Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1159–1166
Marseille, 11–16 May 2020

What Speakers really Mean when they Ask Questions : Classification of Intentions with a Supervised Approach

Angèle Barbedette, Iris Eshkol-Taravella
Université Paris Nanterre, MoDyCo UMR 7114
200 avenue de la République, 92001, Nanterre, France
angele.barbedette@gmail.com, ieshkolt@parisnanterre.fr

Abstract

This paper focuses on the automatic detection of hidden intentions of speakers in questions asked during meals. Our corpus is composed of a set of transcripts of spontaneous oral conversations from ESLO's corpora. We suggest a typology of these intentions based on our research work and the exploration and annotation of the corpus, in which we define two "explicit" categories (REQUEST FOR AGREEMENT and REQUEST FOR INFORMATION) and three "implicit" categories (OPINION, WILL and DOUBT). We implement a supervised automatic classification model based on annotated data and selected linguistic features and we evaluate its results and performances. We finally try to interpret these results by looking more deeply and specifically into the predictions of the algorithm and the features it used. There are many motivations for this work which are part of ongoing challenges such as opinion analysis, irony detection or the development of conversational agents.

Keywords: intention analysis, speech act theory, dialog acts, sentiment analysis, opinion mining

1. Introduction

1.1. Motivations

Many of today's NLP challenges focus on the use of figures of speech to express an idea, an opinion, a will, etc., such as irony (a mocking way of saying something when you think the opposite) or litany (a way of reducing what you say : you say less than what you really think). Most of the existing work on this problem focuses on written or oral non-spontaneous data, which include more easily identifiable clues to identify and understand these cases, particularly through prosody. For this work, we decided to focus only on oral transcriptions, and more specifically on transcriptions of recordings of natural and spontaneous conversations during meals. In our opinion, these are more challenging because it is less easy to interpret the intentions of the speakers' utterances in them. The main problem is to be able to determine, without relying on oral and prosodic indicators, what the speaker means in a non-literal way, that is, his intentions, when he asks a question. The aim is to try to meet a NLP need to define and identify the implicit content, which does not only require to find textual clues. This work also provides us with a corpus of annotated natural and spontaneous oral data, a type of data that is difficult to obtain and yet very needed in NLP. There are many possible outcomes to this work like the improvement of human-machine dialogue and chatbots, a deeper opinion analysis and irony detection.

1.2. Organization of the paper

We will start by contextualizing this work and defining what is an intention. After presenting our corpus, we will suggest a typology, show how we pre-processed the data and how we chose the features to be added. Finally, we will present our results and try to interpret and discuss them.

2. Related work and contributions

Austin (1962) associates actions with language : when a speaker says something, he does something. He introduces three categories of speech acts which are the locutionary act (act of saying something), the illocutionary act (act performed in saying something) and the perlocutionary act (act performed by saying something). Searle (1975) distinguishes cases where the speaker produces a statement with exactly what he or she is saying from other cases involving irony or metaphors. The first type of statement is intended to convey a particular illocutionary force. In the second type, several illocutionary forces may be involved. Searle refers to *indirect speech acts* as cases where an illocutionary act is performed indirectly by another act. These indirect speech acts are composed of a primary, i.e. non-literal, illocutionary act and a secondary, i.e. literal, illocutionary act. In the statement "Can you pass the salt to me ?", the secondary language act (the literal meaning) would be "Do you have the ability to pass the salt to me ?" and the primary language act (the non-literal meaning) would be "Give me the salt". As Kerbrat-Orecchioni (1986) pointed out, the propositional content of the statement carries an illocutionary force that is similar to the intention expressed by the speaker. Allen and Perrault (1980) use the term *intentional action* to define speech acts : "A speech act is an intentional action that has as parameters a speaker [...], a hearer, and a propositional content, and whose execution leads to the production of an utterance. Their preconditions and effects are defined in terms of the beliefs and wants of the speaker and hearer". The notion of *intention* is part of both the fields dealing with dialog (or speech acts) and opinion analysis. Opinion analysis is a very active field in NLP since it is becoming increasingly easy to collect opinions on the web, particularly through social networks. According to Karoui et al. (2019), an opinion can either be explicit, i.e. it can be identified using textual clues (words, symbols or subjective expressions of language), or implicit, i.e. based on cultural or pragmatic knowledge shared by the sender of the message and its receiver. The purpose of dialog acts is to contribute to the fine analysis of the conversation and all the types of statements that compose it. Several works on

