Discourse and Prosody in spoken French: Why, what and how should one count? A comparative statistical perspective

Julie Beliao, Anne Lacheret, Sylvain Kahane

Discourse and Prosody in spoken French: Why, what and how should one count? A comparative statistical perspective

Julie Beliao, Anne Lacheret, Sylvain Kahane
MoDyCo Lab - UMR7114
University of Paris Ouest, France
<julie@beliao.fr, anne@lacheret.com, sylvain@kahane.fr>

Résumé

Dans cet article, nous présentons un résumé de certains aspects méthodologiques et statistique de notre récent travail sur un corpus de français parlé. Pourquoi compter? Nous montrons comment une étude basée sur des données linguistiques de l'oral, peut avoir une visée technique, où l'objectif serait par exemple d'obtenir des erreurs de classification faibles, ou bien avoir une visée plutôt interprétative, où l'objectif serait d'étudier les relations entre niveaux linguistique afin d'en tirer des conclusions. Comment compter? En fonction de cette distinction, on montre que certaines méthodes statistiques sont parfois plus ou moins adéquates. Par exemple, les méthodes de classification automatique de machine learning s'avèrent souvent très efficaces, mais peuvent se révéler beaucoup moins satisfaisante lorsqu'il s'agit de comprendre et d'expliquer l'influence d'une variable linguistique sur une autre. De la même manière, nous montrons qu'une analyse statistique peut être utile pour évaluer et même réduire la tâche d'annotation, en permettant par exemple l'identification des variables redondantes qui pourraient alors être évitées. La discussion est basée sur de nombreux exemples tirés de notre récent travail sur le corpus Rhapsodie, qui est un corpus de français ordinaire, annoté en syntaxe, discours et en prosodie.

Keywords: *prosody, macrosyntax, SVM, decision tree, PCA*

1. Why count?

1.1. Engineering perspective: classification tasks

The purpose of machine learning is to build a general model from specific data in order to predict discourse behavior with new data. In discourse genre studies, machine learning seeks to group a set of items into several classes (communicational variables) according to their linguistic characteristics (descriptive variables, also called features). Classes may be unknown, as in the case of unsupervised classification, or known, as in supervised classification. In this study, we consider supervised learning only.

A machine learning method for classification operates as follows: in

a first step, a set of observations is provided for the purpose of *learning*. Each observation consists of some numerical features along with the corresponding value for the situational variable (e.g. +/- planned speech). The objective of the learning algorithm, which varies depending on the model considered, is to learn how to predict the situational variable from observation of the features only. In a second *testing* step, the value of the situational variable is not known, only the value of the features. The learned model can then be used to predict this situational variable. In order to assess the performance of a machine learning technique, part of the corpus is generally used to train the model, while the other part is used to evaluate it, by comparing the estimated situational variables to the true ones. Doing so permits not only to evaluate the model's ability to explain known and observed data, but also its ability to be applied on unknown data, that is to say its ability to generalize. It must also be emphasized that the aim of an engineering task is, above all, to achieve good performances, regardless of the features considered.

1.2. Functional perspective

The functional interpretation of features with respect to situational variables implies a thorough understanding of the relations between the features and the situational variables. In this case, the main objective is not to achieve high classification scores, but rather to explain how these features are organized and to characterize a kind of speech, for example. Classification performance using machine learning methods, in contrast, is poorly suited for understanding the structure of language. Still, although it may seem somewhat artificial, performing such automatic classification with manual linguistic annotations as an input is useful, because it can validate these annotations as relevant to characterize the situational variable under study, but not much more (Beliao, Lacheret, & Kahane, 2014).

Functional theories of grammar focus on the functions of language and on its elements as a key to understanding linguistic structures and processes. These theories suggest that since language is primarily a tool, it is reasonable to assume that its structures are better analyzed and understood with reference to the functions they perform (François, 1998, 2008, 2010). This means that functional theories of grammar tend to pay attention to how the language is actually used in a communicative context, which is not something that can be efficiently captured by classical machine learning methods for classification. Instead, a functional analyst may formulate distinct linguistic hypotheses and try to elicit one based on statistical tests performed over a corpus. While this methodology does not lead to automatic engineering methods to achieve a particular classification task, it does provide valuable insight

from a functional perspective (Beliao & Lacheret, 2013; Beliao, 2014).

