# DECODA: a call-center human-human spoken conversation corpus

*F. Bechet*[1], *B. Maza*[2], *N. Bigouroux*[3], *T. Bazillon*[1], *M. El-Bèze*[2], *R. De Mori*[2], *E. Arbillot*[4]

[1] Aix Marseille Univ, LIF-CNRS, Marseille, France
frederic.bechet@lif.univ-mrs.fr thierry.bazillon@lif.univ-mrs.fr
[2] Universite d'Avignon, Avignon, France
renato.demori@univ-avignon.fr marc.elbeze@univ-avignon.fr benjamin.maza@univ-avignon.fr
[3] Sonear, Paris, France
nicolas.bigouroux@sonear.com
[4] RATP, Paris, France
eric.arbillot@ratp.fr

**Abstract**

The goal of the DECODA project is to reduce the development cost of Speech Analytics systems by reducing the need for manual annotation. This project aims to propose robust speech data mining tools in the framework of call-center monitoring and evaluation, by means of weakly supervised methods. The applicative framework of the project is the call-center of the RATP (Paris public transport authority). This project tackles two very important open issues in the development of speech mining methods from spontaneous speech recorded in call-centers : robustness (how to extract relevant information from very noisy and spontaneous speech messages) and weak supervision (how to reduce the annotation effort needed to train and adapt recognition and classification models). This paper describes the DECODA corpus collected at the RATP during the project. We present the different annotation levels performed on the corpus, the methods used to obtain them, as well as some evaluation of the quality of the annotations produced.

**Keywords:** Speech Analytics, Conversational Speech, Syntactic and Semantic annotations

## 1. Introduction

Speech Analytics is a new term used to describe the extraction of information from speech data. Three main research domains are involved: Automatic Speech Recognition (ASR), Natural Language Processing (NLP) and Data Mining. The first level in the Speech Analytics process is to translate a speech signal into several sequences of symbols. These symbols can describe both the lexical hypotheses about the linguistic content of a speech segment and the acoustic and prosodic properties of it such as: environmental noise, quality of speech, speaker, prosodic cues, etc.
This first level is made by the Signal Processing and Automatic Speech Recognition processes (ASR). The lexical hypotheses produced by the ASR modules are processed by the NLP modules. Then the Data Mining processed can extract from a large corpus some information about the behaviors of the callers in the context of a call-center.
The multiplication of call-centers and the low storing cost of audio data have made available the recording of very large audio database containing conversations between operators and customers. From the companies point of view, these call-centers are a strategic interface toward their customers. Therefore the need for automatic tools allowing to monitor such services and mine the dialogs recorded is very high, as pointed out by the development of industrial products, mainly by American companies, in the field of Speech Analytics: Nuance, Verint, CallMiner, BBN-Avoke, Nexidia, Autonomy eTalk.
These products can provide two kinds of services:

- punctual analysis of large dialog corpora for data mining purposes, like detecting a problem in the call-center behavior, or extracting knowledge about the call-center performance;
- periodic analysis, or monitoring of the call-center by a day-by-day analysis of the call-center dialog logs.

An example of a speech analytics process is displayed in figure 1, where a conversation between an agent and a caller is analyzed according to several dimensions: from the segmentation of the dialogs, the extraction of concepts and meanings to the summarization process and the behavioral analysis.
All the products need the manual annotation of large amount of speech data in order to train and adapt the recognition and classification models. This task has to be done regularly in order to adapt the models to the new behaviors of a call-center, leading to an increase in the cost development of Speech Analytics tools.
The goal of the DECODA[1] project is to reduce the development cost of Speech Analytics systems by leveraging the manual annotation effort needed. This project develops robust speech data mining tools in the framework of call-center monitoring and evaluation, by means of weakly supervised methods. This project tackles two very important open issues in the development of speech mining methods from spontaneous speech recorded in call-centers: robustness (how to extract relevant information from very noisy and spontaneous speech messages) and weak supervision (how to reduce the annotation effort needed to train and adapt recognition and classification models). The applicative framework of the project is the call-center of the RATP (Paris public transport company).
This paper describes the DECODA corpus collected at the RATP during the project. We present the different annotation levels performed on the corpus, the methods used to

