# Text and Speech Corpora for Text-To-Speech Synthesis of Tales

**David Doukhan[1], Sophie Rosset[1], Albert Rilliard[1], Christophe d'Alessandro[1], Martine Adda-Decker[2]**

[1] LIMSI-CNRS, UPR 3251, 91403 Orsay – France
[2] LPP-CNRS, UMR 7018, 75005 Paris - France
E-mail: {doukhan,rosset,rilliard,cda,madda}@limsi.fr

**Abstract**

Text and speech corpora for training a tale telling robots have been designed, recorded and annotated. The aim of these corpora is to study expressive storytelling behaviour, and to help in designing expressive prosodic and co-verbal variations for the artificial storyteller). A set of 89 children tales in French serves as a basis for this work. The speech annotation principles and scheme are described, together with the corpus description in terms of coverage and inter-annotator agreement. Automatic analysis of a new tale with the help of this corpus and machine learning is discussed. Metrics for evaluation of automatic annotation methods are discussed. A speech corpus of about 1 hour, with 12 tales has been recorded and aligned and annotated. This corpus is used for computing expressive prosody in children tales, above the level of the sentence.
**Keywords:** corpus, expressive TTS, annotation agreement, storytelling

## 1. Introduction and motivations

This work, in the framework of ANR GV-LEX project (Gelin *et al.*, 2010), aims at providing a humanoid robot with storytelling abilities. The situation envisaged is a NAO robot telling tales to 7-9 years hold children. Our task in the project concerns mainly raw text analysis, prosodic prediction, and time line definition for coordinating the robot's stream of gestures and speech, mimicking the situation of an adult telling a story to children.

A performance of storytelling is an aesthetic piece of art in itself, which reflects the storyteller's cognitive representation of the story, mixed with effects intended to capture audience attention. The storyteller, while reading the tale, is planning a series of expressive effects to make his performance entertaining. Expressive effects come partly from the text itself (the lexical and stylistic choices of the author), and consequently are given to the reader. The textual material is augmented in live performances with the performer's involvement in reading, expressed through gestures, vocal and prosodic variations. Many possible strategies are available to the reader including gestural and kinesic behaviours, and a wide range of impersonated voices.

Previous works on storytelling addressed some aspects of these expressive possibilities. Alm & Sproat (2005) analysed emotions induced by prosodic variations in storytelling. Klabbers &Van Santen, (2004) studied prosodic patterns variation at the foot level for English tales. Levin, (1982) investigated the linguistic and prosodic differences between storytelling and story reading. Theune *et al.* (2006) studied prosodic variations associated to a growing suspense in Dutch tales, while Adell *et al.* (2005) described the changes associated to with discourse modes (narrative, descriptive, dialogue), and with the expression of basic emotions (anger, neutral, sadness, surprise). All these studies emphasized the important role played by planning for expressive and consistent tale reading.

Planning of expressive storytelling seems to encompass the level of the entire tale, taken into account its structure. Although they are can contain a high degree of polysemy, tales exhibit interesting structural properties that can be exploited for our purpose. Their structures are somewhat constrained. They have been described by linguists with the help of several structural schemes (Propp, 1968; Greimas, 1966). Tales usually feature several stereotypical characters ("hero", "villain", "helper"...). Taletellers often vary their expressivity according to character impersonation at a *supra-sentential level*.

Nowadays, state of the art Text-To-Speech synthesizers (**TTS**) are producing high naturalness speech sentences, and have led to several real-world applications (vocal servers, human-computer interfaces for blind users...). However, prosodic synthesis strategies are generally restricted to the sentence level. Each sentences is synthesized in isolation, without taking into account its context or higher order information. These limitations generally result in monotonous speech streams, which eventually may become boring when the listening to texts larger than a few sentence. Synthesizing entertaining speech for storytelling is thus a particularly tough problem, since it requires inferring expressive prosodic variations, based on a global understanding of the input text. Another well-known difficulty in expressive TTS lies in the trade-off between sound naturalness, large prosodic variation, and artefacts related to sound processing procedures (Burkhardt & Stegman, 2009). Automatic expressive storytelling requires procedures for automatic semantic analysis, and procedure for prosodic and gestural synthesis according to this knowledge.

