Rapid FrameNet annotation of spoken conversation transcripts

Jeremy Trione, Frederic Bechet, Benoit Favre, Alexis Nasr
Aix Marseille University, CNRS, LIF
Marseille, France
firstname.lastname@lif.univ-mrs.fr

Abstract

This paper presents the semantic annotation process of a corpus of spoken conversation transcriptions recorded in the Paris transport authority call-centre. The semantic model used is a FrameNet model developed for the French language. The methodology proposed for the rapid annotation of this corpus is a semi-supervised process where syntactic dependency annotations are used in conjunction with a semantic lexicon in order to generate frame candidates for each turn of a conversation. This first hypotheses generation is followed by a rule-based decision module in charge of filtering and removing ambiguities in the frames generated. These rules are very specific, they don't need to generalize to other examples as the final goal of this study is limited to the annotation of this given corpus, on which a statistical frame parser will finally be trained. This paper describes this methodology and give examples of annotations obtained. A first evaluation of the quality of the corpus obtained is also given on a small gold corpus manually labeled.

1 Introduction

Parsing human-human conversations consists in enriching text transcription with structural and semantic information. Such information include sentence boundaries, syntactic and semantic parse of each sentence, para-semantic traits related to several paralinguistic dimensions (emotion, polarity, behavioral patterns) and finally discourse structure features in order to take into account the interactive nature of a conversation.

The applicative context of this work is the automatic processing of human-human spoken conversations recorded in customer service telephone call centers. The goal of processing such data is to take advantage of cues in order to automatically obtain relevant summaries and reports of such conversations for speech mining applications. These processes are needed because coarse-grained analyses, such as keyword search, are unable to capture relevant meaning and are therefore unable to understand human dialogs.

Performing semantic parsing on spoken transcriptions is a challenging task Coppola et al. (2009). Spoken conversation transcriptions have characteristics that make them very different to process from written text Tur and De Mori (2011).

- non-canonical language: spontaneous speech represents a different level of language than the *canonical* one used in written text such as newspaper articles;
- *noisy messages*: for spoken messages, automatic speech transcription systems make errors, especially when dealing with spontaneous speech;
- relevant and superfluous information: redundancy and digression make conversation messages prone to contain superfluous information that need to be discarded;
- conversation transcripts are not self-sufficient: for spoken messages, even with a perfect transcription, non-lexical information (prosody, voice quality) has to be added to the transcription in order to convey speakers' intention (sentiment, behavior, polarity).

The general process of parsing conversations can be divided into three levels: conversational data pre-processing; syntactic parsing; semantic parsing.

The pre-processing level involves the transcription (automatic or manual) of the spoken content and the segmentation into speakers' turns and sentence-like units.

The syntactic parsing level aims to uncover the word relationships (e.g. word order, constituents) within a sentence and support the semantic layer of the language-processing pipeline. Shallow syntactic processes, including part-of-speech and syntactic chunk tagging, are usually performed in a first stage. One of the key activities described in this paper is the adaptation of a syntactic dependency parser to the processing of spontaneous speech. The syntactic parses obtained are used in the next step for semantic parsing.

The semantic parsing level is the process of producing semantic interpretations from words and other linguistic events that are automatically detected in a text conversation or a speech signal. Many semantic models have been proposed, ranging from formal models encoding *deep* semantic structures to shallow ones considering only the main topic of a document and its main concepts or entities. We use in this study a FrameNet-based approach to semantics that, without needing a full semantic parse of a message, goes further than a simple flat translation of a message into basic concepts: FrameNet-based semantic parsers detect in a sentence the expression of frames and their roles Gildea and Jurafsky (2002). Because frames and roles abstract away from syntactic and lexical variation, FrameNet semantic analysis gives enhanced access to the meaning of texts: of the kind *who does what, and how where and when ?*.

We describe in this paper the rapid semantic annotation of a corpus of human-human conversations recorded in the Paris public authority call-centre, the *RATP-DECODA* corpus presented in Bechet et al. (2012). This corpus is presented in section 2. The methodology followed is a semi-supervised process where syntactic dependency annotations are used in conjunction with a semantic lexicon in order to generate frame candidates for each turn of a conversation. This first hypotheses generation is followed by a rule-based decision module in charge of filtering and removing ambiguities in the frames generated. Section 3 describes the adaptation of syntactic parsing models to the processing of spontaneous speech. Section 4 presents the FrameNet semantic model derived for annotating these call-centre conversations, and finally section 5 reports some evaluation results on a small gold corpus manually annotated.