[1] http://decoda.univ-avignon.fr

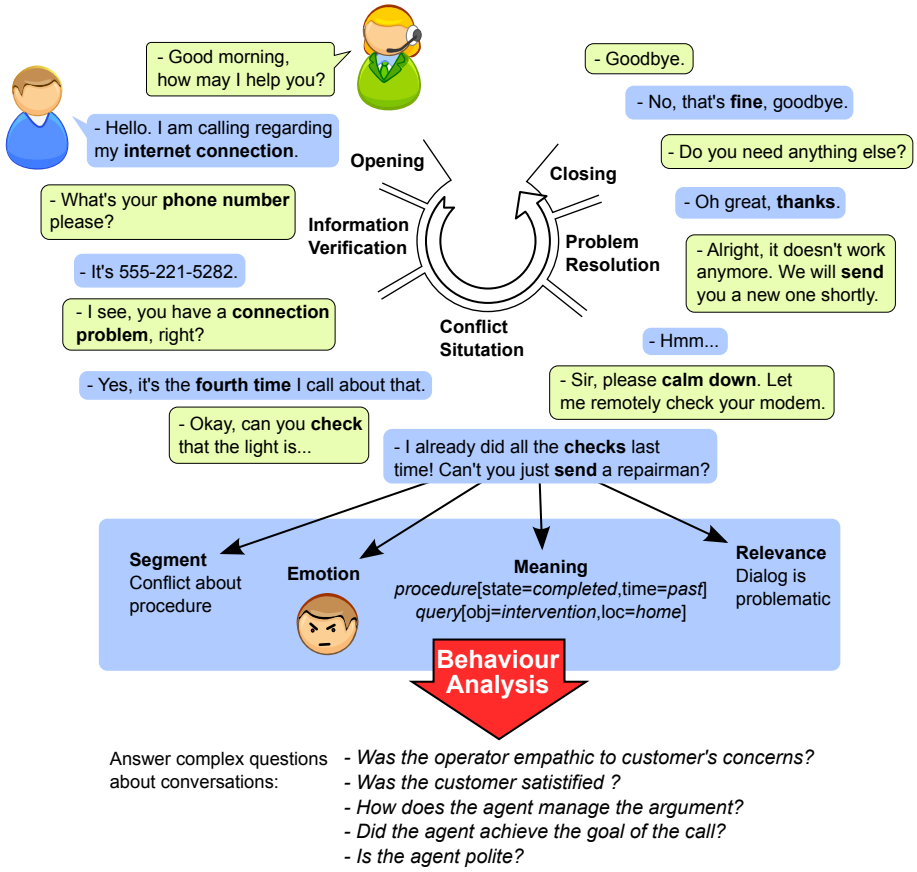


Figure 1: Example of speech analytic processes on a call-center conversation

obtain them, as well as some evaluations about the quality of the annotations produced.

## 2. Related work: previous corpora of spoken conversation

Most of the studies of human-human spoken conversations have been limited in scope due to either size of the corpora or realistic task scenarios with only a few exceptions (Tur and Hakkani-Tür, 2011). Several corpora have been collected in the past and used for conversation analysis. The first example is probably Switchboard (Godfrey et al., 1992), a large corpus of telephone conversations involving about 500 speakers who were given a pre-determined topic to talk about. The research originally focused on speech and speaker recognition rather than speech understanding. More natural and multilingual/accented conversational speech was collected later in the Call-Home and Call-Friend corpora. Goal-orientated human-human dialogues are available in the TRAINS (Allen, 1994) corpus for transportation planning, in the Monroe (Stent, 2000) corpus for disaster handling, and in the MapTask (Anderson et al., 1991) corpus for giving directions on a map.