Our approach is aiming at automatic analysis of raw texts, for improving tales reading by a non-uniform units TTS synthesiser, and to provides the robot's gestural controller with relevant information for managing its kinesic behaviour. The present paper describes the

corpus of text tales used to train the automatic tale analyser, and the corpus of read tales used to infer realistic prosodic strategies *above the sentence level* to drive the TTS system. The expressive gestural aspects of this project are described in Pelachaud *et al.* (2010).

The paper is organized as follows. Section 2 provides a description of the tales contained in the text corpus. In section 3 we present the manual annotation tasks that were performed on the corpus, together with inter-annotator agreement estimations for these tasks. Section 4 present the annotations obtained through automatic procedures, and used to enrich the corpus. The read tale corpus, designed for analysing the correlations between prosodic properties and the annotations made on texts, in presented in section 5. Section 6 presents a discussion on the reliability of such annotation tasks, in the light of our first prosodic analyses on the speech corpus. It also provides some perspectives for enhancement of the annotation scheme, and discussed of future work in the context of the GV-LEx project.

## 2. Text Corpus

### 2.1. Text corpus

A corpus of 89 freely available tales in French was collected from the website http://www.contes.biz. Tales were selected in order to meet the following criteria:

- They are suited to a 7-8 years old audience
- They are readable in about 5 minutes
- They contain at least 2 speaking characters

### 2.2. Normalization

Tales were converted from HTML to text using LIMSI's processing tool *Wmatch* (Galibert, 2009; Rosset *et al.*, 2009). Line breaks and paragraph markers were mapped to carriage return symbols, while other structural of formatting tags were removed. Orthographic and grammatical errors were hand-corrected.

Tale texts were normalized using LIMSI's normalization software (Adda *et al.*, 1997). Word and punctuation markers were separated using a single blank space. "Carriage return" symbols inferred from initial texts were kept, to provide paragraph information. A "carriage return" symbol was inserted after each sentence. Ambiguous punctuation marks (-, ') were recognized and processed. Compound words joined with hyphens ("*chauve-souris*") were considered as single words. Clitics were split ("*j'ai*" was split into word "*j'*" and word "*ai*"). Capital characters were kept only for proper nouns.

### 2.3. Description

Table 1 present the main features of tales constituting the corpus. Tales contain an average of 752 word and 61 sentences, and 20 paragraphs. Each tale contains at least two paragraphs, one corresponding to tale title, and the other corresponding to tale text. A large variability was observed on the usage of paragraph marks. While some tales texts are split in two paragraphs, some other contain up to 40 different paragraphs. For tale containing a large number of paragraphs, we observed author indentation strategies consisting to insert paragraph marks between each transition between narrator and tale character impersonation. Low mean sentence sizes usually correspond to tale containing many character turns, associated to simple post-quotation patterns:

```
"- Oui !
dit la souris."
("-Yes !
Said the mice."
```

Long sentence sizes were observed for tales having repetitive structures, for narrative purposes:

```
"viens ronger le cordon , qui refuse d' étrangler le forgeron , qui refuse de briser cette lame , qui refuse de tuer le taureau , qui refuse de boire l' eau , qui refuse d' éteindre ce feu , qui refuse de brûler ce bâton , qui refuse de frapper Brirouch , qui refuse de dîner , parce qu' il a perdu son chevreau ."
("Just come to gnaw the cord, which refuses to strangle the blacksmith, who refuses to break the blade, which refuses to kill the bull, who refuses to drink the water, which refuses to extinguish this fire that refuses to burn the stick, who refuses to strike Brirouch, who refuses to dinner, because he lost his kid. ")
```

| | Mean | Min | Max | Total |
|---|---|---|---|---|
| Nb words | 751.9 | 309.0 | 1054 | 66922 |
| Nb sentences | 61.1 | 26.0 | 131 | 5438 |
| Nb paragraphs | 19.6 | 3.0 | 41 | 1741 |
| Mean sentence size | 13.1 | 6.8 | 22.1 | |
| Max sentence size | 40.2 | 22 | 116 | |

Table 1: Main features of the 89 tales in the Text Corpus

## 3. Manual Annotations

### 3.1 Protocol

Two annotators, trained in linguistics, referred as A1 and A2, labelled normalized tales texts manually.

Both annotators (in order to compute estimates of inter-annotator agreement and get insights on the reproducibility of the tasks at hand) labelled the same 7 tales of the corpus. Annotator A1 labelled a total of 61 tales, and annotator A2 labelled 35 tales, leading to a total of 96 distinct annotations.