the task of automatic classification into dialog acts exist : it consists in classifying statements in a category of dialog act chosen from a set of predefined categories that fulfil specific functions of social discourse (Moldovan et al., 2011). To complete this task, several taxonomies of dialog acts have been suggested. We can refer in particular to the DIT ++ annotation scheme (Bunt, 2005) or other taxonomies such as DAMSL (Allen and Core, 1997), SWBD-DAMSL (Jurafsky and Shriberg, 1997), HCRC Map Task (Anderson et al., 1991) or VERBMOBIL (Alexanderssony Bianka et al., 1997). Since our subject matter is questions, our work will focus on statements that Austin describes as performative, that is, statements that have both an illocutionary aspect and a perlocutionary effect on the communication situation, since we consider that when a speaker asks a question, he always means something more than what the locutionary act of the question actually conveys. The questions therefore also fall within the scope of Searle's work and its definition of indirect speech acts since, in our view and as defined by Searle, they always perform a primary and a secondary illocutionary act that the hearer must be able to interpret. The combination of the characteristics of opinions and dialog acts leads us to define what is an intention in this study. It is the illocutionary activity expressed by a statement that makes it possible to characterize it according to its purpose, whether it is explicit, i.e., directly identifiable in the statement, or implicit, i.e., based on the common knowledge of the conversation's participants. Illocutionary activity is not just limited to the expression of opinions : it also applies to all types of purposes involved in the production of a statement. Chen et al. (2013) distinguish explicit intentions, that is, intentions that are clearly and explicitly stated and for which it is not necessary to infer anything, from implicit intentions. Examples are given to illustrate this difference : "I am looking for a brand new car to replace my old Ford Focus" in which the speaker explicitly states that he wants to buy a new car, and "Anyone knows the battery life of iPhone ?" in which the speaker may have the idea of buying a new phone. This work will focus on the implicit intentions expressed by speakers when they ask a question and will attempt to automatically classify them.

3. Corpus

3.1. General presentation

The data used for this work comes from the ESLO1 and ESLO2 oral corpora, created as part of the ESLO scientific project, Enquêtes SocioLinguistiques à Orléans, of the Laboratoire Ligérien de Linguistique of the University of Orléans (Baude and Dugua, 2011; Eshkol-Taravella et al., 2011).

3.2. Corpus architecture and transcription format

The ESLO1 corpus is composed of about 300 hours of recording and the ESLO2 of 400 hours of recording. For ESLO1, we can count about the same number between interviews and more diverse recordings such as phone calls or conversations during meals. ESLO2 includes more various recordings of conversations in public places (e.g. bakeries, markets, shops, counters and cinemas) or private places than interviews. Transcriptions of the ELSO1 and ESLO2 recordings are available on the ESLO website and open to the public. They are in *.xml* format and specify the identity of the speakers and transcribers, the start and end times of the recording and the details of the speaking turns during the recording.

3.3. Use of corpus data

As part of this work, we will use all transcriptions of recordings made during meals and available on the ESLO project website, both in ESLO1 and ESLO2, which corresponds to a total of 28 files in *.xml* format. Among these 28 files, we can count seven of them as part of the ESLO1 corpus and 21 as part of the ESLO2 corpus. These files form a whole of about 19 hours of recording. The choice of this specific category is linked to the desire to use the most spontaneous data. Indeed, since one of the aims of this work is to predict the speaker's intention through questions, it seemed more logical not to use transcriptions of recordings made during interviews or other formal situations, but rather to use data collected in natural conversations. Meals are also a category in which we can expect to have more questions since we can assume that part of the conversations will focus on requests and questions about the content of the meals served or the actions involved in the preparation of a meal.

4. Intention modeling in questions

4.1. Reference corpus

The data from the original transcripts were formatted using a script that included :

- the cleaning of the corpus ;
- the extraction of the speaking turns, represented by *Turn* tags, and information from the speaker(s) in each of them, represented by a *speaker* attribute ;
- the extraction of the questions, using the transcribed question marks ;
- the extraction of the left and right contexts of each question, assuming that a context is a maximum of ten speaking turns (this means that the goal is to pick up ten turns if possible and to pick up as many as possible if there are less than ten turns, for example in the case of a question asked at the beginning of the recording which then appears in the first few turns of the file) and that a turn is a list of sequences transcribed and found in the original data, more precisely within the same *Turn* tag ;
- the writing to an output file (figure 1), with empty attributes that will allow us to manually annotate the data.

