

# Affective Interaction with a Companion Robot for Hospitalized Children: a Linguistically based Model for Emotion Detection

Marc Le Tallec<sup>1</sup>, Jean-Yves Antoine<sup>1,3</sup>, Jeanne Villaneau<sup>2</sup>, Dominique Duhaut<sup>2</sup>

<sup>1</sup>LI, Université François Rabelais de Tours, France

<sup>2</sup>Valoria, Université Bretagne Sud, France

<sup>3</sup>Lab-STICC, CNRS, France

{marc.letallec,jean-yves.antoine}@univ-tours.fr, {jeanne.villaneau,dominique.duhaut}@univ-ubs.fr

## Abstract

This paper presents a system which aims at characterizing emotions in speech by only considering linguistic content. It is based on the assumption that emotions can be compound: simple lexical words have an intrinsic emotional value, while verbal and adjectival predicates act as a function on the emotional values of their arguments. The paper describes the compositional computation algorithm of the emotion, as well as the lexical emotional lexicons used by this algorithm. A quantitative and qualitative analysis of the differences between system outputs and expert annotations is given, which shows satisfactory results, with a good detection of emotional valence in 82.8% of the test utterances.

**Keywords:** Emotion detection, spoken language understanding

## 1. Scientific context

This paper deals with emotion detection, also called affect analysis, in spoken dialogs within a specific application context: the interaction between hospitalized children and an emotional companion robot whose purpose is to give its users some comfort<sup>1</sup>. The interaction consists of a child speaking to and manipulating the robot, a cuddly bear which reacts to the child's speech. The task consists of identifying the emotional state of the child during this interaction.

When considering spoken human-machine dialog, emotion detection is generally done through acoustic or prosodic clues (Devillers *et al.*, 2005). Several works have shown that system performances can be improved by the complementary use of linguistics clues (Devillers *et al.*, 2003). On the whole, these linguistic features restrict however to some limited cues such as contextual valence shifters (Ptaszynski *et al.*, 2010) which designates negation and intensifying words (very...). On the opposite, a deeper consideration of the linguistic content should be found in Natural Language Processing researches on written/electronic texts. This is particularly the case with sentiment analysis or opinion mining, where the task consists of identifying positive and negative opinions, emotions and evaluations in a document. For instance (Neviarouskaya, 2011) works on affect expressed by text messages on blogs or instant messaging, by using symbolic cues but also syntactic structures at phrase and sentence levels. Similarly, our approach aims at investigating a linguistic detection of emotions based on the semantic structure of the utterance in a different context: the system has to handle spoken language (ASR errors, speech disfluences), as well as an additional one: child language.

After a brief presentation of the context of this research, section 2 describes a pilot experiment where human subjects were asked to detect emotion by only considering linguistic content. Then, section 3 describes our detection model which combines a robust spoken language understanding system, an emotional lexical norm and the emotion recognizer based on a compositional calculus. The results given in Section 4 show that this approach is effective in this application context.

## 2. Human detection of emotions

There is currently no consensus about what an emotion is and how an emotion has to be characterized. An emotion is a complex cognitive state, which is strongly dependent on various contexts: the short-term context includes the type and the circumstances of the interaction, while the long-term context is related to cultural and personal life. To characterize emotion, two approaches are mainly used. In the first one, emotions are classed into emotional modalities. The selected modalities may be different according to the authors, but most of them agree with a classification into seven emotional modalities: anger, disgust, enjoyment, fear, surprise, sadness and neutrality (Ekman, 1999; Cowie and Cornelius, 2003). The second approach uses ordinal classification in a multidimensional space. For example, some psycholinguists use excitement level and emotional valence (negative/positive).

Whatever the approach is, all works show that a strong majority of speech turns do not convey a perceptible emotion in real dialogue. Besides, the perception of the emotion is very variable: inter-annotator agreement is always weak, what explains that a reference annotation must be achieved with a majority vote only (Devillers *et al.*, 2005; Callejas *et al.*, 2008).

This work investigates the benefits of achieving linguistic detection of emotions, which should be complementary to

---

<sup>1</sup> This work is part of the EmotiRob project founded by the French National Agency of Research.

acoustic or prosodic detection. Pure linguistic detection of emotions remains an original approach in speech technologies, the evaluation campaigns of which have also shown that the detection of emotional modalities is still a difficult task (Schuller *et al.*, 2009). Thus, this research is focused on the detection of valence and intensity (and not activation (Russell, 1980)), in which valence can be negative, positive or neutral and intensity is measured by an integer from 0 to 2.