More recently, multi-party human-human conversations (or meetings) and lectures have been considered, for example in the European funded projects AMI (Augmented Multi-party Interaction) and AMIDA (Renals et al., 2007), and, in the USA, the CALO (Cognitive Assistant that Learns and Organizes) project that concentrated on conference-room meetings with small scale in-vitro experiments. A corpus of university lectures, some of them including interactions with the audience, was used in the CHIL (Computers in the Human Interaction Loop) European funded project. These types of corpora inspired research on human-human conversation mostly with specific purposes such as browsing and learning dialogue structures (e.g. dialogue act tagging, topic segmentation, summarization).

Another source of human-human conversation corpora is broadcast conversations such as debates or interviews. This kind of data has the advantage of being easily accessible from corpus distribution agencies such as LDC or ELDA through large corpora of broadcast shows. However broadcast conversations are far from being truly spontaneous as many of the speakers are professionals.

Telecommunication and call-center companies involved in Customer Relationship Management (CRM), internally record call-center conversational speech data, which are the focus of this paper. Recently some European and national research programs have devoted resources to acquiring corpora through collaboration between service providers and research labs.

In the LUNA FP6 project human-machine and human-human conversations have been collected in the context of Information Technology (IT) help-desk interactions.

In France the CallSurf project (Garnier-Rizet et al., 2008) has collected a large corpus of conversations through an EDF call-center. Part of this corpus has been manually tran-

scribed and annotated.

## 3. The DECODA corpus

The main problem with call-center data is that it often contains a large amount of personal data informations, 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.
The DECODA corpus currently collected within the project has been fully anonymized, manually segmented and transcribed, then annotated at various linguistic levels. The current state of the corpus is made of 1514 dialogs, corresponding to about 74 hours of signal.
The average duration of a dialog is about 3 minutes. The distribution of the duration is given in table 1.
As we can see most of dialogs are quite short, however 12% of them are longer than 5 minutes.
The collect of the first 1514 dialogs of the DECODA corpus has been done during 2 days of traffic, in order to have a good picture of all the possible requests expressed by the callers during a whole day.

## 4. Speech transcription

The speech transcription process involved three steps: the *anonymization* process; the *segmentation* process and the *transcription* process itself. The goal of the anonymization process is to suppress from the audio files all the personal references such as person names, phone numbers or personal addresses. This is a fully manual process, made before any other tasks, and consisting of replacing all these personal references by a beep. Thankfully, as mentioned before, there is very little personal information in the corpus, except in the first sentence where the two speakers identify themselves.
The second step is a segmentation process. The whole audio file is first segmented into chapters: *opening*, *problem presentation*, *problem resolution*, *closing*. Then a speaker segmentation process is performed in order to separate the turns of each speaker. This step is necessary since unfortunately the DECODA corpus is recorded with only one channel mixing the voices of the callers and the operators. Finally each turn is segmented into sentence-like units (called *segments*), separated within each other by a pause longer than a given threshold.
All the sentence-like units have been manually transcribed using the tool Transcriber and the conventions of the ESTER (Galliano et al., 2009) program.
The DECODA corpus contains so far 1514 dialogs, resulting into 96103 speakers turns, 106691 segments for a total of 482745 words after tokenization. The most frequent word in this corpus is the discourse marker "*euh*" (*hum*), indicating the spontaneity of the speech collected. The total vocabulary of the corpus is 8806 words.

## 5. Semantic annotation

At the semantic level each dialog has been manually tagged according to the RATP ontology. The repartition of the 10 top call-types is given in table 2.

| Calltype | % |
|---|---|
| Info Traffic | 22.5 |
| Route planning | 17.2 |
| Lost and Found | 15.9 |
| Registration card | 11.4 |
| Timetable | 4.6 |
| Ticket | 4.5 |
| Specialized calls | 4.5 |
| empty | 3.6 |
| New registration | 3.4 |
| Price info | 3.0 |

Table 2: Repartion of the 10 top calltypes in the DECODA corpus

As we can see, 4 call-types represent 67% of the dialogs. The top call-type is *Info Traffic* which contains requests about delays, roadworks and strikes in the transportation network. This is an interesting category as a large variety of customers behaviors occur, like for example stress and angriness sentiments from people being late because of a problem on the transportation network.