The annotation started even before we had a definitive version of our typology of intentions and was divided into several steps :

- the quick look through the annotation file to try to find the common points between the questions and thus be able to have a first idea of the categories to be defined ;

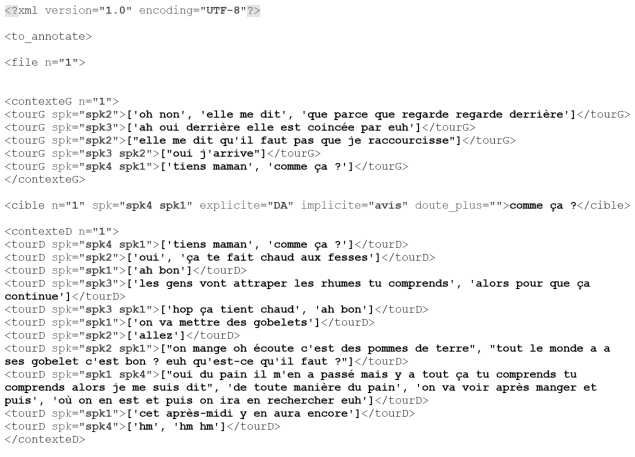

Figure 1: Extract of the annotation file

- the back and forth between the annotation task and the typology ;
  - several tests of annotation of some questions based on this first idea of the identified groups,
  - discussions between the different annotation steps to find and refine the chosen labels and associated definitions, and thus finalize the typology ;
- the annotation of the 3647 questions of the corpus by regularly referring to the established typology.

4.2. Typology elaboration

The different steps of this methodology have allowed us to build a typology of intentions in questions divided into two parts : the first, focusing on the type of answer expected for each of the questions, and the second, focusing on the intention expressed by the speaker through the question. For the first part, we can define two categories related to what is literally said or explicit that help classify questions according to the type of answer expected from the receiver of the message. A question could therefore be :

- a *request for agreement*, that is, a question whose answer may be "yes" or "no" ;
- a *request for information*, that is, a question whose answer is something other than "yes" or "no".

Some examples of these two categories can be found in table 1.

The second part implies more complex categories to be determined since they require an interpretation. They are what the sender suggests to the receiver of the message and that must be decoded and understood : it is a non-literal or implicit message. As a result of the several steps that led to our reference corpus, we were able to identify three categories reflecting the intention expressed by the speaker by producing a statement and more precisely a question : *opinion*, *will* and *doubt*. These three types of intentions can be detected and distinguished through criteria that they may or may not match. We can find some examples in table 2.

REQUEST FOR AGREEMENT	REQUEST FOR INFORMATION
alors Joy tu en veux ? (*so Joy, do you want some ?*)	regarde là-dedans c'est quoi ? (*look in there, what's that ?*)
je peux mettre ça là ? (*can I put this here ?*)	c'est où Saint-Raphaël ? (*where is Saint-Raphaël ?*)

Table 1: Examples of questions for categories related to what is explicit

When the speaker's intentions are to express an *opinion*, the questions meet the following criteria :

- they express positive or negative judgments ;
- they do not necessarily imply any action by any of the speakers.

When his intentions are to express a *will*, the questions have the following characteristics :

- they correspond to the will of the speaker or his interlocutor(s) to do something or to behave in a certain way (we don't necessarily know what the speaker wants but we know that he wants something) ;
- they often assume the use of a verb expressing an action ;
- they imply an answer corresponding to an action in the present, in the near future or at a given time.

Finally, when the speaker's intentions correspond to the expression of a *doubt*, the questions meet other criteria :

- they are a questioning of what is being said, of the truth or falsity of a thing or an external event ;
- they may be similar to a repetition, a request for confirmation, a request for clarification but also to surprise or astonishment ;
- they do not necessarily imply any action by one of the speakers.