A hierarchical annotation scheme inspired by (Propp, 1928; van Dijk, 1982; Golden, 1985) was defined to represent tale episodic structure, speech acts, references to tale characters, and linguistic information. A simplified example is shown in figure 1. Remember that these annotations are intended for designing prosodic and gesture synthesis rules. Automatons and machine learning algorithms in a second stage of the project should ultimately infer them automatically.

## 3.2. Episodes

### 3.2.1. Definition

Top-level structural annotations consist in segmenting the text into episodes. The segmentation was done based on several standard clues: paragraph indentation, time or place change markers, and introduction of new characters. Special kinds of episodes were tagged using the following labels: title, exposition (Proppian initial situation; Propp, 1928), triggering event (text between the exposition and the Proppian Departure of the hero), epilogue (ending of the tale that may contain a moral, and include Proppian recognition, exposure, transfiguration, punishment, wedding), refrain (episodes having nearly identical surface manifestation within the tale). Other kinds of episodes were tagged as scene.

This information is potentially useful for TTS: van Dijk, (1982) observed long pauses or hesitations phenomena associated to the beginning of episodes. Triggering events of the tale can possibly be associated with suspense patterns described by Theune *et al.* (2006).

### 3.2.2. Repartition

Table 2 shows the number of tales containing the episodic structure that were defined. All tales were tagged as having at least a title and a scene. Some of them were not considered as having an exposition, a triggering event, or an epilogue. Refrains were observed in 19% of the tales. Table 3 displays the coverage of each episode type.

| Title | Exposition | Trigger | Scene | Refrain | Epilogue |
|---|---|---|---|---|---|
| 96 | 94 | 89 | 96 | 18 | 95 |

Table 2: Number of tales containing at least one occurrence of the scenic categories

| Title | Exposition | Trigger | Scene | Refrain | Epilogue |
|---|---|---|---|---|---|
| 0.6 | 11.6 | 9.0 | 71.3 | 2.6 | 4.9 |

Table 3: Corpus Coverage percentage for each scenic category

### 3.2.3. Inter-annotator Agreement

Episodic labelling is a particular case of text segmentation into block, associated with a labelling of the blocks. We used the normalized sentence as the atomic unit for evaluating the reliability of this task. Title annotation was trivial, and ignored in our observations. The mean episodic segment length observed was 6.95 sentences for Annotator A1 and 4.91 for A2.

Table 4 shows the confusion matrix of episodic segmentation. The agreement on refrains was maximal. The annotators did not necessarily agree on the presence and on the difference of the exposition and of the triggering event. They agree on the beginning of the epilogue for 5 tales out of 7. While significant agreement was observed for scene boundaries, the tendency of A2 to consider smaller segments lead to several differences.

While this these information provide a qualitative description of the agreement, it cannot be considered as it as an agreement metric. Moreover, it does not take into account the difference between small and large error boundaries. Several metrics specific to text segmentation tasks that do not take block labels into account were proposed in the literature, In (Hearst, 1997), Cohen's Kappa was used to compute inter-annotator agreement.

Text segmentation evaluation methods were also proposed (*WindowDiff*, Pevzner and Hearst, 2002; *Generalized Hamming Distance*, Bestgen, 2009), requiring segmentation as a reference. These measures are difficult to interpret, and may be biased by degenerated examples.

| A1\A2 | expo | trigg | scene | epilog | refrain | inside |
|---|---|---|---|---|---|---|
| Expo | 6 | 0 | 0 | 0 | 0 | 0 |
| Trigger | 1 | 2 | 1 | 0 | 0 | 2 |
| Scene | 0 | 1 | 25 | 0 | 0 | 7 |
| Epilog | 0 | 0 | 1 | 5 | 0 | 1 |
| Refrain | 0 | 0 | 0 | 0 | 8 | 0 |
| Inside | 0 | 4 | 29 | 2 | 0 | 322 |

Table 4: Confusion Matrix of Episode boundary segmentation for Annotator A1 and A2. *intro*, *trigger*, *scene*, *epilogue*, *refrain* stands for the first line considered to be enclosed in the corresponding category; *inside* stands for sentences not considered to be related to a boundary.