## 2.1. Emotion annotation scheme

At first, we conducted a pilot experiment to assess whether or not such a linguistically-based detection of emotion was feasible for humans. This experiment consisted in measuring inter-annotator agreement when asking experts to determine the emotion conveyed in sentences by only considering their linguistic content. Although researches already focused on child language like in the CHILDES initiative (MacWhinney, 2000), an emotional corpus of child-robot interaction does not currently exist in French (unlike the German FAU AIBO robot (Steidl, 2009)). The choice was made to use fairy tales, working on the assumption that children could tell stories to the robot. Such corpora have already been used in research related to emotion detection (Alm *et al.*, 2005; Volkova *et al.*, 2010).

These experiments were carried out on an adaptation of the French fairy tale “Comment le Grand Nord découvre l’été” (How the big North discovered the summer). 31 subjects were asked to attribute an emotional value to every sentence through a single scalar value including valence and intensity between -2 (very negative) and 2 (very positive). In a first experiment, the sentences were given in a random order to investigate an out-of-context detection of emotion. In a second experiment, the story of the fairy tale was followed. A majority vote was used as gold standard to create a test reference from the annotations.

	Agreement between		
	100%-70%	70%-50%	50%-25%
<b>Without context</b>			
Valence	66%	29%	5%
Valence+intensity	31%	46%	23%
<b>With context</b>			
Valence	67%	31%	2%
Valence+intensity	28%	46%	26%

Table 1: Distribution of the annotation according to percentage of total votes of the 1<sup>st</sup> ranked vote.

Table 1 presents the distribution of the sentences according to the percentage of votes obtained by the winning annotation. For instance, when considering out-of-context annotation, there are only 5.4% of the sentences for which the winning annotation did not reach a majority of 50% annotation. A satisfactory correlation between annotators can be observed for the valence task, with an absolute majority agreement for 95% of out-of-context sentences and 98% of sentences in context. If intensity is considered then such agreement is more difficult to obtain, which was expected (Schuller *et al.*, 2009). Results point out that emotion can be seen as a

continuum, where votes are distributed on a segment of the emotion scale.

However this agreement is not perfect. Kappa’s  $\kappa$  measures (Fleiss, 1971) show indeed a moderate agreement on both situations (0.48 without context and 0.5 in context). If you consider intensity in addition to valence, such an agreement is more difficult to obtain. A higher dispersion of the annotations is indeed observed by a Kappa’s  $\kappa$  measure (0.31 without context and 0.33 in context). This shows a low agreement that has already been found in other studies (Schuller *et al.*, 2009; Vidrascu and Devillers, 2007). One should however observe that there is rarely a complete dispersion of votes from very negative to very positive values. This relative agreement is shown by a Krippendorff’s  $\alpha$  (Krippendorff, 1995) measure ( $\alpha = 0.72$  without context and 0.76 in context) which accounts better for an ordinal classification than  $\kappa$ , since it makes possible to introduce a weighted distance between the classes.

## 2.2. Influence of contextual information

Experimentations with contextual information have shown that annotators changed their opinions on some sentences when the sentences were presented in the regular order of the story. Figure 2 shows the global distribution of votes without and in context. A translation of votes can be seen from the neutral class to positive classes. It can be supposed that information from the

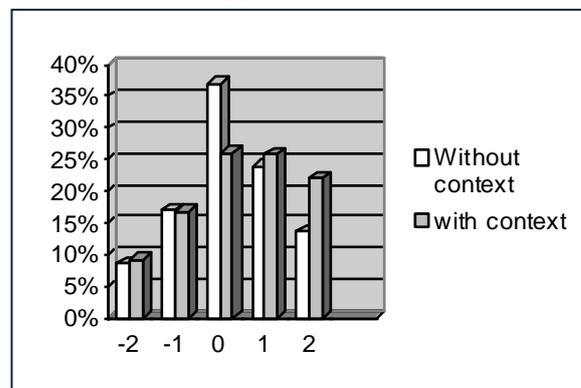


Figure 1: Distribution of emotions in function of percentage of votes in context and without context.

story and the evolution of characters modify the emotion felt. Indeed, one could easily find a correlation between the progressive shift towards a positive emotion and the fact that this fairy tale has a happy ending.

