SynPaFlex Corpus Annotation Manual

Aghilas Sini[1], Damien Lolive[1], Élisabeth Delais-roussarie [2], Marie Tahon[1], and Gaëlle Vidal[1]

[1]IRISA, University of Rennes 1 (ENSSAT), Lannion, France
[2]UMR 7110-LLF & Université Paris-Diderot, France

October 4, 2019

Contents

1 Corpus files

1.1 Audio files

Global particularities of the speech audio files and their texts were written, for each book, in a corpus description file, including an estimation from the text of amount of acted direct speech, and references to manually produced information.

1.1.1 Global description

The SynPaFlex corpus is comprised of dozens of hours of speech recordings found on the free domain server Librivox. Files formatted with the best quality, mp3 128-kbps, can be download at: archive.org. It is monospeaker, performed by a female reader who reads French classic literature (written in the 19th century for most of them). Each file corresponds to the recording of a book chapter, or of a small literary work (poem, tale, fable, ...). The speaker can personify her voice for some characters in direct style. Only files of good audio quality were selected[1].
Table 1 proposes a literary genre classification and shows, for each genre, duration for whole and best audio quality speech as well as for different manual processing. Chapters distributed over five historic novels constitute the majority of the corpus: 70% of the 87 hours. fantastic and adventure ones represent 15% and 10% (13 and 9 hours) of the corpus. A group of tales from different authors, and a few chapters of a symbolist short story concern for each one to two hours of recording. The rest of the speech corresponds to few minutes readings of other original literary works: an epistolary novel, a fantastic epic poetic prose, a fantastic dramatic, a pamphlet, and also verse forms of fables and poems.

SynPaflex corpus French mono-speaker	Whole corpus audio-files duration		Manual annotation subcorpora audio-files duration		
Genres	Whole corpus	Best Quality	Characters	Emotion and Prosody	Phonetic segmentation validation
Historic novels	61h 18m	57h 21m	24h 43m	8h 41m	20m
Fantastic novels and short stories	12h 55m	10h 08m	1h 08	1h 08m	-
Adventure novels and short stories	9h 01m	3h 37m	3h 37m	3h 37m	20m
Symbolist short story	1h 42m	1h 18m	1h 42m		-
Tales	1h 06m	1h 06m	1h 06m	10m	-
Epistolary novel	20m	20m	20m	-	-
Fables	18m	13m	18m	-	1m
Fantastic epic	18m	11m	18m	-	-
Fantastic dramatic	11m	10m	10m	-	3m
Poems	12m	12m	12m	-	1m
Pamphlet	6m	-	6m	-	-
TOTAL	87h 28m	74h 35m	33h 34m	13h 36m	47m

Table 1: Presentation of the audio literary corpus and manual annotation

[1] Unselected recordings remain available, for example the 12 hours novel *Les temps modernes* (link).

Genre	Title, author (full reading *)	Duration	Lower audio quality	Abbr.
Historic novels	*Les Mystères de Paris* vol.1 and 2,, Eugène Sue *	25h 32m	22m	HNmy
	Les misérables, Victor Hugo	14h 16m	2h 18m	HNmi
	Madame Bovary, Gustave Flaubert *	13h 12m	26m	HNma
	Le roman de la momie, Théophile Gautier *	6h 33m	-	HNro
	Germinal, Émile Zola	1h 46m	50m	HNge
Fantastic novels and short stories	*La vampire*, Paul Féval *	10h 02m	2h 26m	FNva
	Voyage au centre de la terre, Jules Verne	1h 52m	22m	FNvo
	La Vénus d'Ille, Prosper Mérimée *	1h 02m	-	FSve
Adventure novel and short stories	*La fille du pirate*, Maurice Chevalier *	6h 43m	4h 31m	ANfi
	Carmen, Prosper Mérimée *	2h 18m	53m	ASca
Symbolism short stories	*Contes cruels*, Auguste Villiers de l'Isle-Adam	1h 42m	24m	SYco
Tales	*La malle volante*, Andersen *	12m	-	TAan
	Le monstre Yatama, Claudius Ferrand *	8m	-	TAfx
	Les sept chevreaux, Claudius Ferrand *	16m	-	TAfy
	Ourashima Taro et la déesse de l'Océan, Claudius Ferrand *	16m	-	TAfz
	La Hyène, l'Hippopotame et l'Éléphant, Franz de Zeltner *	11m	-	TAzx
	L'histoire de Koli, Franz de Zeltner *	4m	-	TAzy
Epistolary Novel	*Lettres persanes*, Montesquieu	20m	-	ENle
Fables	*Fables de La Fontaine*	18m	5m	FAfo
Fantastic epic	*Les chants de Maldoror*, Comte de Lautréamont	18m	7m	FEch
Fantastic dramatic	*Infernaliana*, Charles Nodier	11m	-	FDin
Poems	*L'albatros*, Charles Baudelaire *	1m	-	POal
	Chanson d'automne, Paul Verlaine *	1m	-	POch
	Le Dormeur du val, Arthur Rimbaud *	1m	-	POdo
	Fiez vous y !, Charles d'Orléans *	1m	1m	POfi
	Gaudriole en six couplets, unknown *	3m	-	POga
	Un matin, Emile Verhaeren *	1m	-	POma
	Perles, Jean Courdil *	1m	-	POpe
	La veuve indienne, Eugène Fouques *	4m	-	POve
Pamphlet	*Le cerf-volant aux six têtes*, Guillaume Taillerand-Perigord *	6m	6m	PAce

Table 2: Literary works readings and their audio quality

1.1.2 Later changes made to the audio files

The following changes have been made to the audio files after their collection and their manual annotation.