4.3. Reference corpus and typology evaluation

To ensure the reliability of our categories, their definitions and therefore our annotations, we have decided to set up a collaborative method in order to evaluate the reference corpus. Using an online form created from Google Forms and associated with instructions, examples and counter-examples, we were able to solicit contributions to complete an annotation task of fifteen questions from our corpus. Figure 2 shows an example of what we could find in the form. It is a question ("tu me donnes un petit peu maman ?", *Can you give me a little Mum ?*), surrounded by its left and right contexts (ten speaking turns before and ten speaking turns after, if possible), with the speaker's identifiers added in order to get a better understanding of the conversation. Each example is followed by two questions :

OPINION	WILL	DOUBT
ils veulent passer pour des boulets ou euh ? (*they want to look like assholes or, uh ?*)	il a fini sa côte d'ailleurs ? (*he finished his rib by the way ?*)	quelle casserole ? (*which pan ?*)
tu as vu comment elle prend soin de moi ? (*did you see how she takes care of me ?*)	tu en veux toi ? (*do you want some ?*)	et sinon le et et la Caf c'est réglé ? (*and by the way, the and and the problem with the Caf is fixed ?*)
pourquoi tu es malpolie ? (*why are you being rude ?*)	bon ça y est tu as bien saccagé le le journal ? (*well, have you finished trashing the newspaper ?*)	alors il est plus avec sa nana lui ? (*so he's not with his girlfriend anymore ?*)

Table 2: Examples of questions for categories related to what is implicit

P3	'traînent', 'la mode traîne', 'parce que la'
P3/P1	'oh là là', "tu as vu l'orage ?"
P1	'horrible hein'
P3	"tu crois pas qu'il faut quelque chose de marrant ?"
P1/P3	'il y a', 'oh non non'
P3/P1	-
P2/P1	"ce qui est le plus joli c'est carrément le le le le à mi-mollet", 'tiens tu en veux euh', '?'
P2	'vraiment moi je'
P1	'Suzanne tu en veux ?'
P3	'non merci'
P4	'tu me donnes un petit peu maman ?'
P3	"moi je trouve qu'on on", 'pareil à', 'à mi-mollet là oh moi tu sais je crois que ça reviendra', 'plus ou moins ça reviendra', 'aussi bien', "
P1/P3	'moi je crois que y aura'
P3	"y aura la mode pour l'hiver et puis la mode pour l'été", "le court pour l'été le long pour l'hiver"
P3/P2	'tu te rends compte ?', 'ah oui'
P5	"mais ce qui est le plus marrant c'est qu'en été y a"
P1	'ah oui'
P5	'encore une autre'
P3/P1	'eh bah dis donc', 'dis donc elle va'
P1	'hein'
P5	'encore'

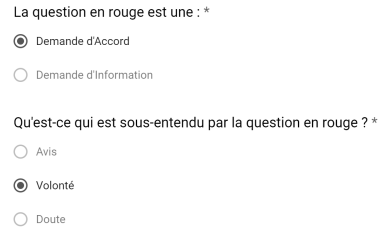

Figure 2: Extract of the evaluation form

the first one is about its literal aspect and the second one is about the intention which is expressed. This form collected a total of twenty-six contributions, i.e. participants who annotated the fifteen questions, and we were able to compare the answers with the annotations from the reference corpus. We calculated an inter-annotator agreement with a Cohen's Kappa coefficient (generally used to measure the agreement between two qualitative judgments) between the reference and the answers of each of the participants. The results, which we can see in table 3, show a certain consistency since, as we can see from the median measurement, we obtained for half of the annotations an inter-annotator agreement higher than 0.73 for the explicit categories and higher than 0.6 for the implicit categories. This means, according to the Cohen's Kappa interpretation table of Landis and Koch (1977), that we have a strong agreement (between 0.61 and 0.8) or an almost perfect agreement (between 0.81 and 1) for half of the participations for our two types of categories. We can also see, thanks to the calculation of the 3rd quartile, that about a quarter of the participants obtained an almost perfect agreement, since it is higher than 0.86 for the explicit and higher than 0.8 for the implicit.