2 The RATP DECODA corpus

The RATP-DECODA[1] corpus consists of 1514 conversations over the phone recorded at the Paris public transport call center over a period of two days Bechet et al. (2012). The calls are recorded for the caller and the agent, totaling over 74 hours of French-language speech.

The main problem with call-center data is that it often contains a large amount of personal data information, belonging to the clients of the call-center. The conversations collected are very difficult to anonymized, unless large amounts of signal are erased, and therefore the corpus collected can't be distributed toward the scientific community. In the DECODA project we are dealing with the call-center of the Paris transport authority (RATP). This applicative framework is very interesting because it allows us to easily collect large amount of data, from a large range of speakers, with very few personal data. Indeed people hardly introduce themselves while phoning to obtain bus or subway directions, ask for a lost luggage or for information about the traffic. Therefore this kind of data can be anonymized without erasing a lot of signal.

Conversations last 3 minutes on average and usually involve only two speakers but there can be more speakers when an agent calls another service while putting the customer on wait. Each conversation is anonymized, segmented and transcribed. The call center dispenses information and customer services, and the two-day recording period covers a large range of situations such as asking for schedules, directions, fares, lost objects or administrative inquiries.

Because speech that can be found in a call-centre context is highly spontaneous, many speech-specific phenomenon such as disfluencies appear with a high frequency. In the RATP-DECODA corpus the

[1] The RATP-DECODA corpus is available for research at the Ortolang SLDR data repository: http://sldr.org/sldr000847/fr

disfluencies considered correspond to repetitions (e.g. *le le*), discourse markers (e.g. *euh*, *bien*) and false starts (e.g. *bonj-*).

Table 1 displays the amount of disfluencies found in the corpus, according to their types, as well as the most frequent ones. As we can see, discourse markers are by far the most frequent type of disfluencies, occurring in 28% of the speech segments.

disfluency type	# occ.	% of turns	10 most frequent forms
discourse markers	39125	28.2%	[euh] [hein] [ah] [ben] [voila] [bon] [hm] [bah] [hm hm] [coutez]
repetitions	9647	8%	[oui oui] [non non] [c’ est c’ est] [le le] [de de] [ouais ouais] [je je] [oui oui oui] [non non non] [a a]
false starts	1913	1.1%	[s-] [p-] [l-] [m-] [d-] [v-] [c-] [t-] [b-] [n-]

Table 1: Distribution of disfluencies in the RATP-DECODA corpus

Because of this high level of spontaneity, syntactic models such as Part-Of-Speech models or dependency models that were trained on written text have to be adapted. This semi-supervised annotation method is presented in the next section.

3 Semi-supervised syntactic annotation

It has been shown in Bechet et al. (2014) that a great improvement in tagging and parsing performance can be achieved by adapting models to the specificities of speech transcripts. Disfluencies can be integrated into the models without negative impact on the performance, if some annotated adaptation data is available.

In order to adapt the tagger and parser to the specificities of oral French, we have parsed the RATP-DECODA corpus with the MACAON tagger and dependency parser Nasr et al. (2011) and developed an iterative process consisting in manually correcting errors found in the automatic annotations thanks to a WEB-based interface Bazillon et al. (2012).

This interface allows writing regular expressions on the POS and dependency tags and the lexical forms in order to correct the annotations on the whole RATP-DECODA corpus. Then the parser is retrained with this corrected corpus. When the error rate computed on a development set is considered acceptable, this correction process stops. The resulting corpus, although not perfect, constitutes our training corpus, obtained at a reasonably low price compared to the whole manual annotation process of the corpus. This process is described by figure 1.

The accuracy of the new parser is far above the accuracy of the parser trained on written text (French TreeBank) : from 65.8% to 85.9% for Unlabeled Attachment Score (UAS) and from 58.3% to 83.8% for Labeled Attachment Score (LAS). The performances of the parser can be compared to the performances of a parser for written data despite the fact that the parser has been trained on a partially manually corrected corpus.

