Annotation and Classification of French Feedback Communicative Functions

Laurent Prevot, Jan Gorish, Sankar Mukherjee

Annotation and Classification of French Feedback Communicative Functions

Laurent Prévot
Aix-Marseille Université
Laboratoire Parole et Langage
Aix-en-Provence, France
laurent.prevot@lpl-aix.fr

Jan Gorisch
Institut für Deutsche Sprache
Mannheim, Germany
gorisch@ids-mannheim.de

Sankar Mukherjee
Istituto Italiano di Tecnologia
Genova, Italy
sankar1535@gmail.com

Abstract

Feedback utterances are among the most frequent in dialogue. Feedback is also a crucial aspect of all linguistic theories that take social interaction involving language into account. However, determining communicative functions is a notoriously difficult task both for human interpreters and systems. It involves an interpretative process that integrates various sources of information. Existing work on communicative function classification comes from either dialogue act tagging where it is generally coarse grained concerning the feedback phenomena or it is token-based and does not address the variety of forms that feedback utterances can take. This paper introduces an annotation framework, the dataset and the related annotation campaign (involving 7 raters to annotate nearly 6000 utterances). We present its evaluation not merely in terms of inter-rater agreement but also in terms of usability of the resulting reference dataset both from a linguistic research perspective and from a more applicative viewpoint.

1 Introduction

Positive feedback tokens (*yeah, yes, mhm ...*) are the most frequent tokens in spontaneous speech. They play a crucial role in managing the common ground of a conversation. Several studies have attempted to provide a detailed quantitative analysis of these tokens in particular by looking at the form-function relationship (Allwood et al., 2007; Petukhova and Bunt, 2009; Gravano et al., 2012; Neiberg et al., 2013). About form, they looked at lexical choice, phonology and prosody. About communicative function, they considered in particular grounding, attitudes, turn-taking and dialogue structure management.

Despite the previous attempts to quantify that form-function relationship of feedback, we think that more work needs to be done on the conversational part of it. For example, Gravano et al. (2012) used automatic classification of *positive cue words*, however the underlying corpus consists of games, that are far off being "conversational" and therefore do not permit to draw any conclusions on how feedback is performed in conversational talk or talk-in-interaction. What concerns the selection of the feedback units, i.e. utterances, more work that clarifies what consists of feedback is also needed, as an approach that purely extracts specific lexical forms ("okay", "yeah", etc.) is not sufficient in order to account for feedback in general. Also, the question of what features to extract (acoustic, prosodic, contextual, etc.) is far from being answered. The aim of this paper is to shed some more light on these issues by taking data from real conversations, annotating communicative functions, extracting various features and using them in experiments to classify the communicative functions.

The study reported in this paper takes place in a project (Prévot and Bertrand, 2012) that aims to use, among other methodologies, quantitative clues to decipher the form-function relationship within feedback utterances. More precisely, we are interested in the creation of (large) datasets composed of feedback utterances annotated with communicative

29th Pacific Asia Conference on Language, Information and Computation pages 302 - 310
Shanghai, China, October 30 - November 1, 2015

functions. From these datasets, we conduce quantitative (statistical) linguistics tests as well as machine learning classification experiments.

After presenting feedback phenomena and reviewing the relevant literature (Section 2), we introduce our dataset (Section 3), annotation framework and annotation campaign (Section 4). After discussing the evaluation of the campaign (Section 5), we turn to the feature extraction (Section 6) and our first classification experiments (Section 7).