Participant	Kappa for explicit	Kappa for implicit
P1	0.86	0.7
P2	0.59	0.6
P3	0.29	0.8
P4	1	0.5
P5	0.59	0.4
P6	0.86	0.6
P7	0.86	0.8
P8	0.36	0.4
P9	0.74	0.8
P10	1	0.9
P11	0.39	0.6
P12	1	0.9
P13	0.47	0.5
P14	0.86	0.7
P15	0.86	0.5
P16	1	0.9
P17	0.47	0.7
P18	0.47	0.6
P19	1	0.6
P20	0.05	0.4
P21	0.72	0.8
P22	0.74	1
P23	0.62	0.4
P24	1	0.8
P25	0.62	0.6
P26	0.19	0.6
Median	0.73	0.6
3rd quartile	0.86	0.8

Table 3: Inter-annotator agreements

5. Pre-processing and linguistic features

5.1. Morpho-syntactic labelling and lemmatization

Before starting the automatic classification task, we had to do some pre-processing on our data :

- morpho-syntactic labelling ;
- lemmatization.

For lemmatization, we chose to use *TreeTagger*, an annotation tool that allows us to obtain both lemma and POS tagging information for each word (Schmid, 1994), and more specifically we used the parameter files of the *PERCEO* project, (Projet d'Étiqueteur Robuste pour l'Écrit et pour l'Oral), a robust labeller project for written and oral data made of resources whose goal is the automatic annotation

of written and oral data in lemma and parts of speech (Benzitoun et al., 2012). The reason for choosing these different resources for our work is simply the desire to use tools adapted to our data, which are transcriptions of recordings that took place during informal meals, in which we will certainly find dissimilarities with traditional written data.

5.2. Vectorization

The vectorization of questions and their contexts is necessary to get a vector representation of the text that can be used as input to our classification algorithm, in addition to other linguistic features. In contrast to other types of vectorizations that are supposed to capture the meaning of words (such as *word2vec* with the models *CBOW* and *Skip-Gram* or the pre-trained vectors of *Flair*), the one obtained with a *TF-IDF* (*term frequency-inverse document frequency*) allowed us to obtain slightly better classification results. In addition, the TF-IDF measurement allows us to assign a weight to the words that compose the questions and their contexts according to their frequency in a given document and in all the documents of the corpus. In other words, it is a measure that takes into account their importance, the rarity and therefore the discriminative function of words within the corpus.

5.3. Lexical features

5.3.1. Lexicons

The elaboration of our typology of intentions in questions and the annotation phases of our corpus have pointed out lexical aspects that allow us to discriminate between our different categories. These elements are indicators that we have decided to group into six lexicons detailed below :

- *speech verbs* (34 occurrences) : they are generally used to report, introduce discourse and words (*say, ask, propose, suggest, explain,* etc.) ;
- *movement verbs* (1003 occurrences from the lexical resource DinaVmouv, (Stosic and Aurnague, 2017)) : they express an idea of movement, displacement and therefore action (*hook, follow, sit, fill, go,* etc.).) ;
- *interrogative words* (24 occurrences) : these are adverbs and interrogative pronouns (*who, how many, which, why, when,* etc.) ;
- *interjections* (73 occurrences) : they correspond to the list of interjections presented in the ESLO corpus transcription guide ;
- *feelings* (190 occurrences) : these can be nouns, verbs, adjectives or adverbs that allow emotions or opinions to be expressed (*appreciate, delighted, hatred, null, disturb,* etc.) ;
- *modal adverbs and adjectives* (24 occurrences) : they allow us to affirm or question something (*really, impossible, certainly, perhaps, true,* etc.).

5.3.2. Other features

In addition to vectorizations and information related to lexicons, i.e. the frequency of appearance of words in each question and left or right context, we have integrated into our lexical features the number of words and the number of characters in each question and in each context.

6. Automatic classification of intentions in questions

6.1. Experiences and results

6.1.1. Weka

To start our experiments, we first used *Weka*, which allows the use of many algorithms and machine learning tools, especially for automatic classification and data mining. We were therefore able to test the *Random Forest* algorithm for the classification of vectorized questions, on the one hand in explicit categories (*request for agreement*, or *RA*, and *request for information*, or *RI*) with 2538 questions to classify and on the other hand in implicit categories (*opinion*, *will* and *doubt*) with 858 questions to classify. We obtained 82.782% of well classified items and 17.218% of misclassified items for the first case and 60.14% of well classified items and 39.86% of misclassified items for the second case. Table 4 shows the results obtained with precision, recall and f-score measures for each category, as well as an overall average of these measurements by typology (explicit or implicit). These numbers give an idea of the results to be obtained and improved by implementing the Random Forest algorithm.