To help interpretation and comparison of these results, we defined several segmentation automatons. The NoB (No boundary) associates single boundary to each new tale. AlB (All boundaries) associates a boundary to each sentence. R1 and R2 set random boundaries, using the same boundary frequency of annotator A1 and A2. PB set a scenic boundary at new paragraph mark, identified by a double carriage return symbol. PSB: Set boundary for new paragraphs only if the first sentence correspond to the narrator speech turn (more details on speech turns in section 3.3).

Table 5 reports segmentation similarity between annotator A1 and A2 using several metrics. The *WindowDiff* measure was obtained using a window size defined as:

$$k = int\left(\frac{nb\ sentences}{nb1+nb2}\right)$$

With nb1 and nb2 being the number of boundaries of the sequences being compared. The Generalized Hamming Distance was obtained using a shift cost of 2, with insertion and deletion cost set to:

$$cost = int\left(\frac{2 * nb\ sentences}{nb1 + nb2}\right)$$

Both measures showed a highest agreement between both annotators than between any other automatons.

### 3.3. Speech Turns

#### 3.3.1. Definition

The speech turn structural level was used for distinguishing the narrator's speech from tale character's speech. Each speech turn was labelled using a distinct identifier for the narrator and each tale character.
A restricted definition of sentences was used, such that they cannot cover adjacent speech turns.

#### 3.3.2. Description

A total of 418 distinct speaking characters were annotated in the corpus. The mean number of speaking character per tale is 4.35, and the maximum number of speaking characters found was 14. Tale character's speech turns covered a minimum of 5% of tale texts, and a maximum of 72%, with a mean cover- age percentage of 30.5%.

| | A1 | A2 | NoB | AlB | R1 | R2 | PB | PSB |
|---|---|---|---|---|---|---|---|---|
| *Cohen's Kappa* | | | | | | | | |
| A1 | 1.00 | 0.89 | 0.87 | 0.14 | 0.76 | 0.71 | 0.77 | 0.82 |
| A2 | 0.89 | 1.00 | 0.81 | 0.20 | 0.71 | 0.67 | 0.78 | 0.83 |
| *WindowDiff* | | | | | | | | |
| A1 | 0.00 | 0.27 | 0.89 | 0.85 | 0.66 | 0.65 | 0.60 | 0.60 |
| A2 | 0.27 | 0.00 | 0.94 | 0.79 | 0.63 | 0.68 | 0.52 | 0.42 |
| *Generalized Hamming Distance* | | | | | | | | |
| A1 | 0 | 197 | 659 | 624 | 421 | 451 | 380 | 362 |
| A2 | 197 | 0 | 707 | 551 | 426 | 422 | 284 | 273 |
| *F-Measure* | | | | | | | | |
| A1 | 1.00 | 0.69 | 0.21 | 0.25 | 0.20 | 0.14 | 0.44 | 0.46 |
| A2 | 0.69 | 1.00 | 0.15 | 0.34 | 0.21 | 0.24 | 0.54 | 0.58 |

Table 5: Similarity comparison of scene boundary localization, between annotator A1, annotator A2 and simple automatons (NoB, AlB, R1, R2, PB, PSB).

#### 3.3.3. Inter-annotator Agreement

Computing agreement for the speech turn labelling task consist in checking if speech turn boundaries are the same, and if references to tale characters are consistent between annotators. Estimates of agreement for this task were obtained using the implementation of MUC, B-Cubed, CEAF and Blanc metrics provided in Uzuner *et al.* (2012). Table 6 displays F-measures for these 4 metrics, considering the average obtained using annotator A1 and A2 as reference.

| MUC | B-Cubed | CEAF | Blanc |
|---|---|---|---|
| 0.98 | 0.965 | 0.94 | 0.99 |

Table 6: Agreement estimation for the speech turn labelling task.

| MUC | B-Cubed | CEAF | Blanc |
|---|---|---|---|
| 0.94 | 0.88 | 0.72 | 0.97 |

Table 7: Agreement estimation for tale character references.

### 3.4. Lexical Level Annotations

The last structural level refers to passages with enumerations, such as: "*pas le blé, ni les noix ni le pain dur.*" ("not wheat, nor nuts nor stale bread."), and to the elements of enumerations which start the lexical level of annotation (in the above example, elements of the enumeration are: *<pas le blé>*, *<ni les noix>*, *<ni le pain dur>*).
Lexical level tagging was performed for named entities (time and place), and other entities such as nominal group and adverbial locutions (MWE).