2 Feedback utterances

Definition and illustration Concerning the definition of the term *feedback utterance*, we follow Bunt (1994, p.27): *"Feedback is the phenomenon that a dialogue participant provides information about his processing of the partner's previous utterances. This includes information about perceptual processing (hearing, reading), about interpretation (direct or indirect), about evaluation (agreement, disbelief, surprise,...) and about dispatch (fulfillment of a request, carrying out a command, ...)."*

As a working definition of our class *feedback*, we could have followed Gravano et al. (2012), who selected their tokens according to the individual word transcriptions. Alternatively, Neiberg et al. (2013) performed an acoustic automatic detection of potential feedback turns, followed by a manual check and selection. But given our objective, we preferred to use perhaps more complex units that are closer to *feedback utterances*. We consider that feedback functions are expressed overwhelmingly through short utterances or fragments (Ginzburg, 2012) or in the beginning of potentially longer contributions. We therefore automatically extracted candidate feedback utterances of these two kinds. Utterances are however already sophisticated objects that would require a specific segmentation campaign. We rely on a rougher unit: the Inter-Pausal Unit (IPU). IPUs are stretches of talk situated between silent pauses of a given duration, here 200 milliseconds. An example of an *isolated feedback IPU* is illustrated in Figure 1a. In addition to isolated items, we added sequences of feedback-related lexical items situated at the very beginning of an IPU (see section 3 for more details and Figure 1b for an example).

Related work The study of feedback is generally associated with the study of *back-channels* (Yngve, 1970), the utterances that are not produced on the *main* communication channel in a way not to interfere with the flow of the main speaker. In the seminal work by Schegloff (1982), back-channels have been divided into *continuers* and *assessments*. While *continuers* are employed to make a prior speaker continue with an ongoing activity (e.g. the telling of a story), *assessments* are employed to evaluate the prior speaker's utterance.

A formal model for feedback items was proposed by Allwood et al. (1992). It includes four dimensions for analysing feedback: (i) Type of reaction to preceding communicative act; (ii) Communicative status; (iii) Context sensitivity to preceding communicative act; (iv) Evocative function. The first dimension roughly corresponds to the functions on the grounding scale as introduced by Clark (1996): *(contact / perception / understanding / attitudinal reaction)*. The second dimension corresponds to the way the feedback is provided *(indicated / displayed / signalled)*. The third dimension, *Context sensitivity*, is divided into three aspects of the previous utterance: mood *(statement / question / request / offer)*, polarity and information status of the preceding utterance in relation to the person who gives feedback. The fourth dimension, *Evocative function*, is much less developed but relates to what the feedback requires / evokes in the next step of the conversation.

Grounded in this previous work but more concerned with annotation constraints, especially in the context of multi-modal annotations, Allwood et al. (2007) use a much simpler framework that is associated with the annotation of turn management and discourse sequencing. The feedback analysis is split into three dimensions: (i) basic *(contact, perception, understanding)*; (ii) acceptance; (iii) emotion / attitudes that do not receive an exhaustive list of values but include *happy, surprised, disgusted, certain,* etc.

Muller and Prévot (2003; Muller and Prévot (2009) have focused on more contextual aspects of feedback: function of the feedback target and feedback *scope*.The work relies on a full annotation of communicative functions for an entire corpus. The annotations of feedback-related functions and of feedback scope are reported to be reliable. However, the dataset analysed is small. and the guide-