## 6. Syntactic annotation

Five levels of linguistic annotation have been performed on the transcriptions of the DECODA corpus:

1. Disfluencies
2. Part-of-Speech (POS) tags
3. Chunks
4. Named Entities
5. Syntactic dependencies

The first level (*disfluencies*) corresponds to repetitions (e.g. *le le*), discourse markers (e.g. *euh*, *bien*) and false starts (e.g. *bonj-*). These annotations have been manually performed on the corpus thanks to a WEB-interface allowing to write regular expressions to recognize some word patterns corresponding to a kind of disfluencies. For each regular expression written, the interface displays all the matches found in the corpus in order to check the relevance of the expression. Once validated, all the occurrences matching the expression are tagged with the chosen label. Table 3 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.
The Named Entity annotation level has been performed manually with the same WEB-interface. In our corpus,

| Duration (min.) | $<= 1$ | $1-2$ | $2-3$ | $3-4$ | $4-5$ | $5-6$ | $6-7$ | $7-8$ | $8-9$ | $9-10$ | $> 10$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # of dialogs | 597 | 367 | 230 | 139 | 68 | 43 | 27 | 13 | 10 | 6 | 14 |

Table 1: Repartition of the duration

| **disfluency type** | # occ. | % of turns | 10 most frequent forms |
|---|---|---|---|
| *discourse markers* | 39125 | 28.2% | [euh] [hein] [ah] [ben] [voilà] [bon] [hm] [bah] [hm hm] [écoutez] |
| *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] [ça ça] |
| *false starts* | 1913 | 1.1% | [s-] [p-] [l-] [m-] [d-] [v-] [c-] [t-] [b-] [n-] |

Table 3: Distribution of the disfluencies manually annotated on the DECODA corpus

most of the named entities correspond to location expressions such as addresses, bus stops and metro stations.
The POS and chunk levels of annotation have been first automatically produced by the NLP tool MACAON (Nasr et al., 2011). At the same time a fully manual annotation process is performed on a subset of the corpus, called the *GOLD corpus*. An iterative process is then applied, consisting in manually correcting errors found in the automatic annotations, retraining the linguistic models of the NLP tools on this corrected corpus, then checking the quality of the adapted models on the fully manual annotations of the *GOLD corpus*. This process iterates until a certain error rate is reached (Bazillon et al., 2012). Table 4 shows the POS error rate before and after this adaptation process on the *GOLD corpus*. As we can see the error rate is reduced by almost 60%.

| POS error rate | Baseline | After adaptation |
|---|---|---|
| DECODA GOLD corpus | 21.0% | 8.5% |

Table 4: POS error rate before and after the iterative adaptation process performed on the DECODA training corpus

The same kind of semi-supervised method has been used to obtain the syntactic dependency annotations. These kinds of annotation are very useful in order to train new approaches to parsing based on dependency structures and discriminative machine learning techniques (Nivre, 2003; McDonald et al., 2005). These approaches are much easier to adapt to speech than context-free grammar approaches since they need less training data and the annotation with syntactic dependencies of speech transcripts is simpler than with syntactic constituents. The annotation process of the DECODA corpus is described in details in (Bazillon et al., 2012). A first evaluation of a statistical dependency parser trained on this annotated corpus and applied to the *GOLD corpus* has obtained encouraging results.
An example of a full annotated segment is displayed in table 5. The corpus can be found either on this semicolon format or in an XML format, similar to the one used in MACAON (Nasr et al., 2011).

## 7. Conclusion

We have described in this paper the acquisition, anonymization, transcription and annotation process of the DECODA corpus collected through the RATP call-center in Paris. The first experiments we carried out have controlled the quality of the annotations produced (Bazillon et al., 2012) and the usefulness of the corpus for spoken language understanding tasks (Maza et al., 2011). We are going now to fully exploit this annotated corpus in order to propose advanced speech analytics systems taking into account all the levels of annotation produced.