2. Rhapsodie Model: corpus design and annotation schemes

In this section, we first present the data used in our experiments: the Rhapsodie treebank. We highlight the choices that were made when building this corpus, both for sampling and for the definition of the situational variables used for analysis. The linguistic features can be divided into primary and secondary data: broadly speaking, primary data were built automatically, while secondary data relied on manual annotations.

2.1. Sampling and metadata

Two features characterize the Rhapsodie project (Lacheret et al., 2014)[1]: (i) it aims to model the intonosyntactic interface (i.e. relative to the interaction between syntax and prosody) from various constructions, annotated according to prosodic and syntactic levels, that are sufficiently numerous to allow descriptive generalizations; (ii) it assumes that there is a close relationship between the typological characteristics of speech, i.e. spoken patterns defined on the basis of strictly formal criteria, and discourse genres, or rather the situational features that characterize them[2].

Accordingly, the guidelines for sampling the Rhapsodie corpus were as follows: (i) collect a set of sufficiently diversified samples in terms of types of text, (ii) have a sufficiently large panel of speakers to avoid individual idiosyncrasies, (iii) given the first two constraints and given the huge time cost implied by a robust annotation of macrosyntax and prosody, select short samples only (five minutes on average). All these points explain the relatively small size (3 hours and 34 000 tokens) of our corpus in comparison with automatically annotated or written corpora.

Since a corpus achieving this diversity and textual balance does not exist at the present time, data were first extracted from existing sources (Durand, Laks, & Lyche, 2009; Branca-Rosoff, Fleury, Lefeuvre, & Pires, 2009; Eshkol-Taravella et al., 2011; Avanzi, Simon, Goldman, & Auchlin, 2010)[3]. This first set of samples was then supplemented by other data types (multimedia, movie descriptions, itineraries, among others),

---

[1] The corpus has been produced as part of the ANR Rhapsodie 07 Corp-030-01, http://www.projet-rhapsodie.fr

[2] See also the concept of "register" in (Biber & Conrad, 2009), characterized by a frequent and recurring series of lexical-grammatical features in the texts of some variety and serve major communicative functions.

[3] Exhaustive description of the sources is available on the Rhapsodie website: http://www.projet-rhapsodie.fr/propriete-intellectuelle.html

collected for the Rhapsodie project to ensure the balance of the samples, which was fixed in advance.

The situational variables considered were the speaking type, the degree of discourse planning, interactivity, and discourse genre. See Table 1 for more details.

SPEAKING TYPE	Monologue, Dialogue
SPEECH PLANNING	Spontaneous, Semi-planned, Planned
INTERACTIVITY	Interactive, Semi-Interactive, Non-interactive
DISCOURSE GENRE	Argumentative, Procedural, Descriptive, Oratory

Table 1: *Situational variables for the Rhapsodie corpus*

2.2. Choice of features

The numerical features for each excerpt in the corpus include[4]: (i) primary data (for example, the number of tokens, the length of the excerpt, etc.) (ii) secondary data, i.e. the number of different annotated linguistic units, including macrosyntactic, microsyntactic (Benzitoun, Dister, Gerdes, Kahane, & Marlet, 2009; Benzitoun et al., 2010; Gerdes & Kahane, 2009; Pietrandrea, Kahane, Lacheret, & Sabio, 2014), prosodic units (Avanzi, Lacheret-Dujour, Obin, & Victorri, 2011; Lacheret, Obin, & Avanzi, 2010; Lacheret & Victorri, 2002), etc.

Four questions underlie the processing of the data: (i) how can one compare the discriminative power of each type of each manipulated variable, i.e. syntactic, prosodic and intonosyntactic? (ii) how can complementary or redundant features among these variables be detected? (iii) is it relevant to perform an intonosyntactic processing, i.e. a combined processing of prosodic and syntactic variables (for example a correlation between syntactic and prosodic units), instead of handling them separately? (iv) what is the contribution of secondary (annotated) data to these characterization and classification operations?

3. Results

3.1. Engineering tasks

In this study, we consider two automatic classification models: decision trees and Support Vector Machines. The aim of this task is to test machine learning methods on manual annotations. As was highlighted in section 1.2., this is more to validate the features that are discriminative of the phenomenon under study than to propose an engineering method to perform classification.

[4] The process is performed automatically with tools implemented in OOPS (Beliao & Liutkus, 2014).