Table 8 displays the overall inter-annotator agreement (IAA) for these annotations using the Kappa metric and the F-measure.

| | Kappa | F-measure |
|---|---|---|
| location | 0.75 | 0.77 |
| time | 0.71 | 0.73 |
| MWE | 0.70 | 0.76 |
| enum | 0.71 | 0.73 |

Table 8: Inter-annotator agreement in term of Kappa and F-Measure for all lexical entities

### 3.5. Tale character References

Tale character references (noun, pronoun, lexical groups) were tracked using an identifier per character (humans, animals, plants, speaking objects). Table 7 displays agreement estimations corresponding to average F-Measure using MUC, B-Cubed, CEAF and Blanc metrics.

### 3.6 Tale character Meta Information

Meta information was associated to each speaking tale character, having at least one speech turn (see section 3.3). Character's age was encoded using categories: kid, teenager, adult, old. Possible gender categories were male, female, or neutral (e.g. the element "fire"). Character's kind was described using categories like human, wolf, fairy, knife... Valence was described as good, bad or neutral. Relative height was discretized using the labels small, medium, large, extra large. Description based on Proppian "actant" theory (Propp, 1928 – aggressor, donor, auxiliary, princess and the father, committer, hero, bogus hero) was used. Greimas (1966) definition of "actants" (sender, subject, supporter, object, receiver, oppositionist) was also used for each speaking character.
Tale character meta-information tagging was only per-formed for the tales selected in the recorded speech corpus (see section 5.).

## 4. Automatic Annotations

Grammars based on regular expressions were defined in the framework of the project and used with *Wmatch* (Galibert, 2009; Rosset *et al.*, 2009) to obtain automatically three complementary classes of sentence-level annotations. The first class is based on

agent communicative acts (Berger & Pesty, 2005; Rivère *et al.*, to appear), providing labels assertive, attractive, directive and expressive. Other classes related to speaking mode (laugh, cry, shout, exclamation) and classical dialog acts (inform, reject, request, request order, interdiction) were used only for sentences uttered by tale characters. This distinction was done because we considered tale characters able to laugh or cry, while the narrator can only tell that somebody is crying or laughing.

Identification of post-quotation clauses (e.g. "I'm hungry! Said the dog.") was motived by the assumption that it would be associated with shorter pauses between the two sentences, and changes in pitch register and voice quality. These clauses were identified with automatons using speech turn labels and part of speech tags.

Grammars were defined to mark intensification patterns, which could be associated with prosodic or gestural emphasis. These patterns were defined as intensification adverbs followed by adjectives ("He was so hungry").

Part Of Speech tags (POS) were obtained using LIMSI internal tools (Allauzen & Bonneau-Maynard, 2008). Tree-Tagger (Schmid, 1994) was used as a way to obtain lemma, and another source of POS. Word stemmed-forms were obtained using NLTK (Loper & Bird, 2002) implementation of the *SnowBall* Algorithm.

| Deletions | Insertions | Modification | Shifting |
|---|---|---|---|
| 40 | 31 | 41 | 4 |

Table 9: Number of differences between original tale texts (9664 words), and the recordings of the audio corpus.

## 5. Speech Corpus

### 5. 5.1. Motivation

The speech corpus has been built for analysing correlations between tales annotations described in the previous section, and their prosodic realization. These analyses serve a double goal: inferring mapping rules from the proposed annotation scheme to prosodic instructions used by TTS synthesizers; and validating the relevance of our working hypotheses for improving TTS synthesis.

### 5.2. Recording

12 tales of the text corpus were selected, and recorded in a studio by a professional speaker, resulting in a corpus of about 1 hour of speech. The speaker was an experienced professional, well acquainted to studio recording procedures. A fellow sound engineer assisted him. The speaker was informed of our goals, and was instructed to avoid excessive dramatization: his tasks was to read the tales as if the audience was his own children. He was also allowed to change small portions of texts, which were difficult to tell with sufficient fluency. To obtain optimal speech material, he told the tales using overdubbing recording techniques. They consist in stopping the recording, going backward, playback, and overwriting problematic segments.

Figure 2: Intonation Stylization of tale Little Red Riding Hood based on a model of tonal perception. Each vowel nucleus is associated to one or more tonal segment. Segment widths correspond to tone intensity (green plot) at segment boundaries.