A Student’s t-Test on the difference between both annotations (without context and contextual) for each sentence leads to a value of 2.22. This result, with the degree of freedom, is statistically significant, with a risk threshold inferior to 0.05. Such a result suggests the importance of context during the process of emotion detection.

## 3. The EmoLogus system

### 3.1. Basic principles of the EmoLogus system

The previous experiments suggest that linguistic detection of emotion is pertinent for human subjects, especially when it is restricted to detection of valence.

Our objective is precisely to detect emotions which are conveyed in the propositional content of a linguistic message. Some works have shown that many words convey a positive or negative emotion by themselves (Bestgen, 2008). Thus, a simple first approach to obtain an emotional measure of a message is to add up the emotional measure of each of its words. An advantage of this solution is that no in-depth understanding of the message is required. Nevertheless, although such a method can be used to obtain the emotional content of a text (as shown by research on sentiment analysis) or, combined with prosody, to detect the emotional state of a speaker, it is obvious that it can also lead to serious errors in our application context. For example, in the utterance “the nasty witch is dead”, adding up the emotional measure of each word can lead to an erroneous very negative emotion. On the contrary, the system can assume that the emotional potential of the concept “to die” is dependent on its related subject: the semantic dependencies between the involved concepts must be taken into account.

Our work is based on these principles and aims at realizing a compositional calculation of the emotional content of the utterances. To do so, a semantic treatment of the utterances is required, which achieves their “understanding”, i.e. specifying the semantic linkages between concepts. More precisely, the treatment of an utterance is split into two steps: in the first step, a Spoken Language Understanding (SLU) system provides a logical formula to represent the meaning of the word sequence, which is provided as input by the Automatic Speech Recognition system (1-best sequence). Then, the system uses this logical formula and two emotional lexicons to calculate the emotional value of the utterance (Fig. 2).

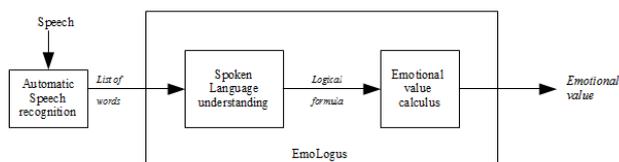


Figure 2: The architecture of the EmoLogus system.

Our SLU system is based on logical formalism and performs incremental deep parsing (Villaneau and Antoine, 2009). It builds a logical formula which represents the meaning of the word list, with which Automatic Speech Recognition provides it as input. It uses an ontology of the concepts known by the system, built from corpus studies, especially fairy tales (El Maarouf, 2009). This ontology is completely handcrafted, which remains feasible because it is built from a restricted vocabulary, the usual basic language of a 4 or 5 year child. This language is defined by using Bassano's studies related to the development of child language (Bassano *et al.*, 2005). Around 400 nouns, 250 verbs and 150 adjectives are classified into 30 semantic classes which are organized in a hierarchical tree. The logical formula that the SLU provides is built from these concepts, but its capacity of understanding is wider: the child lexicon used by the SLU system as its source language contains about 8000 lemmas selected from the child lexical Manulex (Lété, 2004) and Novlex, French-

speaking databases. SLU carries out a projection from this source language into the concepts of the ontology.

Consider for instance the utterance:

(1) “*Il était une fois une petite marmotte qui n'avait jamais vu le soleil*” (Once upon a time there was a little groundhog who had never seen the sun)

the logical formula provided by the system is:

(2) (*narrative* (*neg* (*to\_see* [(*subject*: (*groundhog*[(*size*: *little*)]), (*object*: (*sun*)]))]))

The predicate *narrative* points out that the context is a tale, as is shown by the sequence “*il était une fois*” (*once upon a time*). The negation is expressed by the predicate *neg*, which is applied to the action *to\_see*; the predicate *to\_see* has two arguments: a subject *groundhog*, which has the property of being small-sized (*size*: *little*) and an object: the *sun*. The calculation related to the emotion detection of this utterance from this formula is presented in the next section.

The emotional annotations described in the previous section have shown that the context plays an important part in the perception of the emotion conveyed in a linguistic message. Currently, our SLU includes a resolution of co-references: pronouns, possessive and demonstrative adjectives are linked to the entities to which they are related.