	RA	RI	AVG.	OPINION	WILL	DOUBT	AVG.
Precision	0.894	0.781	0.837	0.737	0.636	0.516	0.630
Recall	0.744	0.912	0.828	0.462	0.622	0.720	0.601
F-score	0.812	0.841	0.827	0.568	0.629	0.601	0.599

Table 4: Classification results with Random Forest in Weka

6.1.2. Random Forest implementation

The corpus was first cut into two representative samples of the data :

- a training sample, which is the learning set (in our case it is 75% of the data set) by which the model or algorithm fits the data and learns ;
- a sample of test (in our case it is 25% of the data set) to provide a final evaluation of the model.

We used the cross-validation method *k-fold cross-validation* which consists in separating the corpus into *k* samples (parameter for which we chose the value of 8) and using each of them one after the other as a test set and the other ones as a training set. In order to get the best results from our model, we tested several hyperparameters for our algorithm and selected the most optimal values. The hyperparameters we have decided to test are as follows :

- *n_estimators* which corresponds to the number of trees in the forest of decision trees ;
- *criterion* which corresponds to the type of measurement chosen to evaluate the quality of each separation point of a tree, i.e. each node ;
- *bootstrap* which corresponds to the choice of using or not using new data samples selected in the initial sample.

All our experiments were tested by balancing the categories to avoid a bias in the results. The performance of the model was evaluated by averaging the performance of each category using precision, recall and f-score measures and by visualizing the predictions of the algorithm using a confusing matrix.

6.1.3. Results

The goal is to have an idea of the features that can help us with the task of automatically classifying our questions, so we decided to test our algorithm with several combinations of different features that are presented in table 5. This table is an overview of all the precision, recall and f-score measures obtained for each experiment. Each of them represents a selected set of features, which are ticked in the table. We can see for example that all experiments include the vector of the question but only experiments 8 and 9 include the number of characters for the left and right contexts, and that experiment 3 only takes into account two features corresponding to the vector and the POS tagging of the question. This table shows very similar results overall as we see

		EXPERIMENTS								
		1	2	3	4	5	6	7	8	9
FEATURES	vector q.	×	×	×	×	×	×	×	×	×
	vector l.		×							
	vector r.		×							
	POS tagging q.			×	×	×	×	×		
	POS tagging l.									
	POS tagging r.									
	feelings q.				×	×	×	×	×	×
	feelings l.					×				
	feelings r.					×				
	interjections q.				×	×	×	×	×	×
	interjections l.					×				
	interjections r.					×				
	interrogative q.				×	×	×	×	×	×
	interrogative l.					×				
	interrogative r.					×				
	movement q.				×	×	×	×	×	×
	movement l.					×				
	movement r.					×				
	speech q.				×	×	×	×	×	×
	speech l.					×				
	speech r.					×				
	modal q.				×	×	×	×	×	×
	modal l.					×				
	modal r.					×				
	explicit						×	×	×	×
	nb. words q.							×	×	×
	nb. words l.								×	×
	nb. words r.								×	×
	nb. char. q.							×	×	×
	nb. char. l.								×	×
	nb. char. r.								×	×
RESULTS	Precision	0.622	0.492	0.618	0.632	0.6	0.63	0.621	0.631	0.613
	Recall	0.612	0.493	0.612	0.623	0.592	0.622	0.617	0.624	0.606
	F-score	0.611	0.489	0.61	0.622	0.59	0.62	0.616	0.622	0.603

Table 5: Overview of experiments and results

for experiments 4, 6 and 8 for which we get a f-score near 0.62. In particular, we can see the exact distribution of the predictions of our algorithm for experiment 8 with a confusion matrix (figure 3). However, some measures seem to differ, such as those in experiment 2, which are lower than 0.5, or those in experiment 5, which are lower than 0.6, with both experiments taking into account features related to the context of the question (vectorization in the first case and the presence of words belonging to lexicons in the second one).