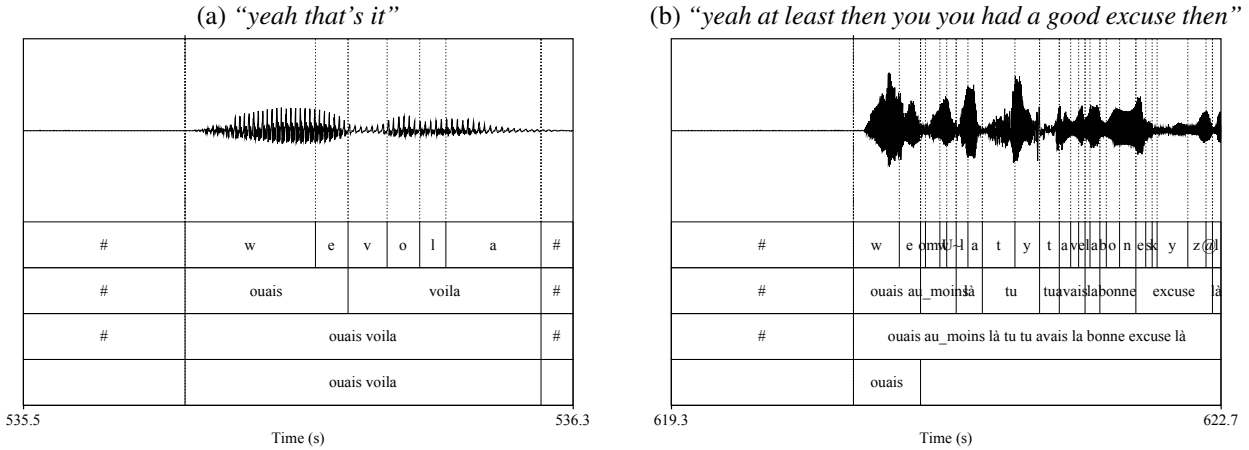

Figure 1: Approximation of feedback items. Isolated feedback (left); Initial feedback item sequence (right).

lines are genre-specific (route instruction dialogues) while we intend here a generalisable approach.

More recent frameworks include work by Gravano et al. (2012) who propose a flat typology of affirmative cue word functions. This typology mixes *grounding* functions with *discourse sequencing* and other unrelated functions. It includes for example *Agreement, Backchannel*, discourse segment *Cue-Beginning* and *Cue-Ending* but also a function called *Literal modifier*. The reason for such a broad annotation is that every instance of an affirmative cue word is extracted following a completely form-driven simple rule. Such an approach allows to create high-performance classifiers for specific token types but hardly relates to what is known about feedback utterances in general. Their dataset is therefore much more homogeneous than ours in terms of lexical forms but more diverse in terms of position since we did not extract feedback related tokens occurring for example in a medial or final position of an IPU. A token-based approach forbids to give justice to complex feedback items such as reduplicated positive cue words, and obvious combinations such as *ah ouais (=oh yeah)*, *ok d'accord (=okeydoke)*. Their strategy is simply to annotate the first token and ignore the other. Our strategy is to capture potential compositional or constructional phenomena within feedback utterances. Moreover, even within a word-based approach, it is debatable to use space from a transcription to delineate the units of analysis. Some of these sequences could already be lexicalized within the actual spoken system. A final point concerns reduplicated words. It is often difficult to determine whether an item is *mh*, *mh mh* or *mh + mh*. While treating IPUs does not completely resolve this issue, it is more precise than only annotating the first token.

The form-driven approach by Neiberg et al. (2013) also combines automatic data selection with lexical and acoustic cues. As for the function annotation, they identify five scalar attributes related to feedback: *non-understanding – understanding*, *disagreement – agreement*, *uninterested – interested*, *expectation – surprise*, *uncertainty – certainty*. This scalar approach is appealing because many of these values seem to have indeed a scalar nature. We adopt this two tier approach to characterize communicative functions. We first identify a BASE function and when this function is taken to hold some deeper evaluative content such as agreement or the expression of some attitude, a second level EVALUATION is informed. Moreover, our approach considers that a crucial aspect of feedback utterances is their contextual adequacy and dependence. To test this hypothesis, we included an annotation for the *previous utterance* in our annotation framework (more detail in section 4).

3 Dataset

All data used in this study come from corpora including conversational interactions (CID) and task oriented dialogues (Aix-MapTask). Both corpora include native French speaking participants.

CID Conversation Interaction Data (CID) are audio and video recordings of participants having a

conversational interaction with the mere instruction of talking about strange things to kick-off the conversation (Bertrand et al., 2008; Blache et al., 2010). The corpus contains 8 hours of audio recordings[1].

Aix-MapTask Remote The Aix-MapTask (Bard et al., 2013; Gorisch et al., 2014) is a reproduction of the original HCRC MapTask protocol (Anderson et al., 1991) in the French language. It involves 4 pairs of participants with 8 maps per pair and turning roles of giver and follower. The remote condition (MTR) contains audio recordings that sum up to 2h30 with an average of 6 min. 52 sec. per map[2].