### 3.2. Evaluation of the emotional value

The detection of emotions relies on a major principle: the emotion conveyed by an utterance is compositional. It depends on the emotion of every word supposed to convey an emotion by our emotional lexicon, as well as the semantic relations characterized by the SLU system. More precisely, an intrinsic emotional value is given to the words - generally common nouns - which are used as simple arguments of predicates in the logical formula, while the words used as predicates - generally verbal and adjectival predicates - act as a function on the emotional values of their arguments.

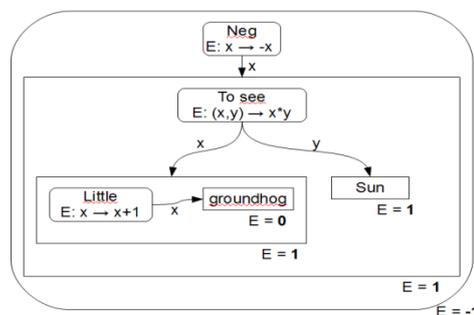


Figure 3: Compositional calculation for example (1).

As an illustration, consider sentence (1) given in the previous section and its related logical formula (2). The calculation of the emotion begins with the consideration of the emotional value of the words which are simple arguments of the formula, *groundhog* and *sun*:  $E = 0$  for *groundhog* and  $E = 1$  for *sun*. Then, the adjective, *little*, and the verb, *to see*, act as predicates on these initial values. For instance, *little groundhog* is assigned  $E = +1$

as emotional value. The adjective *little* has a sophisticated definition which is explained in the following section, since it is defined as the predicate  $E: x \rightarrow \text{if animate}(x) \text{ then } x+1 \text{ else } x$ . The last applied function is related to the negative, which is defined as  $(E: x \rightarrow -x)$  and transforms an emotional value into the opposite value. The successive applications of the predicates provide the global emotional value of the sentence:  $E = -1$  (figure 3). Our approach can be seen as original because all the words of the sentences are taken into account, whereas in most studies, only words conveying affective content are considered. Now, words which are considered as neutral can even so have an influence on the global emotional value conveyed by the sentence. For instance, the verb “survivre” (to survive) become negative with a negative subject (“*the nasty witch survived*”) and positive with a neutral one (“*a man survived*”).

Neutral words can also be used in desambiguation. Consider for instance utterances (3) and (4) with verb “to eat”:

(3) “*un ogre mange une pomme*” (an ogre is eating an apple)

(4) “*un ogre mange un homme*” (an ogre is eating a man)

In sentence (3), “apple” is neutral but it represents food, whereas in sentence (4) “man” is a person. The ontology makes it possible to use this type of information, and to precise the function associated with the verb “to eat”: neutral if food is eaten as in sentence (3) and negative in sentence (4).

These elements constrain us to take into account all elements of the sentence, including neutral words.

### 3.3. Emotional norm

In order to implement the above described calculation, two types of emotional data are needed: those which are related to simple arguments of the logical formula are, as previously stated, constant values. They are generally related to common nouns. The other type of data is related to the predicates, which correspond more or less to adjectives and verbs: As said before, they are modeled by functions which apply to the emotional values of their arguments. We will now present how these two emotional databases are obtained.

The first database describes the emotional valences which are associated to the lexicon words of children. This is precisely the aim of emotional lexical norms which have been used for a long time in experimental psychology. These normative word lists compile subjective evaluations of a representative population of judges on one or several emotional characteristics of words. Some of them are related to emotional characteristics, such as the duration of the emotion caused by a word (Zammuner, 1998; Niedenthal *et al.*, 2004). However, in all emotional lexical standards, two characteristics are always estimated: valence and intensity. Both characteristics are estimated using nominal scales of judgment (positive, neutral, negative) or ordinal (i.e., -5 very negative through 5 very positive).