3.1.1. Two supervised classification methods

A decision tree takes as input a feature vector and outputs the value of the situational variable of interest, e.g. speech planning. The construction is top-down: at the beginning, all the samples are pooled and the algorithm is then applied recursively, by splitting samples from each group as effectively as possible using a simple test over the features. In practice, a partitioning is considered good if it separates the samples into two groups, each with the same value for the explanatory variable (i.e., without noise), or at least with as small a variance as possible. The algorithm stops when all the samples from each subset belong to the same class. Decision trees have three qualities that are particularly interesting from a linguistic interpretation perspective: (i) the classification is very fast; (ii) the decision is made in a dichotomous manner, i.e. it is binary; (iii) as a result, decisions are easily interpretable.

Support vector machines (SVM) adopt a different approach to train classifiers on the data. They are mainly based on a geometrical interpretation of the classification task: if each sample is associated with a feature vector (i.e., x-y coordinates), training a classifier amounts to finding a way to best separate these points into two groups, while making as few errors as possible. An SVM conducts the task by essentially identifying the few samples that can cause problems because they are on the border between two groups, but generalizes this selection in an arbitrary dimension, i.e. when the number of features is large. This ability of SVMs to take only a small number of samples into account for the processing makes them particularly attractive when handling large amounts of data. Thus, unlike decision trees that seek the best criterion several times in a greedy fashion, SVMs seek the best global combination of features.

3.1.2. Classification performance

We applied the two classification methods described above with a *leave-one-out cross-validation* (LOOCV). This means that the model is trained using 56 samples of the corpus and tested on the latter, iteratively through all 57 samples. Generally, LOOCV can correctly estimate the error with a small bias, but with a higher variance than with other cross-validation approaches. We then calculated the success rate of both methods, using only the primary data, only the secondary data and finally both. The corresponding results are given in 2 for decision trees, and in Table 3 for SVM.

Table 2 shows that the performance of decision trees in terms of generalization are often worse (type of speech, planning) than those of SVM. One classical interpretation is that decision trees are more sensitive to the "curse of dimensionality": if the number of features is high,

Input data	Monologue vs dialogue *(2 values)*	Planning *(3 values)*	Interactivity *(3 values)*	Discourse genre *(4 values)*
Primary	*86.4*	*66.4*	*60.5*	39.3
Secondary	74.2	54.7	46.7	39.3
Both	77.5	56.4	*60.5*	*57.1*

Table 2: *Succes rates of correct categorization for* DECISION TREES *(in percentages) obtained for each class based on different data sets*

Input data	Monologue vs dialogue *(2 values)*	Planning *(3 values)*	Interactivity *(3 values)*	Discourse genre *(4 values)*
Primary	82.45	56.14	66.66	36.84
Secondary	77.19	*64.91*	*68.42*	64.91
Both	*84.21*	59.64	64.91	*66.66*

Table 3: *Succes rates of correct categorization for* SVM *(in percentages) obtained for each class based on different data sets*

the volume of the feature space increases dramatically and many samples are required to correctly train a model. Consequently decision tree algorithms suffer from a small corpus size if the number of features is high. In contrast, SVMs have a better generalization performance, even (and especially) when the number of features increases, even for a limited set of samples. This is because the corresponding learning algorithm determines a global optimum to the problem of classification, while the decision tree learning algorithm proceeds greedily (iteratively partitioning), assuming that short-term optimal solutions will yield the long-term one, which is not necessarily the case. Several properties of SVMs explain their good practical performances. First, automatic data normalization is performed during the learning phase. This automatically compensates for the dynamics observed between features. Then, the margin parameter roughly determines the support vectors needed to correctly learn the model. It is significant that in our tests, on 56 samples used for training, 51 were used as support vectors. Indeed, given the low density of the learning samples in the features, it is normal that almost all observations were kept, because the data are not redundant.

In most cases, the gain induced by the inclusion of the secondary data is noticeable, meaning that the manually annotated prosodic and syntactic features are indeed relevant to study the situational variables considered.

3.2. Functional analysis tasks

Why is language structured the way it is? Is it because it reflects constraints on language use? But then, how is this "reflection" operated? Such questions arise when studying at the syntactic and prosodic level.