Two reasons can explain this result. The first one is that the DECODA corpus has a quite restricted and specific vocabulary and the parser used is quite good at learning lexical affinities. The second one is that the DECODA corpus has a rather simple syntax with utterances generally restricted to simple clauses and less common ambiguities, such as prepositional attachment and coordination, than written texts.

One crucial issue is the amount of manual supervision needed to update the models. If a whole annotation of the corpus is needed, the process will be too costly whatever gain in performance is achieved. We display in 2 the learning curve of the POS tagger, starting from a generic model trained on the French

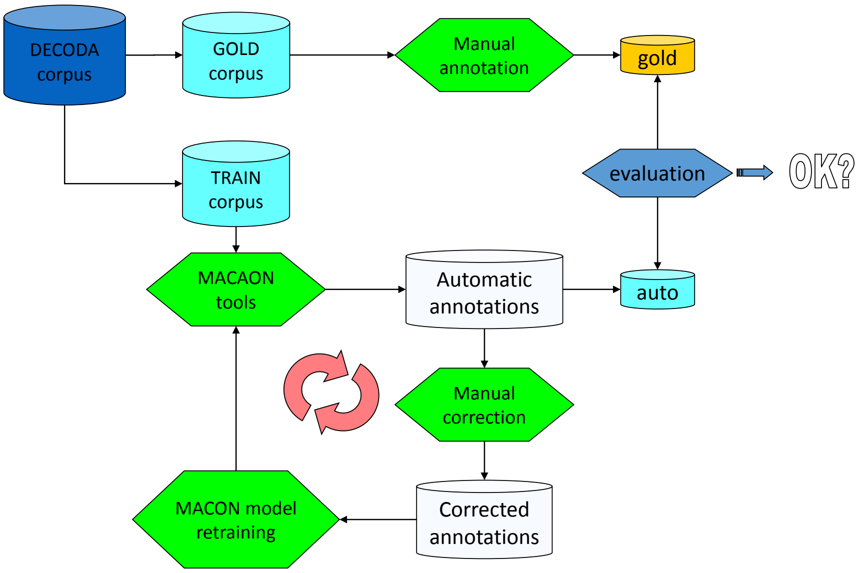

Figure 1: Semi-supervised adaptation process

TreeBank, and including some manual annotation on the target corpus. As we can see, even a very limited annotated subset of the corpus can boost performance: by adding as little as 20 dialogs, the POS error rate drops by more than half (green curve) from 19% to 8%.

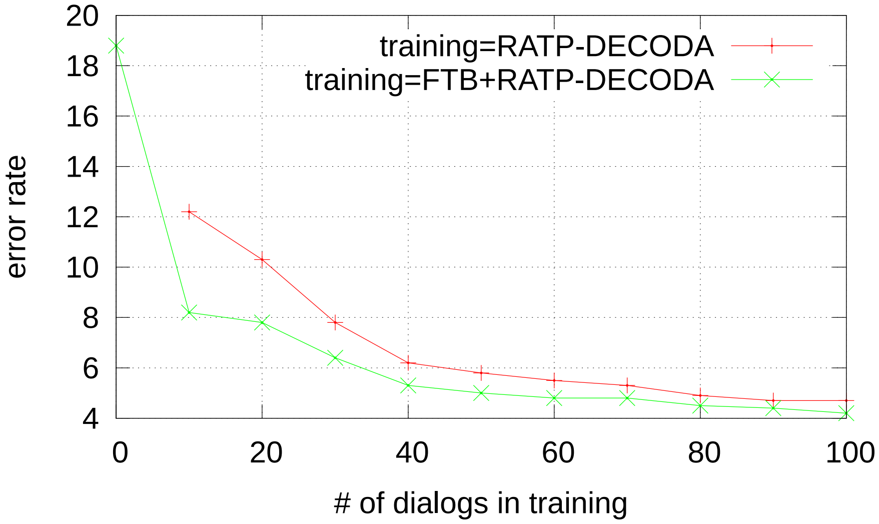

Figure 2: Learning curve of the POS tagger with and without the FTB on the RATP-DECODA corpus

4 From syntactic to semantic annotation

Annotating manually with frame labels a corpus like the RATP-DECODA corpus is very costly. The process we followed in this study is to take advantage of both syntactic annotations and external semantic resources for performing this annotation at a very low cost.