Data extraction Our objective is to obtain a dataset that covers as completely as possible feedback utterances. We exploited our rather precise transcriptions (aligned with the signal at the phone level with the tool SPPAS (Bigi, 2012)) that include laughter, truncated words, filled pauses and other speech events. We started from the observation that the majority of feedback utterances are IPUs composed of only a few tokens. We first identified the small set of most frequent lexical items composing feedback utterances by building the lexical tokens distribution for IPUs made of three tokens or less. The 10 most frequent lexical forms are : *ouais / yeah (2781), mh (2321), d'accord / agree-right (1082), laughter (920), oui / yes (888), euh / uh (669), ok (632), ah (433), voil / that's it-right (360)*. The next ones are *et / and (360), non / no (319), tu / you (287), alors / then (151), bon / well (150)* and then follow a series of other pronouns and determiners with frequency dropping quickly. We excluded *tu, et* and *alors* as we considered their presence in these short isolated IPUs were not related to feedback. We then selected all isolated utterances in which the remaining items were represented and treated now each IPU as an instance of our dataset. As mentioned in the introduction, we also extracted feedback related token sequences situated at the beginning of IPUs. This yielded us a total of more than 7000 candidate feedback utterances.

In terms of coverage, given our heuristics for selecting feedback utterances, we miss most of the short utterances that are uniquely made of repetitions or reformulations (not including feedback related tokens). Our recall of feedback utterances is therefore not perfect. However, our final goal is to combine lexical items with prosodic and acoustic features. Therefore, our heuristics focus on these tokens. About lexical items, our coverage is excellent. Although there are some extra items that are not in our list, such as (*vachement* (a slang version of *'a lot'*) or *putain* (a swear word that is used as a discourse marker in rather colloquial French), these items remain relatively rare and moreover, they tend to co-occur with the items of our list. Therefore, most of their instances are part of our dataset in the *complex* category.The plus sign in *ouais+* and *mh+* stands for sequences of 2 or more *ouais* or *mh*. The token *complex* corresponds to all other short utterances extracted that did not correspond to any item from the list, e.g. *ah ouais d'accord*, *ah ben @ ouais*,...). For more details on the dataset, see Prévot et al. (2015).

4 Annotation of communicative functions

We ended up with 5473[3] cross-annotated candidate utterances from CID and MTR corpora. Although the initial annotation schema was fairly elaborate, not all the dimensions annotated yielded satisfactory inter-annotator agreement[4]. In this paper we focus on two articulated dimensions: the BASE, which is the base function of the feedback utterance *(contact, acknowledgment, evaluation-base, answer, elicit, other)*, and EVALUATION, which was informed when the *evaluation-base* value was selected as the BASE function (evaluations could be: *approval, unexpected, amused, confirmation*). The details for these two dimensions are provided in Table 1. We also asked annotators to rate what the function of the previous utterance of the interlocutor was *(assertion, question, feedback, try, request, incomplete, uninterpretable)*. Although circular, this last annotation was gathered to tell us how useful this kind of contextual information was for our task.

[1]The CID corpus is available online for research: `http://www.sldr.org/sldr000027/en/`.

[2]The description of MTR is available online: `http://www.sldr.org/sldr000732`.

[3]The difference from the original data points comes from missing annotation values and technical problems on some files.

[4]Dimensions related to feedback scope and the structure of the interaction were not consistently annotated by our naive annotators and will not be discussed here further.

To conduct the annotation campaign, seven undergraduate and master students were recruited. The campaign was realized on a duration of 2 months for most annotators. Annotating one feedback instance took on average 1 minute. We made sure that every instance received 3 concurrent annotations in order to be able to set-up a voting procedure for building the final dataset.

5 Evaluation

5.1 Inter-rater agreement

Concerning BASE value annotations, the average κ value for the best pair of raters for all the sub-datasets with enough instances to compute this value was around 0.6 for both corpora: MTR (min: 0.45; max: 0.96) and CID (min: 0.4; max: 0.85). Multi-κ yielded low values suggesting some raters were not following correctly the instructions (which was confirmed by closer data inspection). However, we should highlight that the task was not easy. There is a lot of ambiguity in these utterances and lexical items are only part of the story. For example, the most frequent token *ouais* could in principle be used to reach any of the communicative functions targeted. Even after close inspection by the team of experts, some cases are extremely hard to categorize. It is not even sure that the dialogue participants fully determined their meaning as several functions could be accommodated in a specific context.