Table 9 shows the differences between original text, and speaker's transcription, described in terms of the number of word deletion, word insertion, word modification (including synonym substitution), and word shift within sentences. The amount of observed differences was low (about 1%) and occurred only at the sentence level. It did not affect tale structure, neither sentence structure. Levin *et al.*, (1982) reported prosodic differences between storytelling and story reading speech. The recorded material of our corpus should rather be described as "studio story reading" speech. This methodological difference provides more control to the speaker on its performance. Consequently, we may hope this material to contain more controlled prosodic patterns, which would be more suited to the task of speech synthesis.

### 5.3. Lexical and Phonetic Alignment

Phonetic transcription, and phoneme alignment of the speech transcription were obtained using the LIMSI semi-automatic software (Adda-Decker & Lamel, 1999; Gauvain *et al.*, 2005). Textual annotations were aligned with the speech signal and stored in Praat TextGrids (Boersma, 2002).

An example of transcription alignment is given in figure 2. Pitch was obtained using Yin estimator (De Cheveigné & Kawahara, 2002), and stylized using (d'Alessandro & Mertens, 1995) model of tonal perception. Note that more than 20 semi-tones (about 2 octaves) for pitch variation were observed in this 6 second long example. These variations are considered as large, and highlight the variety and complexity of tale prosody. Note also that the typical tale opening formula "*Il était une fois*" ("once upon a time") exhibits a prototypical pitch pattern. A clear prominence can be seen for the lexical element "*la plus jolie*" (the prettiest), marked with an intensifier marker in the corpus.

## 5. Discussion and Future Work

In this paper, we presented a text and speech corpus designed for studying expressive Text-To-Speech and Gestural Synthesis. A focus was made on the description of our corpus-annotating scheme, in term of coverage and inter-annotator agreement. This shows the difficulty of the labelling task we defined. This task should ultimately be done automatically for tale synthesis.

Another goal of this article was to define metrics for evaluation of the future automatic annotation methods.

A preliminary prosodic analysis of the speech corpus (Doukhan *et al.*, 2011) showed that storytelling indeed induces important prosodic variations. Prosody exhibits more variations than those measured for e.g. the political address and radio news speaking styles.

Significant correlations were observed between prosodic properties and tale episodic structure (e.g.: higher pitch dynamics in tale exposition and triggering event, lower loudness and dynamics in epilogue). Agreement measures reported in section 3.2.3 showed that scenes segmentation and identification is not trivial, and may be subject to several interpretations. They also showed that a significant proportion of scene boundaries could be detected using automatons assigning a scene boundary to each paragraph mark associated to a narrator speech turn. Further analysis will compare pause durations between scenes and between paragraphs, to measure the impact of text presentation on the speaker performance.

Prosodic effects depending on the age and gender of the speaking characters were observed. These observations invite us to add the tale character meta-information proposed in section 3.6. to the whole corpus. While strong inter-annotator agreement was observed for speech turn identification (section 3.3.3.), it was significantly lower for the references to tale characters within text (section 3.5.). Difficulty to define tale character reference boundaries, and to define what a tale character is (e.g. 'object', 'sun'...) may explain these differences.

The first results of prosodic analyses also point out the need for more local annotations. To that aim, speaker mode and dialog acts were added at the sentence level, and intensifiers were used at the lexical level. Since those annotations were obtained automatically, they have the advantage of staying consistent over the whole corpus. Further prosodic analyses on the speech corpus will empirically confirm whether or not those new annotations are relevant for improving the speech quality.

An expressive prosodic generation prototype was implemented, in conjunction with the Acapela non-uniform-unit text-to-speech synthesizer, whose full description goes beyond the scope of this paper.

Current work is devoted to prosodic analysis, prosodic prediction, and perceptual evaluation of automatic tales synthesis.

## 6. Acknowledgements

This work has been funded by the French project GV-LEx (ANR-08-CORD-024 http://www.gvlex.com).

## 7. References

Acapela (2012). http://www.acapela-group. com/.

G. Adda, M. Adda-Decker, J.L. Gauvain, & L. Lamel (1997). Text normalization and speech recognition in French. *In EUROSPEECH*, pp. 2711–2714.

M. Adda-Decker & L. Lamel (1999). Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29(2-4):83–98.

J. Adell, A. Bonafonte, & D. Escudero (2005). Analy- sis of prosodic features towards modelling of emotional and pragmatic attributes of speech. *Procesamiento de Lenguaje Natural*, 35:277–284.