We use in this study a FrameNet model adapted to French through the ASFALDA project. The current model, under construction, is made of 106 frames from 9 domains. Each frame is associated to a set of Lexical Units (LU) that can trigger the occurrence of a frame in a text. The first step, in annotating a corpus with FrameNet, is to detect triggers and generate frame hypotheses for each detection. We did this process on the RATP-DECODA corpus and found 188,231 potential triggers from 94 different frame definitions.

The semi-supervised annotation process presented in this paper consists, for each LU, in searching in the output of the parser for the dependencies (such as subject or object) of each trigger. This first annotation is further refined thanks to semantic constraints on the possible dependent of a given LU, considering the domain of the corpus.

The first step in our annotation process is to select a set of triggers on which frame selection will be applied. In this study we limited our trigger set to the 200 most frequent verbs. By analyzing several triggers on the corpus, we have defined the main domains and frames that we will use to annotate the corpus.

Seven domains were considered:

- Motion.

  Motion frames involve a theme which goes to an area. A vehicle can be use, and several other parameter can be used like the source, the path used, the time, ...
  Most used frames: Motion, Path_shape, Ride_vehicle, Arriving.
  Examples:

Je	voudrais	aller	a Juvisy.
Theme		Motion	Area

I	would like to	go	to Juvisy.
Theme		Motion	Area

- Communication.

  Communication frames involve a communicator sending a message to an addressee. While our corpus is about call center conversation these frame are really important to describe the structure of the call.
  Most used frames: Communication, Request, Communication_response.
  Examples:

je	vous	appelle	parce qu'on m'a redirige vers vous.
Communicator	Addressee	Request	Message

I	call	you	because I was redirected to you.
Communicator	Request	Addressee	Message

- Sentiment expression.

  Sentiment expression frames involve a communicator and an addressee. In our case these sentimental detections can evaluate the behavior of the people in the conversation.
  Most used frames: Judgment_direct_address, Desiring.
  Examples:

je	vous	remercie	beaucoup.
Communicator	Addressee	Judgment_direct_address	Degree

I	thank	you	a lot.
Communicator	Addressee	Judgment_direct_address	Degree

- Commerce.

  Commerce frames involve a buyer, some goods and sometimes a seller. These frames are pretty frequent in every call about tariff or fine paying.
  Most used frames: Commerce_Buy, Commerce_pay.
  Examples:

Vous	devez	acheter	un ticket.
Buyer		Commerce_buy	Goods

You	have	to	buy	a ticket.
Buyer			Commerce_buy	Goods

- Action.

  We call action frames every frames that involve an action linked to a person. These kind of frames are frequent in conversations that deal with misfortune of the caller.
  Most used frames: Losing, Giving, Intentionally_affect.
  Examples:

J'	ai	perdu	mon telephone	dans le bus 38.
Owner		Losing	Possession	Place

I	lost	my phone	in the bus 38.
Owner	Losing	Possession	Place

As mentioned above, all these frames are triggered by the 200 most frequent verbs in the corpus. However, FrameNet was not specially designed for spoken conversations and we had to extend it with two new frames specific to this kind of data:

- Greetings.

  This frame is triggered to represent the opening and the closing of a call. We use the same frame in both cases (”Hello!”, ”Goodbye!”).
  Examples:

Bonjour	monsieur.
Hello	Addressee

Hello	sir.
Hello	Addressee

- Agreement.

  Agreement is a crucial frame in a dialog context. Detecting positive or negative answers to direct Boolean questions in the context of a call-centre dialog is very important. The Agreement frames refer to every mark of agreement (”yes”, ”of course”, . . .).
  Examples:

[...]	d'accord	merci.
	Agreement	

[...]	alright	thank you.
	Agreement	

Once this Frame selection process has been done, we are able to produce Frame hypotheses directly from our trigger list of verbs and derive Frame elements from syntactic annotations. There can be only one frame candidate by trigger. If a trigger can correspond to several frames, we use a rule-based approach to choose one frame according to the context. Because the semantic domain of our corpus is rather limited, there are not many ambiguities and most of the verbs only corresponds to one frame, therefore the set of rules needed to remove ambiguities is very limited and restricted to the 5 most frequent verbs (such as *aller - to go*).