While best pair's κ seems to be a very favorable evaluation measure, most of our samples received only 3 concurrent annotations. Moreover, aside a couple of exceptions, always the same two raters are excluded. As a result, what we call "best-pair kappa" is actually simply the removal of the annotation of the worse two raters from the dataset, which is a relatively standard practice. There could be a reason for these raters to behave differently from others. Because of timing issues, one annotator could not follow the training sessions with the others and had to catch up later. The other annotator did the training with the others but had to wait almost 2 months before performing the annotation.

Concerning EVALUATION values annotations, it is more complex to compute reliably an agreement to the sporadic nature of the annotation (evaluation values are only provided if the rater used this category in the BASE function). Since the set of raters that annotate a given sample varies, in most cases of MTR the number of instances annotated by a given set of raters is too small to compute reliably agreement. On the CID corpus, which has much larger samples, κ-measures of EVALUATION can be computed but exhibit huge variations with a low average of 0.3. This is indeed a difficult task since raters have to agree first on the BASE value and then on the value of the EVALUATION category. But, as we will see later, our voting procedure over cross-annotated datasets still yielded an interesting annotated dataset.

5.2 Quality of the reference dataset

In order to better understand the choice we have about data use and selection, we evaluated several datasets built according to different confidence thresholds.

For the `base` level, we started with the whole dataset and then built sub-datasets made of the same data but restricted to a certain threshold based on the number of raters that employ this category (threshold values: $\frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1$). More precisely, we computed a confidence score for each annotated instance. We then use these different datasets to perform two related tasks: classifying the functions of the whole dataset (using a `None` category for instances that did not reach the threshold) and classifying the functions within a dataset restricted to the instances that received an annotated category of a given threshold. In the case of the classification of `eval`, we first restricted the instances to the ones that received the `evaluation` value as value for the `base` category.

These datasets are ranging from noisy datasets (low threshold, full coverage) to cleaner ones but without full coverage. They correspond to two main objectives of an empirical study: (i) more linguistic / foundational studies would probably prefer to avoid some of the noise in order to establish more precise models to match their linguistic hypotheses, (ii) natural language engineering has no other choice than to work with the full dataset.

Composition of the dataset As for the BASE category distributions, the CID dataset is made of bit more than 40% of *ack* and *eval*, almost 15% of *others* and only 2% of *answer* ($\sim$2%). The MTR

Table 1: Annotated categories of communicative functions and their paraphrases.

Base Function	Paraphrase
contact	I am still here listening.
acknowledgment	I have heard / recorded what you said but nothing more.
evaluation-base	I express something more than mere acknowledgement (approval, expression of an attitude,...).
answer	I answer to your question / request.
elicit	Please, provide some feedback.
other	This item is not related to feedback.
Evaluation	
approval	I approve vs. disapprove / agree vs. disagree with what you said.
expectation	I expected vs. did not expect what you said.
amusement	I am amused vs. annoyed by what you said.
confirmation / doubt	I confirm what you said vs. I still doubt about what you said.

dataset, has a similar amount of *ack*, about 20% of *eval* and *answer*, 10% of *others* and 5% of the *elicit* category (that was basically absent from CID).

As for the EVALUATION category, CID is mostly made of *approbation* (46%) and *amused* (38%), then *confirmation* (8%) and *unexpected* (6%) while MTR has over 60% *confirmation*, only 13% *amused* feedback and 17% *approbation*.