3.2.1. Methodology

For this approach, one habitually chooses two features. Doing so permits to first observe how they are related and also to test for an a priori hypothesis on these variables. In this section, we present three experiments conducted on functional analysis grounds. The first one consisted in a comparative analysis of prosodic disfluencies (Beliao & Lacheret, 2013) — annotated according to the protocol described in (Avanzi, Bordal, Lacheret, Obin, & Sauvage-Vincent, 2014) — and syntactic disfluencies (or discursive markers), described in (Kahane & Pietrandrea, 2009). The second one was a combined analysis (Beliao, 2014) of the relations between intonational periods (IPes) (Lacheret & Victorri, 2002) and illocutionary units (IUs) (Pietrandrea et al., 2014). Unlike in a machine learning approach, the purpose here is not to obtain high performances over a categorization task, but rather to understand how two particular linguistic features interact. In the third experiment, we sought to detect redundancy in terms of information among all the annotations of the corpus (Beliao et al., 2014).

For the first study, we collected the prosodic and syntactic disfluencies (DM) in order to check whether the two kinds of disfluencies were related. To this purpose, we proposed a statistical analysis focusing on two separate aspects. The first one focused on the average number of prosodic disfluencies and DM per minute (hes/min and DM/min). Studying the scatter-plot showing one versus the other across all samples, we performed a correlation study. Then, we studied whether prosodic disfluencies and DM were systematically synchronized. We demonstrated through a synchronization analysis that this is not in fact the case.

The second study aimed at testing whether the density of IPes compared to that of IUs in a sample was characteristic of a particular speech genre (oratory, argumentative, descriptive and procedural). To this end, we identified two relational quantities: the first one was the degree of synchronization of IPes and IUs, and the second one was the ratio between the frequency of IUs and the frequency of IPes. This ratio was given in log scale to make the variable symmetrical and hence does not favor the rate of IUs per IPe compared to the rate of IPes per IU (because log (a/b) = -log (b/a)). This information may indicate the respective potential for inclusion of these two types of units: an IU/IPe ratio greater than 1 indicates that the sample contains more IUs than IPes, therefore IUs will probably be the unit that include IPes.

$$Ratio\,(sample) = \log \frac{number\,of\,IPe}{number\,of\,IU}$$

Finally, an experiment was conducted in order to detect the redundant features among all the annotations made in the Rhapsodie corpus. This study proved valuable in providing guidelines to reduce the burden of the annotation task, by eliminating redundant annotations.

Given the descriptive features of all the samples in the corpus, how can they be simultaneously and graphically represented, when there are for instance 27 features? The difficulty is that the samples are no longer represented in a two-dimensional space, but in a space of dimension 27. The objective then becomes to "summarize" this high-dimensional information in a lower-dimensional space (two dimensions here: x and y-axis). This is achieved through Principal Component Analysis (PCA) (Jolliffe, 2005; Hotelling, 1933), which basically rotates and then projects the data in a low dimensional space so that most of the information is preserved. As a side effect, redundant features are merged. In other words, PCA is a method of transforming the potentially correlated descriptive features into a new set of uncorrelated variables. These new variables are called "main components" and report information in a less redundant way.

3.2.2. Interpretation and results

With the first study, we demonstrated that the density of prosodic disfluencies in a sample is indeed strongly correlated to the density of syntactic disfluencies, even if the two notions are shown not to be equivalent. It is hence our belief that a joint analysis of prosody and syntax may lead to a better understanding of spontaneous speech. Since hesitations are the most frequent type of speech disfluency in many languages, it is possible that the majority of synchronized cases fall within the class of hesitations.

For the second study, the first experiment demonstrated that for a given sample, a much larger number of intonational periods than of IUs is characteristic of oratory while the reverse is true of descriptive speech for example. Furthermore, the synchronization of the prosodic and syntactic meta-units seems to be related to canonical speech, i.e. one in which syntax and prosody coincide regardless of speech genres. An interesting perspective could be to apply this framework to other languages. We hypothesize that the relative frequency of IPes over IUs is a distinguishing criterion to classify and characterize types of discourse. The difference between the observed IPe/IU ratio and the intuitively expected ratio is illustrated by a massive production of IPes compared to the number of IUs.

Lastly, PCA helped to illustrate the redundancy of some selected secondary variables. However, it was found that certain groups of distinctly annotated variables do operate together, leading to the conclusion that some annotations, even if interesting to the linguist, seem to

provide the same information as others and could hence be omitted for annotation efficiency.