To write these rules we selected examples of these ambiguous verbs on our corpus, and wrote rules taking into account the lexical and syntactic context of these verbs. Only six rules were needed, five of them focused on disambiguating motion frames which are the most ambiguous frame in our corpus. Table 2 show an example of rule.

Trigger	Aller	
Rule	Trigger + non verb = motion frame	
Example 1	Un conseiller *va* prendre votre appel.	Not a motion frame
Example 2	Il faut *aller* directement en agence.	Motion frame

Table 2: Example of syntactic rule.

This example illustrates the ambiguity of the trigger verb ”*aller*” (*to go*). This verb is very frequent in French, particularly in spontaneous conversations. Similarly to English, this is a polysemic verb that can means ”*motion*” as well as an ongoing action (e.g. ”*I'm going to do something*”). A simple rule checking if this verb is associated to another verb or to an object can remove this ambiguity (example 1 in 2).

For each rule proposed, we checked on a reference corpus (*gold* corpus presented in the next section) how many ambiguities were correctly resolved, and we kept only the most efficient ones. This process was quite fast as it was done on the Frame hypotheses already produced and checked automatically on a small gold corpus. Just a few iterations allowed us to produce the small set of rules that removed most of the ambiguities of the most frequent verbs.

The Frame selection process consists now, for each trigger in a conversation, to check first if this trigger is ambiguous or not. If it is, a rule should be applied to disambiguate it. If the trigger is not ambiguous, we simply annotate the sentence with the corresponding frame from the dictionary. Due to our very specific corpus, we have a low number of ambiguities and therefore a low number of rules.

5 Evaluation of Frame selection

A small gold corpus was manually defined and annotated. The automatic rule-based Frame selection process is evaluated on this corpus, as presented in figure 1. Our gold corpus is a set on 21 conversations from the RATP-DECODA corpus. These conversations were fully manually annotated by one annotator. The tables below give a representation of the distribution of the frames on this subcorpus, comparing manual annotation and automatic annotation.

Table 3 show us that on average there is at least one trigger per speaker turn. Moreover, we can already tell that the automatic annotation predicts more triggers than the human annotator, and get more variability in the frame chosen. In Table 4 we find our main domain on the RATP-DECODA corpus through the frames. In fact ”Hello” and ”Judgment_direct_address” represent the structure of the call (opening and closing), while ”Request”, ”Losing”, ”Motion” and ”Commerce_buy” can easily represent the reason of the call.

	Manual annotation	Automatic annotation
Number of Frames per Conversation	23.67	31.33
Number of Frames per speaker turn	0.97	1.24
Number of different frames	26	37

Table 3: Frames distribution on the gold corpus.

Manual Annotation		Automatic annotation	
Frame name	Occurrences	Frame name	Occurrences
Agreement	161	Agreement	216
Hello	95	Hello	95
Judgment_direct_address	59	Motion	45
Motion	33	Communication	34
Request	21	Judgment_direct_address	27
Waiting	20	Desiring	20
Awareness	18	Awareness	19
Communication	15	Intentionally_affect	16
Losing	14	Possibility	12
Commerce_buy	9	Waiting	11

Table 4: Top 10 used frames on the gold corpus.

The quality of the automatic prediction, with respect to the gold corpus, is presented in Table 5. There are different levels of evaluation (trigger selection, frame level, frame element level, span, ...). We chose to evaluate our annotation at the frame level. In other words, we evaluate if a trigger produced the correct frame.

	Recall	Precision	f-measure
Automatic annotation	83.33	94.54	88.58

Table 5: Evaluation on the automatic annotation on the gold corpus.

These first results are satisfying at the precision level is 94.5% of Frame predictions are correct. The recall measure is lower but satisfactory considering that we limited the frame selection process to only the most frequent verbs. A bigger gold corpus is now needed in order to assess the final quality of this corpus.

6 Conclusion

We have presented in this paper a methodology for the rapid annotation of spoken conversation corpus recorded in a French call-centre. This semi-supervised process uses syntactic dependency annotations in conjunction with a FrameNet semantic lexicon. The rule-based decision module in charge of filtering and removing ambiguities in the frames generated is evaluated at each learning cycle on a small manually labelled gold corpus. The first evaluation described in this paper validate this approach by showing good precision scores with an acceptable recall. This corpus will now be used to train a statistical frame parser such as Das et al. (2014) that will be evaluated on other call-centre conversation transcriptions.