A. Allauzen & H. Bonneau-Maynard (2008). Training and evaluation of POS taggers on the French multitag corpus. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

C.O. Alm & R. Sproat (2005) Perceptions of emotions in expressive storytelling. *In Ninth European Conference on Speech Communication and Technology*.

A. Berger & S. Pesty (2005). Towards a conversational language for artificial agents in mixed community. In *Proceedings of the 4th Central AND Eastern European conference on Multi-Agent Systems (CEEMAS'05)*, Budapest, Hungary, September.

Y. Bestgen (2009). Quel indice pour mesurer l'efficacité en segmentation de textes. *Actes de TALN*, 9.

PPG Boersma (2002) Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10): 341–345.

F. Burkhardt & J. Stegmann (2009). Emotional speech synthesis: Applications, history and possible future. *Proc. ESSV*.

C. d'Alessandro & P. Mertens (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech & Language*, 9: 257–288.

A. De Cheveigné & H. Kawahara (2002). Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am*, 111(4): 1917–1930.

D. Doukhan, A. Rilliard, S. Rosset, M. Adda-Decker, & C. d'Alessandro (2011). Prosodic analysis of a corpus of tales. *In InterSpeech*, pp. 3129–3132.

O. Galibert (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. Ph.D. thesis, Université Paris Sud, Orsay.

J.L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, & H. Schwenk (2005). Where Are We in Transcribing French Broadcast News? *In Ninth European Conference on Speech Communication and Technology*. ISCA.

R. Gelin, C. d'Alessandro, Q.A Le, O. Deroo, D. Doukhan, J.C. Martin, C. Pelachaud, A. Rilliard, & S. Rosset (2010). Towards a storytelling humanoid robot. *In AAAI Fall Symposium Series on Dialog with Robots*, pp. 137–138.

J.M. Golden (1985). Interpreting a tale:: Three perspectives on text construction. *Poetics*, 14(6): 503–524.

A.J. Greimas (1966). *Sémantique structurale: recherche et méthode*. Larousse.

M.A. Hearst (1997). Texttiling: Segmenting text into multi- paragraph subtopic passages. *Computational linguistics*, 23(1): 33–64.

E. Klabbers & J. van Santen (2004). Clustering of foot-

based pitch contours in expressive speech. *In Proc. 5th ISCA Speech Synthesis Workshop*, pp. 73–78.

H. Levin, C.A. Schaffer, & C. Snow (1982). The prosodic and paralinguistic features of reading and telling stories. *Language and speech*, 25(1): 43.

E. Loper & S. Bird (2002). Nltk: The natural language toolkit. *In ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, volume 1, pp. 63–70.

C. Pelachaud, R. Gelin, J.C. Martin, & Q.A. Le (2010). Expressive gestures displayed by a humanoid robot during a storytelling application. *New Frontiers in Human-Robot Interaction (AISB)*.

L. Pevzner & M.A. Hearst (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1): 19–36.

V. Propp (1968 orig. 1928). *Morphology of the Folktale*. University of Texas Press.

J. Rivère, C. Adam, & S. Pesty (to appear). Langage de conversation multimodal pour agent conversationnel animé. *Technique et Sciences Informatiques*.

S. Rosset, O. Galibert, G. Bernard, E. Bilinski, & G. Adda (2009). The LIMSI multilingual, multitask QAst system. *In Proc. CLEF 2008*, pp. 480–487, Berlin, Heidelberg. Springer-Verlag.

H. Schmid (1994). Probabilistic part-of-speech tagging us- ing decision trees. *In NEMLP*, vol. 12, pp. 44–49.

M. Theune, K. Meijs, D. Heylen, & R. Ordelman (2006). Generating expressive speech for storytelling applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4): 1137–1144.

O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, & B.R. South (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*.