4. Discussion: does quantity always mean quality?

The comparison of these two different approaches raises various questions: Should we be wary of the automatic selection of features? Or should we prefer the arbitrary choice of these features? Lastly, what kind of statistical tests should one make? And what do we think about hypothesis testing? And as an alternative what kind of results can we expect if we use machine learning?

4.1. Engineering perspective

From an engineering point of view, more features means better accuracy (Beliao et al., 2014), basically because machine learning techniques perform better using more data samples. That said, the overfitting problem is hard to quantify and is likely on small-sized corpora.

basically because machine learning techniques perform better when using more data samples. That said, the overfitting problem is hard to quantify and is likely to occur on small-sized corpora.

An engineering task is, above all, performance based. Many metrics can give valuable feedback, such as F-measure, recall and precision, ROC curve, etc. (Dumais et al., 1998; Sebastiani, 2002; Leopold & Kindermann, 2002; Forman, 2003; Pršir, Goldman, & Auchlin, 2013). To evaluate performance, one has to be careful to use different learning and testing datasets, possibly through cross-validation. For example, for the decision trees we found that the CART algorithm (Breiman, Friedman, Olshen, & Stone, 1984) yields significant error rates in cross-validation, which is explained by its high sensitivity to the "curse of dimensionality" (Fu, Carroll, & Wang, 2005), induced by a large number of features. If the number of descriptive variables (twenty-seven in our case) and situational variables (two hundred and sixteen potential combinations from 14 situational variables) increases, then this type of algorithm is less efficient due to a tendency to overfitting. In other words, it is not difficult to construct a decision tree that makes no error on a data set. However, it is likely that such a tree has very poor generalization capabilities. Why? Mainly because the tests it performs are too specific to the training data and do not capture the true ways of classifying the population for another set of test data. Using the same corpus for learning and testing should therefore be avoided at all costs.

When using manually annotated features as inputs to a classifier as in (Beliao et al., 2014), the objective is not really to improve over classification performance. Indeed, real-world use-cases would only exploit automatically annotated features. Rather, the objective is to assess whether those features are representative of some interesting aspects of

spoken language from a functional perspective. In this context, classification accuracy is used as a proxy for evaluating the relevance of some linguistic annotations for understanding spoken language.

4.2. Functional analysis

From a linguistic perspective, more features leads to worse qualitative value, because the main tendencies are sometimes hard to extract. From the point of view of functional analysis, it is tempting to give a functional interpretation of the computed classification models, but this approach is questionable, notably because of overfitting and too many features. Reducing the number of features in this functional analysis context is a good idea, even if it may lead to a poorer performance in terms of engineering accuracy.

To pursue linguistic interpretations further, some studies discard an engineering perspective in favor of hypothesis testing. This approach consists in first making an a priori functional hypothesis, and then in testing for the likelihood of this hypothesis on real data, using statistical tests (Beliao, Kahane, & Lacheret, 2013; Beliao & Lacheret, 2013; Beliao, 2014; Beliao et al., 2014). In this context, performance is no longer quantified through classification error, but rather on the basis of p-values and from a more theoretical linguistic perspective. Even if this direction does not yield direct engineering solutions to improve on the performance of a classification machinery, it provides great insight into the functional interactions within spoken language and may hence be insightful to the computational linguist.

References

Avanzi, M., Bordal, G., Lacheret, A., Obin, N., & Sauvage-Vincent, J. (2014). The annotation of syllabic prominencies and disfluencies. In A. Lacheret-Dujour, P. Pietrandrea, & S. Kahane (Eds), *Rhapsodie: a prosodic and syntactic treebank for spoken french* (chap. 3). New-York/Amsterdam: Benjamins. (in press)

Avanzi, M., Lacheret-Dujour, A., Obin, N., & Victorri, B. (2011). Vers une modélisation continue de la structure prosodique: le cas des proéminences syllabiques. *Journal of French Language Studies*, *21*(01), 53–71.

Avanzi, M., Simon, A.-C., Goldman, J.-P., & Auchlin, A. (2010). C-prom. un corpus de français parlé annoté pour l'étude des proéminences. In *XXVIIIe journées d'étude sur la parole (JEP'10)*.

Beliao, J. (2014). Characterizing speech genres through the relation between prosody and macrosyntax. In *New directions in logic, language and computation* (Vol. 8607). Springer.