6 Feature extraction

For our experiments, we focused on speech data and our dimensions include properties of items themselves: lexical content (LEX), acoustics (ACO); and properties of their context: apparition, timing and position (POS). We also use three more dimensions: contextual information extracted automatically (CTX-AUT), supplied manually by our annotators (MAN)[5] and meta-data (META). Some details about these features are provided here:

LEX transcription string + presence vs. absence of frequent lexical markers (16 features before binarization)

ACO pitch (min/max/stdev/height/span/steepness/slope/NaN-ratio[6]), intensity (quartiles Q1, Q2, Q3), avg aperiodicity, formants (F1, F2, F3) and duration (16 features)

POS speech environment in terms of speech/pause duration before/after the item for both the speaker and the interlocutor; including overlap information (10 features)

CTX-AUT first/last tokens and bigrams of previous utterance and interlocutor previous utterance (18 features before binarization)

MAN function of the interlocutor's previous utterance, a circular information providing a kind of topline (1 feature)

META Corpus, Speaker, Session, Role (4 features)

For the classification experiments, all textual and nominal features have been binarized. All numeric features have been attributed min max threshold values and then normalized within these thresholds.

7 Classification experiments

7.1 Classification of the Base function

Our first task was to classify the BASE function. The dataset we used most intensively was the one in which we retain only the base functions proposed by at least $\frac{2}{3}$ of the annotators[7]. This is computationally difficult because none of the levels involved is enough to perform this task. As we will see, only a combination of dimensions allows us to reach interesting classification scores.

We first compared the impact of the classifier choice on the dataset. We set-up a *baseline* consisting of the majority class for each frequent lexical item. For example, all single *'mh'* are classified as *ack* because the majority of them are annotated

[5]This corresponds to the annotation of the previous utterance of the interlocutor within this list of labels: *assert, question, feedback, try (confirmation request), unintelligible, incomplete*.

[6]The ratio of unvoiced parts (NaN = Not a Number) and voiced parts of the F0 contour.

[7]The majority of the instances have been cross-annotated by three annotators.

with this function. Then, we took our full set of features (LEX+ACO+CTX-AUT) and ran many classification experiments with various estimators (Naive Bayes, Decision Tree, SVM and Ensemble classifiers - Ada Boost and Random Forest) that are part of the SCI-KIT LEARN Python library (Pedregosa et al., 2011) and several parameter sets. The *Random Forest* method performed best. One explanation for this can be that *Tree-based* classifiers have no problem handling different categories of feature sets and are not confused by useless features. A nice consequence is that it becomes easy to trace which features contribute the most to the classification. This point is indeed crucial for us who intend to clarify the combination of the different linguistic domains involved. For this reason, and because all the experiments (varying various parameters) always ended up with an advantage for *Random Forest*, we used this classifier (with 50 estimators and minimum size of leaves of 10 instances) for the rest of the study in this paper.

We also checked the learning curve with this classifier and we have seen that it brings already interesting results with only one third of the dataset.

Our second task was to vary the sets of features used. We wanted however to refine this experiment by looking separately at each corpus. In figures 2a and 2b, the feature sets tested are the BASELINE described above, only LEXical, ACOustic or POSitional featues, the combination of the three (LPA), ALL automatically extracted features and ALL + MANually annotated previous utterance function. All experiments have been conducted with 10-fold cross-validation providing us the standard deviations allowing significance comparison as can be seen with the error bars in the figures (typically these deviations range between 1% to 2% for BASE and from 3% to 4% with some deviations going up to 10% for EVALUATION).

The results illustrated in Figure 2a, once we know what our features are good at, can be largely explained by the distribution of the categories across the corpora. There are therefore not many *answer* instances ($\sim$2%) in this corpus, a category that is not well caught by our features yet. But LEX, POS and ACO are good to separate precisely *ack, eval* and *other*. The MTR dataset has much more *answers*, which explains the jump in f-measure if we add the manual annotation of the interlocutor's previous utterance (MAN). We simply did not manage to catch this contextual information with our features yet and this has a much stronger impact on MTR than on CID.

7.2 Classification of the evaluation function

We ran the same experiments for the EVALUATION category as presented in Figure 2b. The features used by the classifier are different. Within *evaluation cases*, POS becomes less informative while LEX and ACO retain their predictive power. Corpora differences explain the results. CID has much more AMUSED feedback that are well caught by lexical features. MTR has more *confirmations* that can be signalled by a specific lexical item *(voilà)* but that is also strongly dependent on which participant is considered to be competent about the current question under discussion.