T.A. van Dijk (1982). Episodes as units of discourse analysis. *Analyzing discourse: Text and talk*, pp. 177–195.

```
<recit>

  <titre>
    <narr locid="0">
      <phrase phrid="1"><mwe type="gn"><pers
locid="1">trois petits morceaux de nuit </pers></
mwe>\n\n </phrase>
    </narr>
  </titre>

  <situation>
    <narr locid="0">
      <phrase phrid="2"><ne type="time">voilà
bien longtemps </ne>, <ne type="time">un soir de
printemps </ne>, <mwe type="gn"><pers
locid="1">trois petits morceaux de nuit </pers></
mwe>se détachèrent <ne type="loc">du ciel </
ne>et tombèrent <ne type="loc">sur Terre </ne>.
\n </phrase>
      <phrase phrid="3"><pers locid="1">ils </
pers>étaient <enum><mwe type="locadv">tout noirs
</mwe>, <mwe type="locadv">tout bossus </mwe>,
et <mwe type="locadv">tout étonnés </mwe></
enum>de se retrouver <ne type="loc">là </ne>. \n
\n </phrase>
      <phrase phrid="4"><pers locid="2">le
premier </pers>dit : \n </phrase>
    </narr>
    <spkr locid="2">
      <phrase phrid="5">- où suis <pers
locid="2">-je </pers>? \n </phrase>
    </spkr>
    <narr locid="0">

      <phrase phrid="6"><pers locid="3">le
deuxième </pers>dit : \n </phrase>
    </narr>
    <spkr locid="3">

      <phrase phrid="7">- qui suis <pers
locid="3">-je </pers>? \n </phrase>
    </spkr>
    <narr locid="0">

      <phrase phrid="8">et <pers locid="4">le
troisième </pers>piailla : \n </phrase>
    </narr>
    <spkr locid="4">

      <phrase phrid="9">- <pers locid="4">j' </
pers>ai faim ! \n\n </phrase>
    </spkr>
  </situation>

  <element-declencheur>
    <narr locid="0">
      <phrase phrid="10"><pers locid="5">une
souris </pers>passait <ne type="loc">par là </
ne>. \n </phrase>
    </narr>
    <spkr locid="5">
      <phrase phrid="12">hé mais ! \n </phrase>
    </spkr>
    <narr locid="0">
      <phrase phrid="13">se dit <pers locid="5">-
elle </pers>, </phrase>
    </narr>
    <spkr locid="5">
      <phrase phrid="17">il faut retrouver leurs
parents . \n\n </phrase>
    </spkr>
  </element-declencheur>

  <scene id="1">
    <narr locid="0">
      <phrase phrid="18">mais personne ,
apparemment , n' avait perdu de petits . \n\n </
phrase>
      <phrase phrid="22">allez donc élever des
enfants qui volent , quand vous êtes cloué <ne
type="loc">au sol </ne>! \n\n </phrase>
    </narr>
  </scene>

  <refrain id="3" refrid="1">
    <spkr locid="5">
      <phrase phrid="23">ah là là ! \n </phrase>
    </spkr>
    <narr locid="0">
      <phrase phrid="24">se dit <pers
locid="5"><mwe type="gn">dame Souris </mwe></
pers>, </phrase>
    </narr>
    <spkr locid="5">
      <phrase phrid="25">la vie est <mwe
type="locadv">bien compliquée </mwe>. \n </
phrase>
      <phrase phrid="26">mais il y a toujours
moyen de s' arranger . \n\n </phrase>
    </spkr>
  </refrain>

  <scene id="3">
    <narr locid="0">
      <phrase phrid="38">puis <ne type="time">un
beau soir </ne>, dans <mwe type="gn">l' air
tiédi </mwe>, <pers locid="1">ils </
pers><enum>frémirent et déplièrent leurs ailes </
enum>. \n </phrase>
      <phrase phrid="46">et à la queue leu leu ,
sans bruit , <pers locid="1">ils </pers>s'
envolèrent <ne type="loc">dans la nuit </ne>. \n
\n </phrase>
    </narr>
  </scene>

  <epilogue>
    <narr locid="0">
      <phrase phrid="47">peut-être ... peut-être
est -ce ainsi que sont nées les chauves-souris ,
<mwe type="gn">ces hirondelles <ne type="loc">de
la nuit </ne></mwe>? \n </phrase>
      <phrase phrid="48">et même si l' histoire
n' est pas vraie , parions que les chauves-
souris l' aimeraient - si elles savaient lire !
\n </phrase>
    </narr>
  </epilogue>
</recit>
```

Figure 1: A simplified tale annotation example. Markers title, exposition, triggering event, scene, refrain and epilogue correspond to the first structural level defined in section 3.2. *narr*, *spkr* and phrase markers are described in section 3.3. *pers* markers are described in section 3.5.