Beliao, J., Kahane, S., & Lacheret, A. (2013). Modéliser l'interface intonosyntaxique. In *Proceedings of the prosody-discourse interface conference 2013 (IDP-2013)* (pp. 21–27).

Beliao, J., & Lacheret, A. (2013). Disfluencies and discursive markers : when prosody and syntax plan discourse. In *The 6th workshop on disfluency in spontaneous speech.* Stockholm, Sweden.

Beliao, J., Lacheret, A., & Kahane, S. (2014). Interface intono-syntaxique en français parlé : Compter quoi, compter comment, compter pourquoi ? *Langages*. (submitted)

Beliao, J., & Liutkus, A. (2014). Oops: une approche orientée objet pour l'interrogation et l'analyse linguistique de l'interface prosodie/syntaxe/discours. *CMLF2014*.

Benzitoun, C., Dister, A., Gerdes, K., Kahane, S., & Marlet, R. (2009). annoter du des textes tu te demandes si c'est syntaxique tu vois. *Arena Romanistica 4*, 16–27.

Benzitoun, C., Dister, A., Gerdes, K., Kahane, S., Pietrandrea, P., & Sabio, F. (2010). Tu veux couper là faut dire pourquoi. propositions pour une segmentation syntaxique du franạis parlé. *Actes du Congrès Mondial de Linguistique franạise*.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., & Pires, M. (2009). *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000.*

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks.

Dumais, S., et al. (1998). Using svms for text categorization. *IEEE Intelligent Systems*, *13*(4), 21–23.

Durand, J., Laks, B., & Lyche, C. (2009). Le projet PFC (phonologie du français contemporain): une source de données primaires structurées. In J. Durand, B. Laks, & C. Lyche (Eds), *Phonologie, variation et accents du français* (pp. 19–61). Lavoisier.

Eshkol-Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., & Tellier, I. (2011). A large available oral corpus: Orleans corpus 1968-2012. *TAL*, *52*(3), 17-46.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, *3*, 1289–1305.

François, J. (1998). Grammaire fonctionnelle et dynamique des langues: de nouveaux modèles d'inspiration cognitive et biologique. *Verbum*, *3*, 233–256.

François, J. (2008). Les grammaires de construction, un bâtiment ouvert aux quatre vents. *Cahiers du CRISCO*, *26*, 1–19.

François, J. (2010). Trois monographies récentes sur les parcours de grammaticalisation et la linguistique de l'usage. *Syntaxe & sémantique*, *11*, 185–203.

Fu, W. J., Carroll, R. J., & Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*, *21*(9), 1979–1986.

Gerdes, K., & Kahane, S. (2009). Speaking in piles: Paradigmatic annotation of french spoken corpus. *Proceedings of the Fifth Corpus Linguistics Conference*.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Jolliffe, I. (2005). *Principal component analysis*. Wiley Online Library.

Kahane, S., & Pietrandrea, P. (2009). Les parenthétiques comme «unités illocutoires associées». Une perspective macrosyntaxique. *Linx*, *61*, 49–70.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., … others (2014). Rhapsodie: a prosodic-syntactic treebank for spoken french. In *Language resources and evaluation conference (LREC-2014).*

Lacheret, A., Obin, N., & Avanzi, M. (2010). Design and evaluation of shared prosodic annotation for spontaneous french speech: from expert knowledge to non-expert annotation. In *Proceedings of the 4th linguistic annotation workshop* (pp. 265–273).

Lacheret, A., & Victorri, B. (2002). La période intonative comme unité d'analyse pour l'étude du francais parlé: modélisation prosodique et enjeux linguistiques. *Verbum*, *1*(24), 55–72.

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, *46*(1-3), 423–444.

Pietrandrea, P., Kahane, S., Lacheret, A., & Sabio, F. (2014). The notion of sentence and other discourse units in spoken corpus annotation. In H. Mello & T. Raso (Eds), *Spoken corpora and linguistic studies* (p. 331-364). John Benjamins Publishing Company.

Pršir, T., Goldman, J.-P., & Auchlin, A. (2013). Variation prosodique situationnelle: étude sur corpus de huit phonogenres en français. In P. Mertens & A. C. Simon (Eds), *Proceedings of the prosody-discourse interface conference 2013 (IDP-2013)* (pp. 107–111). Retrieved from http://wwwling.arts.kuleuven.be/franitalco/idp2013/Proceedings.html

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.