6.2. Interpretation and discussion

The first observation we make is about the decrease in performance when context-related features are added, as in experiments 2 and 5. To verify the importance of the context of the questions, we have reproduced experiment 2, which takes into account only the vectorizations of the questions

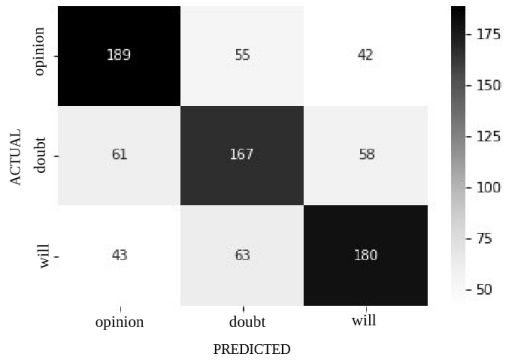

Figure 3: Predictions for experiment 8

and its contexts, and experiment 8, which considers the number of words and the number of characters in the contexts, with two and five speaking turns before and after the question in order to compare the results obtained with ten speaking turns. The scores for experiment 2 (table 6) show an improvement in performance when there are fewer speaking turns in each context, which can be explained by the excessive amount of irrelevant information reported by the vectorization of contexts when they are wider. In contrast, we observe for experiment 8 that the scores are better when the context window is larger : when the algorithm uses lexical information and in particular the number of words and characters, its performances are higher when more contexts are used. To try to interpret the results we

	Experiment 2			Experiment 8		
Speaking turns for each context	2	5	10	2	5	10
Precision	0.575	0.54	0.492	0.588	0.591	0.631
Recall	0.572	0.536	0.493	0.584	0.587	0.624
F-score	0.571	0.535	0.489	0.582	0.585	0.622

Table 6: Comparison of experiments 2 and 8 with two, five and ten speaking turns in the left and right contexts

focused only on experiment 8, one of the experiments with the highest scores and the most features used : it includes all the features that provide information about questions, as well as the number of words and characters for the left and right contexts, which are the only ones about the contexts of the questions to be classified. To get a better view of the results and understand how the algorithm worked, we calculated the median and average number of words and characters for each question, left context and right context. We also calculated the average number of words belonging to the different lexicons found in each question (table 7). For the *doubt* category, we noticed that the median for the number of characters of the questions in this category is 11, which is a rather short length compared to the questions in the *doubt* category that were misclassified as *opinion* or *will* for which we find respectively medians of 17 and 21. The well classified questions of the *doubt* category are therefore generally shorter than the other questions. For this category, the presence of interrogative words is also significant since the average number of words is higher than

	O→O	O→D	O→W	W→W	W→O	W→D	D→D	D→O	D→W
med. words q.	6	5	5	6	5	5	3	5	6
med. words l.	96	92	93.5	87.5	95	88	92	85	101
med. words r.	97	97	81	88.5	97	88	84	94	92
med. char. q.	21	19	21	21	19	17	11	17	21
med. char. l.	331	326	330.5	312.5	322	302	327	312	341
med. char. r.	344	338	301	310	345	317	299	331	309.5
speech verbs	0.03	0.04	0.07	0.04	0.04	0.02	0.05	0.08	0.07
movement verbs	0.06	0.07	0.12	0.19	0.12	0.1	0.08	0.11	0.16
interrogative words	0.34	0.53	0.47	0.33	0.35	0.37	0.58	0.46	0.59
interjections	0.54	0.27	0.25	0.18	0.37	0.3	0.08	0.3	0.24
feelings	0.41	0.09	0.12	0.03	0.09	0.13	0.04	0.16	0.1
modal words	0.31	0.07	0.09	0.13	0.26	0.13	0.09	0.11	0.07

Table 7: Measures for the predictions of opinion (O), will (W) and doubt (D) categories

elsewhere (0.58). The questions in this category that were predicted in *will* have an average of 0.16 movement verbs, a number that is close to the average number of movement verbs in the correct predictions of the *will* category, which is around 0.19. We observe for the *opinion* category that well classified questions have on average a strong proportion of words from the lexicons of feelings (0.41), interjections (0.55) and modal words (0.31) compared to the other categories and misclassified questions in the *opinion* category. Indeed, when these are classified in *will*, they have a mean of 0.12 words in the lexicon of feelings, 0.25 words in the interjections and 0.09 words being modal words. This is also confirmed for the questions that are *opinion* classified as *doubt*, for which we also see that the average number of interrogative words is 0.53, a number close to the one obtained for the well classified questions from the *doubt* category which is 0.58. Finally, when we focus on predictions for the *will* category, we notice a stronger proportion of movement verbs than elsewhere, with an average of 0.19 for questions that are well classified in *will*. This number is lower for questions misclassified as *doubt* (0.1) and is closer to the average number of movement verbs for good predictions of *doubt* which is 0.08. For the questions of *will* classified in *opinion*, we observe a higher frequency of words from the interjections (0.37) and modal words (0.26) lexicons, which are characteristics of the *opinion* category. In addition to this interpretation of the results, we note that most of the scores are very close since they are around 0.6 for precision, recall and f-score. This shows the difficulties and challenges that can be encountered in identifying relevant features to discriminate between utterances that include implicit content, especially in our situation since we only use written transcriptions of oral recordings that do not include information related to prosody. However, the different linguistic features cited above that seem to explain the presence of a given question in a given category are significant, with a p-value lower than 0.05 for the variables that were compared with a t-test. There is certainly a trend of classification as we can see both from the results obtained with Weka and from our implementation of Random Forest but the results could be improved, in particular by identifying new features and completing existing ones. For example, we could add new terms to our lexicons to make them more precise and complete since they are information that seems relevant to the classification task.