To our knowledge, only two standards have compiled the evaluations made by young children: Vasa *et al.* (2006) for the English language, and Syssau and Monnier (2009)

for the French language (5, 7 and 9 year old children). These works show that children are able to judge the emotional valence of the words with substantial agreement and that their system of emotional value differs significantly with adults. This is why the emotional lexicon built by adults (WordNet Affect (Strapparava and Valitutti, 2004), SentiWordNet (Esuli and Sebastiani, 2006)) are not pertinent here. We have completed the standard of Syssau and Monnier with 80 new words. They were extracted from the Bassano lexicon, for children between the ages of 5 and 7, in order to describe completely our application lexicon. This extension respects psycholinguistic requirements and was conducted in four primary schools in collaboration with Syssau and Monnier. The achievement of such a psycholinguistic norm requires a costly human effort. It guarantees, however, the relevancy of our emotional lexicon and remains quite feasible, because of the restricted size of the lexicon of young children. In addition, one should note that data analysis techniques, like Latent Semantic Analysis, should be used to automatically extend an emotional lexicon from such a restricted norm (Bestgen, 2008).

To complete the characterization of our lexicon, an emotional predicate has been assigned to every verb or adjective of our application lexicon through an agreement procedure among five experts. These functions have been proposed by introspection, without considering the test corpus. More precisely, every expert was asked to propose one or at most two definitions for every predicate from a predefined set of functions. Then, an agreement was sought among these proposals. It is interesting to note that complete agreement was finally reached. As an illustration, here are some examples of predicates which concern adjectives:

*gentil* (kind)  $E: x \rightarrow x + 1$

*énervé* (irritated)  $E: x \rightarrow x - 1$

For these examples, adjectives have a general positive or negative influence on their argument. However, this effect may be related to the type of argument. This is the case of the adjective *petit* (little), the definition of which is  $E: x \rightarrow \text{if animate}(x) \text{ then } x+1 \text{ else } x$ . It can be interesting to note that this definition was not proposed by the experts directly, but has been inferred from the results of the annotations described in section 2.

The emotional value of predicates related to an intransitive verb is modeled by a function of one variable, which may be constant. For example, the emotional functions related to the verbs *sourire* (to smile) and *mentir* (to lie), are respectively  $E: x \rightarrow x+1$  and  $E: x \rightarrow -1$ , which can be interpreted as “*smiling is a way of being perceived as a more positive or less negative character*” and “*lying is always negative*”.

The emotional functions related to transitive verbs are modeled as a function of one or more variables: subject and object. For example,  $E: (x,y) \rightarrow -y$  (*casser* / to break), for which the result is dependent on the broken object, or  $E: (x,y) \rightarrow \min(x,y)$  (*accompagner* / to go with), where the result is the minimum of the emotional values related to both involved characters. This enables one to model correctly the two sentences “the nasty witch goes with the girl” and “the nasty which goes with the monster” which both convey a negative emotion, while “the girl goes with the little cat” conveys a positive one.

## 4. Results

### 4.1. *EmoLogus* System evaluation

Several experiments were conducted to assess the behavior of our system. Since previous experiments have already assessed the robustness of our SLU module (Villaneau and Antoine, 2009), this study focuses on child language, without considering speech recognition errors.

We compared the *EmoLogus* system with the baseline on the reference corpus presented in section 2. The performance measure is accuracy: indeed, we considered all of the annotations, including neutral annotations. *EmoLogus* always takes a decision In a micro-averaged computation framework (Antoine *et al.*, 2011), recall is equal to 100% and accuracy corresponds to precision. Results are shown in Table 2.

Three main observations should be drawn from these results. At first, *Emologus* presents a significantly higher precision than the baseline in all situations (without/in context, valence alone or with intensity). When considering valence alone, this precision remains quite satisfactory (82.8 % without context and 75.3 % in context). This shows the benefit of considering the semantic structure of the sentence rather than performing a bag-of-words approach. The second conclusion concerns the detection of intensity. Clearly, the accuracy of *EmoLogus* is lower on the valence+intensity task. This must not be considered as a weakness of the system. Indeed, the low inter-annotator agreement observed during the experiment, which was presented in section 2, showed that this task is difficult even for human beings. Finally, a slight decrease in precision was observed when considering context. We are precisely working on an even better consideration of the context.

Valence	<i>EmoLogus</i>	Baseline
Without context	82.8 %	64.4 %
With context	75.3 %	53.8 %
Valence + intensity	<i>EmoLogus</i>	Baseline
Without context	69.9 %	52.7 %
With context	59.2 %	39.8 %

Table 2: Precision of the systems.