7.3 Individual features contribution

A close inspection of some of the trees composing the Random Forest allows us to understand some of the rules used by the classifier across linguistic domains. Here are some of the most intuitive yet interesting rules:

- if acoustic values `pitch span` and `F1` increase, attitudinal (EVAL) values are more likely than mere `acknowledgment` (`ack`) and this on various situations.
- `aperiodicity` seems to have been used to catch `amused` values that would not be associated with a *laughter* in the transcription.
- the presence of *mh* and *laughter* in the transcription is a very good predictor of `ack` (in the BASE task) and `amused` (in the EVAL task).
- with an increase of `opb` *(silence duration of the interlocutor channel before the classified utterance)*, `other` than feedback is more likely.

7.4 Impact of the dataset's quality

We checked what happens when one varies the threshold used for proposing a label on the instances and the different results if one uses the whole dataset or only the instances that received a label at a given confidence score (lower score means more labelled

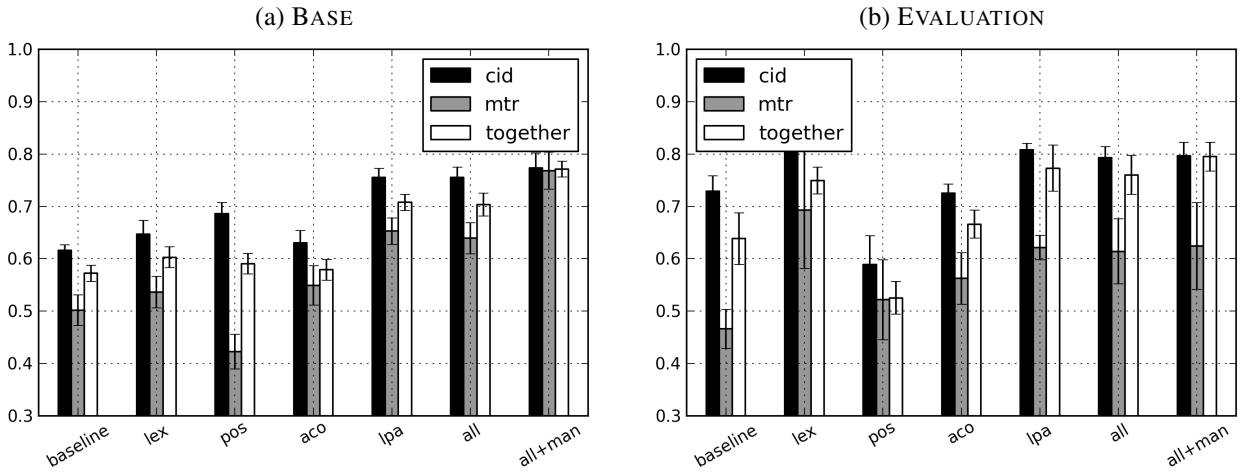

Figure 2: Classification results: f-measure (y-axis) per feature set (x-axis).

data but more noise, higher score means less noise but also less labelled data).

Unsurprisingly, the accuracy on the filtered dataset increases with the employed threshold. We note however that on the `eval` category that has a high score even with a low threshold, the accuracy gain is not fantastic.

About the non-filtered dataset, in the case of `eval` and at threshold $> \frac{3}{4}$, the classifier is focusing on the *None* category to reach a high score (since this category becomes dominant). As for `base`, we note that the changes in threshold have a complex effect on the accuracy. Accuracy is stable for the $\frac{1}{3}$ to $\frac{1}{2}$ shift (reliability on instances is better and coverage still very high), then decrease (with significant decrease of coverage). The shift from $\frac{3}{4}$ to 1 shows a slight increase in accuracy (due to a better recognition of the *None* category).

8 Conclusions

In this paper, the focus was on communicative functions, as they are performed by conversational participants. For everybody who is not directly engaged in the conversation, it is difficult to distinctly categorise such behaviour. In fact, our classification results are getting close to the error rate of the naive raters themselves. On the one hand, we note that some basic important distinctions (in particular the *ack* vs. *eval* divide that can be related to Bavelas et al. (2000) generic vs. specific listener responses) can be fairly efficiently caught by automatic means. This is done thanks to the importance of lexical, positional and acoustic features in determining these differences. On the other hand, our system has to improve as soon as contextual information becomes more important like for identifying *answer* or *confirmation*.