7. Conclusion

The researches conducted as part of this work were only based on transcriptions of oral and spontaneous recordings. The main goal was to detect the implicit and non-literal aspects of questions asked during meals, without using prosodic clues. The contextualization of this work along with the annotation of our corpus of 3647 questions allowed us to develop a typology of intentions in questions, defined by the *opinion*, *will* and *doubt* categories. The evaluation of the annotation and typology using an online form showed its consistency, since the inter-annotator agreements obtained were higher than 0.6 for 50% of them. By integrating selected features into our Random Forest classification algorithm, we were able to reach values of about 0.62 for accuracy, recall and f-score. There are many perspectives for this research. The main goal would be to improve the classification performance of our model and one of the possibilities to try would be to specify and add relevant linguistic features such as new entries in our lexicons. It would also be interesting to specify our typology by defining sub-labels. An interesting outcome of this work would be to extend the typology of intentions in questions to other types of utterances. It would also be a possibility to add prosodic cues to the text that would probably lead to higher and better scores.

8. Bibliographical References

Alexanderssony Bianka, V., Tsutomu, B.-w., Elisabeth Maiery, F., Reithingery Birte, N., Melanie Siegelyy, S., Alexandersson, J., Buschbeck, B., Fujinamiz, T., Reithingery, N., Schmitzx, B., and Siegelyy, M. (1997). Dialogue acts in verbmobil-2. 05.

Allen, J. and Core, M. (1997). Draft of damsl: Dialog act markup in several layers.

Allen, J. F. and Perrault, C. R. (1980). Analyzing intention in utterances. *Artificial intelligence*, 15(3):143–178.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Austin, J. L. (1962). *How to do things with words*. William James Lectures. Oxford University Press.

Benzitoun, C., Fort, K., and Sagot, B. (2012). Tcof-pos: un corpus libre de français parlé annoté en morphosyntaxe.

Bunt, H. (2005). A framework for dialogue act specification. 01.

Chen, Z., Liu, B., Hsu, M., Castellanos, M., and Ghosh, R. (2013). Identifying intention posts in discussion forums. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1041–1050.

Jurafsky, D. and Shriberg, E. (1997). Switchboard swbd-damsl shallow-discourse-function annotation coders manual.

Karoui, J., Benamara, F., and Moriceau, V. (2019). *Détection automatique de l'ironie: Application à la fouille d'opinion dans les microblogs et les médias sociaux*. ISTE Group.

Kerbrat-Orecchioni, C. (1986). *L'implicite*. A. Colin, Paris.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Moldovan, C., Rus, V., and Graesser, A. (2011). Automated speech act classification for online chat. pages 23–29, 01.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees, intl. In *Conference on New Methods in Language Processing. Manchester, UK*.

Searle, J. (1975). Indirect speech acts. *Pragmatics: Critical Concepts*, 5:639–657, 01.

9. Language Resource References

Baude, O. and Dugua, C. (2011). (re) faire le corpus d'orléans quarante ans après: quoi de neuf, linguiste? *Corpus*, (10):99–118.

Eshkol-Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., and Tellier, I. (2011). Un grand corpus oral disponible: le corpus d'orléans 1 1968-2012.

Stosic, Dejan and Aurnague, Michel. (2017). *DinaVmouv: Description, INventaire, Analyse des Verbes de MOUVement. An annotated lexicon of motion verbs in French.*