard.

Le coefficient de Kappa et ses différents modes de calcul sont présentés dans (Siegel et N.J. Castellan, 1988). La mesure de Kappa est détaillée et discutée dans (Carletta, 1996). Il est admis dans la littérature que la fiabilité des annotations est bonne dès lors que l'IAg mesuré par le coefficient de Kappa est supérieur à 80%. Les IAg mesurés dans le cadre du projet atteignent presque 90% dans la phase finale du projet (Bonneau-Maynard et al., 2005), ce qui permet de valider la qualité des annotations dans le corpus. Le tableau 4.5 présente les résultats des IAg mesurés au cours de la phase finale d'annotation du corpus.

Evaluation	1	2	3	4
Nb de dialogues	10	10	10	10
Nb de tours de parole locuteur	165	137	106	163
Nb de segments sémantiques	372	455	342	459
IAg (%)	89,5	83,1	83,9	87,8

TABLE 4.5 – *IAg finales obtenues sur l'annotation du corpus* MEDIA.

4.5 Conclusion

Le corpus de dialogues MEDIA est le matériau d'expérimentation de ce travail. Les dialogues MEDIA ont pour but la réservation d'hôtels et l'obtention d'informations touristiques. Ils sont collectés en utilisant le protocole du Magicien d'Oz dans lequel un opérateur humain assisté d'un outil d'aide à la décision simule les réponses d'un serveur téléphonique.

Le corpus contient 1257 dialogues pour environ 15.000 requêtes utilisateur. Transcrit manuellement, chaque dialogue est également annoté sémantiquement à l'aide de concepts de base. Les requêtes utilisateur sont scindées en segments sémantiques (d'un ou plusieurs mots) auxquels est attribuée une paire concept-valeur.

Le dictionnaire de concepts MEDIA comporte 83 entités. L'annotation est enrichie par l'indication de relations entre les concepts de base du message et le mode de la proposition. Les annotations sémantiques associées aux requêtes ont été conçues pour pouvoir être utilisées par le module de compréhension d'un système de dialogue.

La qualité d'annotation du corpus MEDIA, évaluée tout au long de la phase de travail manuel des experts, est attestée par un IAg final supérieur à 85%. Les données MEDIA représentent donc un support d'expérimentation conséquent et fiable.