Beyond these general results, it is interesting to study the distribution of errors of both systems. We have distinguished four kinds of errors:

- **valence inversion** is certainly the most important kind of error: it corresponds to situations in which the system detects an emotion opposite to the reference (e.g. positive vs. negative)

- **emotion deletion** corresponds to situations in which the system fails to detect a positive or a negative emotion, while classifying the spoken utterance as neutral.

- **emotion insertion** corresponds to the opposite situation: a neutral emotion was expected while the system detects a positive or a negative one.

- finally, **intensity errors** corresponds to less serious errors: here, the valence is correctly detected but the system and the reference disagree on the intensity.

Table 3 presents the resulting distribution of errors on the out-of-context emotion task. As expected, the majority of errors correspond to the less serious ones. Indeed, 43% of the *EmoLogus* errors are only intensity errors. This explains the decrease of accuracy observed between the “valence” and the “valence+intensity” task. Fortunately, valence errors correspond above all to emotion insertions (30%) and emotions deletions (23%). Lastly, serious valence errors only represent 4 % of the errors, which actually correspond to one error in a very specific situation. On the contrary, the behavior of the baseline is less satisfactory. In particular, 3 valence inversions (7 % of the errors) can be observed with this bag-of-words system.

Without context	Valence inversion	Emotion insertion	Emotion deletion	Intensity errors
<i>EmoLogus</i>	4%	30%	23%	43%
Baseline	7%	37%	30%	26%
With context	Valence inversion	Emotion insertion	Emotion deletion	Intensity errors
<i>EmoLogus</i>	3%	5%	52%	40%
Baseline	12%	13%	51%	24%

Table 3: Distribution of errors.

In order to clearly situate the performances of *EmoLogus*, we wanted to verify how human beings would behave on the same “gold standard” reference. The scheme was identical to the experiment with *EmoLogus*: the individual votes of our annotators have been compared with the reference. Table 4 shows the ranking obtained by *EmoLogus* and the baseline inside this test population. This classification is done on 31 human annotations and 2 automatic ones.

Without context, it is very interesting to note that *EmoLogus* obtains better results than many annotators (3<sup>rd</sup> and 7<sup>th</sup> rank). As expected, the results are less satisfactory when context is considered, but they still remain better than some annotators. However, the bag-of-words baseline continues to give poor results.

	Without context		With context	
	Valence+intensity	Valence only	Valence+intensity	Valence only
<i>EmoLogus</i>	3 <sup>rd</sup>	9 <sup>th</sup>	19 <sup>th</sup>	27 <sup>th</sup>
Baseline	31 <sup>st</sup>	31 <sup>st</sup>	33 <sup>rd</sup>	33 <sup>rd</sup>

Table 4: Rank of systems in comparison task.

### 4.2. Influence of speech recognition

As our system is working on speech treatment, it is important to assess the influence of automatic speech recognition errors. A pilot experiment has been conducted on a corpus of fairy tales, compounded of 476 sentences and totalising 4644 words. By means of Dragon Naturally Speaking, a speaker has read tales and produce textual data. We observed a word error rate of

18.3 %. Main errors are substitution of words (81.76%) and insertion (16.47%). It leads to an error of understanding of 34.4% in term of CER (Concept error rate) producing errors in emotion detection (27.96%). Tests have been realized with the standard dictionary of Dragon Naturally Speaking recognition system, which was obviously not adapted to our application. We are now working on the adaptation by eliminating terms which do not belong to children language.

## 5. Conclusion

While most studies on emotion detection focus on prosodic cues, our experiment with human annotators shows that emotion can be detected using only linguistic clues. Considering both the emotional valence of words and the semantic structure of the sentence, our system *EmoLogus* presents a behavior close to that of human beings. There is still room for improvement in terms of context modelling, which restricts for the moment being only to anaphora resolution. For instance, we are now investigating how the emotional value of each lexical concept evolves according to the context. Obviously, the results provided by *EmoLogus* should be combined with the output of a prosodic detection of emotions. Nevertheless, the experiments we are currently conducting in real conditions tends to show that children feel already some comfort while interacting with our EMI robot, while the latter only integrate our linguistically based model.