This methodology is almost completely data-driven and can be therefore applied easily to other languages, given that the corresponding annotation campaign is realized. More precisely, the creation of our feature sets and extractions can be fully automated. The main processing step is the forced-alignment. Most of the lexical features can be derived by extracting token frequency from short IPUs (here 3 tokens or less). The real bottleneck is the annotation of communicative functions. But now that the general patterns are known, it becomes possible to design more efficient campaigns.

Acknowledgements

This work is supported by French ANR (ANR-12-JCJC-JSH2-006-01). The second author also benefits from a mobility from Erasmus Mundus Action 2 program MULTI of the European Union (GRANT 2010-5094-7). We would like to thank Roxane Bertrand for the help on the selection of feedback utterances, Brigitte Bigi for help with the automatic processing of the transcriptions and Emilien Gorene for help with recordings and annotation campaigns. Finally, we would like to thank all recruited students who performed the annotations.

References

J. Allwood, J. Nivre, and E. Ahlsen. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3):273–287.

A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

E. G. Bard, C. Astésano, M. D'Imperio, A. Turk, N. Nguyen, L. Prévot, and B. Bigi. 2013. Aix MapTask: A new French resource for prosodic and discourse studies. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, Aix-en-Provence, France.

J.B. Bavelas, L. Coates, and T. Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952.

R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy. 2008. Le CID-Corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, 49(3):1–30.

B. Bigi. 2012. SPPAS: a tool for the phonetic segmentation of speech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1748–1755, ISBN 978–2–9517408–7–7, Istanbul, Turkey.

P. Blache, R. Bertrand, B. Bigi, E. Bruno, E. Cela, R. Espesser, G. Ferré, M. Guardiola, D. Hirst, E. Muriasco, J.-C. Martin, C. Meunier, M.-A. Morel, I. Nesterenko, P. Nocera, B. Palaud, L. Prévot, B. Priego-Valverde, J. Seinturier, N. Tan, M. Tellier, and S. Rauzy. 2010. Multimodal annotation of conversational data. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV '10)*, Uppsala, Sweden.

H. Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

H.H. Clark. 1996. *Using Language*. Cambridge: Cambridge University Press.

J. Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

J. Gorisch, C. Astésano, E. Bard, B. Bigi, and L. Prévot. 2014. Aix Map Task corpus: The French multimodal corpus of task-oriented dialogue. In *Proceedings of The Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

A. Gravano, J. Hirschberg, and Š. Beňuš. 2012. Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, 38(1):1–39.

P. Muller and L. Prévot. 2003. An empirical study of acknowledgement structures. In *Proceedings of 7th workshop on semantics and pragmatics of dialogue (DiaBruck)*, Saarbrücken, Germany.

P. Muller and L. Prévot. 2009. Grounding information in route explanation dialogues. In *Spatial Language and Dialogue*. Oxford University Press.

D. Neiberg, G. Salvi, and J. Gustafson. 2013. Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication*, 55:451–469.

F. Pedregosa, G. Varoquaux, A Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

V. Petukhova and H. Bunt. 2009. The independence of dimensions in multidimensional dialogue act annotation. 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*, pages 197–200, Boulder, Colorado, USA.

L. Prévot and R. Bertrand. 2012. Cofee-toward a multidimensional analysis of conversational feedback, the case of french language. In *Proceedings of the Workshop on Feedback Behaviors*. (poster).

L. Prévot, J. Gorisch, R. Bertrand, E. Gorene, and B. Bigi. 2015. A SIP of CoFee: A Sample of Interesting Productions of Conversational Feedback. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 149–153.

E. A. Schegloff. 1982. Discourse as an interactional achievement: Some use of uh-huh and other things that come between sentences. *Georgetown University Round Table on Languages and Linguistics, Analyzing discourse: Text and talk*, pages 71–93.

V. H. Yngve. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–578.