References

Bazillon, T., M. Deplano, F. Bechet, A. Nasr, and B. Favre (2012). Syntactic annotation of spontaneous speech: application to call-center conversation data. In *Proceedings of LREC*, Istambul.

Bechet, F., B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, and E. Arbillot (2012). Decoda: a call-centre human-human spoken conversation corpus. In *LREC*, pp. 1343–1347.

Bechet, F., A. Nasr, and B. Favre (2014). Adapting dependency parsing to spontaneous speech for open domain spoken language understanding. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Coppola, B., A. Moschitti, and G. Riccardi (2009). Shallow semantic parsing for spoken language understanding. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 85–88. Association for Computational Linguistics.

Das, D., D. Chen, A. F. Martins, N. Schneider, and N. A. Smith (2014). Frame-semantic parsing. *Computational Linguistics 40*(1), 9–56.

Gildea, D. and D. Jurafsky (2002, September). Automatic labeling of semantic roles. *Comput. Linguist. 28*(3), 245–288.

Nasr, A., F. Béchet, J.-F. Rey, B. Favre, and J. Le Roux (2011). Macaon: An nlp tool suite for processing word lattices. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pp. 86–91. Association for Computational Linguistics.

Tur, G. and R. De Mori (2011). *Spoken language understanding: Systems for extracting semantic information from speech.* John Wiley & Sons.

7 Annex

Abandonment
Accuracy
Activity_pause
Activity_prepare
Activity_resume
Adjusting
Agreement
Agree_or_refuse_to_act
Amalgamation
Arriving
Assessing
Assistance
Attaching
Attempt
Avoiding
Awareness
Becoming
Becoming_a_member
Becoming_aware
Being_in_effect
Being_in_operation
Borrowing
Breaking_apart
Breaking_off
Breathing
Bringing
Building
Bungling
Canceling
Categorization
Causation
Cause_change
Cause_change_of_strength
Cause_harm
Cause_motion
Cause_to_experience
Cause_to_perceive
Certainty
Change_accessibility
Change_event_time
Change_operational_state
Change_position_on_a_scale
Chatting
Choosing
Closure
Coming_to_be
Commerce_buy
Commerce_collect
Commerce_pay
Commerce_sell
Commitment
Communication
Communication_response
Complaining
Compliance
Conferring_benefit
Contacting
Containing
Contingency
Contrition
Control
Cotheme
Deciding
Defending
Departing
Deserving
Desirable_event
Desiring
Difficulty
Duration_description
Duration_relation
Emitting

Emphasizing	Emptying	Erasing
Estimating	Event	Evidence
Existence	Expend_resource	Expensiveness
Experiencer_focus	Experiencer_obj	Explaining_the_facts
Feeling	Filling	Forgiveness
Forming_relationships	Getting	Give_impression
Giving	Givinig	Grasp
Halt	Having_or_lacking_access	Hello
Hiding_objects	Hiring	Impact
Ingestion	Intentionally_affect	Intentionally_create
Judgment	Judgment_direct_address	Justifying
Labeling	Leadership	Lending
Locale_closure	Locating	Location_in_time
Losing	Making_arrangements	Memory
Motion	Name_conferral	Offering
Operating_a_system	Opinion	Participation
Path_shape	Perception_active	Performers_and_roles
Placing	Possession	Possibility
Posture	Practice	Predicting
Preference	Prevarication	Process_continue
Process_end	Processing_materials	Process_start
Questioning	Receiving	Redirecting
Reliance	Removing	rentraire
Repayment	Replacing	Reporting
Request	Required_event	Reserving
Reshaping	Residence	Resolve_problem
Respond_to_proposal	Ride_vehicle	Run_risk
Scrutiny	Self_motion	Self_otion
Sending	Sign	Similarity
Simultaneity	Spelling_and_pronouncing	Statement
Storing	Studying	Subscribing
Success_or_failure	Sufficiency	Surpassing
Taking_sides	Telling	Text_creation
Theft	Topic	Transfer
Trap	Triggering	Using
Using_resource	Verification	Wagering
Waiting	Warning	Work
Working	Negation	

Table 6: Semantic Frames chosen to annotate the RATP-DECODA corpus