## References

- Alm, C. O. , Roth, D., Sproat, R. (2005). Emotions from Text: Machine Learning for Text-based Emotion Prediction, In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada, pp.579-586.
- Antoine, J-Y, Le Tallec M., Villaneau, J.(2011). Evaluation de la detection des émotions des opinions ou des sentiments :Dictature de la majorité ou respect de la diversité d’opinion ?, Proc *Taln’2011*.
- Bassano, D., Labrell F., Champaux C., Lemétayer F.,Bonnet P. (2005). Le DLPF, un nouvel outil pour l’évaluation du développement du langage de production en français, *Enfance*, 2(5):171-208.
- Bestgen, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis, Proc. *LREC’2008*.
- Callejas Z., Lopez-Cozar R. (2008). Influence of contextual information in emotion annotation for spoken dialogue systems, *Speech Communication*, 50 (2008), 416-433
- Cowie R., Cornelius R. (2003). Describing the emotional states that are expressed in speech, *Speech Communication*. 40. 5-32.
- Devillers, L., Vidrascu,L. , Lamel, L. (2003). Emotion Detection in a task-oriented Dialog Corpus, Proc of IEEE International Conference on Multimedia, ICME, Baltimore.
- Devillers, L., Vidrascu,L. , Lamel, L. (2005). Emotion detection in real-life spoken dialogs recorded in call center”, *Journal of Neural Networks*, 18 (4), 407-422.
- Ekman, P. (1999) *Patterns of emotions: New Analysis of Anxiety and Emotion*. Plenum Press.
- El Maarouf I. (2009). Natural ontologies at work: investigating fairy tales, Proc. *Corpus linguistic 2009*.
- Esuli A., Sebastiani F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining, Proc. *LREC’2006* , 417-422.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters, *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382.
- Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology*, 25.
- Lété, B., Sprenger-Charolles, L., Colé, P. (2004). MANULEX : A grade-level lexical database from French elementary-school readers, *Behavior Research Methods, Instruments, & Computers*, 36, 156-166.
- MacWhinney, B., (2000). The CHILDES project : Tools for analyzing talk. 3<sup>rd</sup> edition. Mahwah, NJ: Lawrence Erlbaum associates.
- Neviarouskaya, A., Prendinger, H., Ishizuka M. (2011). Affect Analysis Model: Novel Rule-based Approach to Affect Sensing from Text. In: *International Journal of Natural Language Engineering*, Cambridge University Press, Vol. 17, No. 1, 2011, pp. 95-135
- Niedenthal P. M., Auxiette C., Nugier A., Dalle N.,Bonin P., Fayol M. (2004). A prototype analysis of the French category émotion. *Cognition and Emotion*, 18. 289-312.
- Russell, J. A. (1980). A Circumplex Model of Affect, J. Personality and Social Psychology, vol. 39, no. 6, pp. 1161-1178, 1980.
- Schuller B., Steidl S., Batliner A. (2009). The Interspeech’2009 emotion challenge. Proc. *Interspeech’2009*.
- Steidl S. (2009). Automatic Classification of Emotion-Related User States in Spontaneous Children’s Speech, Logos Verlag, Berlin(chapter 5)
- Strapparava C., Valitutti A. (2004). WordNet-Affect: an affective extension of WordNet, Proc. *LREC’2004* , 1083-1086.
- Syssau A., Monnier C. (2009). Children's emotional norms for six hundred French words, *Behavior, Research, and Methods*, 41, 213-219.
- Vasa R. A., Carlino A. R., London K., Min C. (2006). Valence ratings of emotional and non emotional words in children, *Personality and Individual Differences*, 41, 1169-1180.
- Vidrascu L., Devillers L. (2007). Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features, Proc. *Inter. workshop on Paralinguistic Speech between models and data, ParaLing*.
- Volkova E., Mohler B., Meures D., Gerdemann D. and Bühlhoff H. (2010). Emotional perception of fairy tales: achieving agreement in emotion annotation of text, Proc. *NAACL HLT 2010*.
- Villaneau, J., Antoine, J-Y. (2009) Deeper Spoken Language Understanding for Man-Machine Dialogue on Broader Application Domains, Proc. *EACL workshop SRS’2009*, Athens. 50-57.
- Wilson,T., Wiebe,J., Hoffman, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, Proc. *HLT/EMNLP 2005*, Vancouver, Canada.
- Zammuner V.L. (1998). Concepts of emotion: Emotioness and dimensional rating of Italian emotion words, *Cognition and emotion*, 12, (2), 243-272.