## 8. Acknowledgements

This work is supported by the French agency ANR, Project DECODA, contract no 2009-CORD-005-01, and the French business clusters Cap Digital and SCS. For more information about the DECODA project, please visit the project home-page, http://decoda.univ-avignon.fr

## 9. References

J.F. Allen. 1994. The trains project: A case study in building a conversational planning agent. Technical report, DTIC Document.

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

Thierry Bazillon, Melanie Deplano, Frederic Bechet, Alexis Nasr, and Benoit Favre. 2012. Syntactic annotation of spontaneous speech: application to call-center conversation data. In *Proceedings of LREC*, Istambul.

S. Galliano, G. Gravier, and L. Chaubard. 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

M. Garnier-Rizet, G. Adda, F. Cailliau, J.L. Gauvain, S. Guillemin-Lanne, L. Lamel, S. Vanni, and C. Waast-Richard. 2008. Callsurf-automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. In *Proceedings of LREC*.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Benjamin Maza, Marc El-Beze, Georges Linares, and Renato De Mori. 2011. On the use of linguistic features

| *id* | *word* | *disfluencies* | *NE* | *POS* | *chunk* | *dep. link* | *dep. type* |
|---|---|---|---|---|---|---|---|
| 1 | oui | - | - | adv | B_GADV | 0 | ROOT |
| 2 | je | B_REP | - | cln | - | - | - |
| 3 | je | I_REP | - | cln | - | - | - |
| 4 | je | - | - | cln | B_GN | 5 | SUJ |
| 5 | sais | - | - | v | B_GVfn | 0 | ROOT |
| 6 | pas | - | - | advneg | I_GVfn | 5 | MOD |
| 7 | euh | B_DISM | - | pres | - | - | - |
| 8 | je | - | - | cln | B_GN | 14 | SUJ |
| 9 | ne | - | - | clneg | B_GVfn | 12 | MOD |
| 10 | m’ | - | - | clr | I_GVfn | 12 | AFF |
| 11 | en | - | - | clo | I_GVfn | 14 | DE_OBJ |
| 12 | étais | - | - | v | I_GVfn | 14 | AUX |
| 13 | pas | - | - | advneg | I_GVfn | 12 | MOD |
| 14 | inquiétée | - | - | vppart | I_GVfn | 0 | ROOT |
| 15 | madame | - | - | nc | B_GN | 0 | root |
| 16 | c’ | - | - | cln | B_GN | 17 | SUJ_IMP |
| 17 | est | - | - | v | B_GVf | 0 | ROOT |
| 18 | École | - | B_LOC | np | B_GN | 17 | ATS |
| 19 | Véterinaire | - | I_LOC | np | - | - | - |
| 20 | ou | - | - | coo | B_coo | 18 | COORD |
| 21 | Rungis | - | B_LOC | np | B_GN | 20 | DEP_COORD |

Table 5: Example of annotated segment in the DECODA corpus

in an automatic system for speech analytics of telephone conversations. In *Proceedings of Interspeech 2011*, Florence, Italy.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online Large-Margin Training of Dependency Parsers. In *Association for Computational Linguistics (ACL)*.

Alexis Nasr, Frederic Bechet, Jean-Francois Rey, and Joseph Le Roux. 2011. Macaon:a linguistic tool suite for processing word lattices. In *The 49th Annual Meeting of the Association for Computational Linguistics: demonstration session*.

J. Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*.

S. Renals, T. Hain, and H. Bourlard. 2007. Recognition and interpretation of meetings: The ami and amida projects.

A.J. Stent. 2000. The monroe corpus. *TR728 and TN99-2, Dept. of Computer Science, University of Rochester*.

Gokhan Tur and Dilek Hakkani-Tür. 2011. Human/human conversation understanding. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, Gokhan Tur and Renato De Mori ed. Wiley*.