- mp3 to wav conversion

  As a preparation before automatic processing mp3 were converted to wav and produced slight changes of duration. By the fact, manual annotations, as they had been applied on mp3, are slightly desynchronized from wav files. A 20ms to 50ms desynchronization has been noticed on a few files, varying between the files. The corpus is $87h27m34s$ long for mp3. Further examinations should be done about the audio conversion, and a re-adaptation could have to be done on manual annotations and wav files.

- loudness harmonization

  As recordings were performed in different technical and environmental conditions, loudness has been harmonized using the *FreeLCS* tool[2], offsetting a 1.6 LUFS standard deviation. Despite of that, audio data acoustic features remain more or less heterogeneous.

- new naming of the files

  The original Librivox audio files names had not been changed during collecting and annotation processes, except when the recordings of whole books were performed in a collaborative way (the prefix "NEB-" was added).

  Later on, audio file names have been changed moving author names from the end to the beginning of the file names, and then making them end with their numbers (generally chapters).

1.2 Text files

Once audio data have been selected, the corresponding texts have been collected, and a few manual operations have been applied to simplify further processing:

Textual contents were gathered in text files either for each book (then the name of the corresponding audio file was written as a comment at the beginning of each chapter), or for each chapter (the text file name is the audio file name).

A global textual conformation to the speech was done by inserting transcriptions of introductions and conclusions the speaker had added in the recording, and by placing between check marks footnotes and end-of-book notes where they appear in the reading stream. A few unusual forms can appear in poems (e.g. [encor] for [encore]) False reading are thought to be extremely occasional (e.g. "répondit" said "répondissait", in *Carmen*)

As no transcription convention had been defined, no change was made to the original structures and typography for each book, leading to variations that were reported in the corpus description file (see table).

- Structure
  - carriage return for each paragraph (except for poems and fables)
  - no regular justification for empty lines.

[2] http://freelcs.sourceforge.net/

- Typography
    - Special typefaces: [*], [n°], [œ] (to complete)
    - Variations between the books
        - Character encoding: either [oe] or [œ] [. . .] or [...] (to complete)
        - Abbreviations: [Dr], ... (*HNmy*)
        - Proper names: can be between [_ _] (in *FEch* and *ANfi*, e.g. [_L'évasion_]), or resume to their initial capital (e.g. *SYco*)
        - Common names: some expressions are in capital letters (*ANfi*), or in *poems* and *fables* each first word of a verse has an initial capital letter (*FAfo*)
        - Freedom with punctuation (e.g. [! ! !] (*HNmy*); appositive forms between the same dashes as those used for direct style (*ANfi*). In the poem *FEch* some verses begin with a tabulation (indented lines)
        - Expressions: between either French quotation marks, or underscores. In books where French quotation marks are also used for direct style and focus words (*HNmy*, *HNmi*, *ANfi*, *HNro*, *ASca*), only a double final punctuation distinguishes the situations.
        - Direct style punctuation: textual direct style is indicated by various but identified typographic punctuation characters. Generally a direct style is announced at the beginning of the line either with [--] or [-] or with three kinds of [–], or with French quotation mark opened guillemot. But the can appear in the paragraph stream (e.g. *SYco*, *fables*, *poems*). 2nd level of direct style line begins with a space (*HNmy*, *FNva*, *ANfi*). But some discourse style determinations remain be ambiguous in some cases:
        - A few exceptions are observed in *HNmi*, where some text was read in direct style (characters or personification) but without any written direct style punctuation (e.g. the narrator's voice impersonates the voice of the crowd).
        - Any book written in the first person: any other book than historic works, except the pirate novel *ANfi*, tales and some poems (however, four *ANfi* chapters are written in an epistolary style - audio files 101 111 113 115).
        - Dialogue tags,
        - Reports in a paragraph stream (then without carriage return): during the narration or even during direct style, when a character become the narrator of his own story (e.g. in *ANfi* where punctuation is complex and would need to be validated).

          However in these two last cases a strong punctuation before the closing quotation indicates a high probability for direct style.

1.3 Manual annotations

Manual annotations were applied on some original mp3 files using WaveSurfer[3] Software in HTK format, giving synchronized informations about audio quality, characters corresponding to the read direct style, prosody, and emotion. Also some punctual events were reported (unexpected phonology or noise, change of language).

The annotation method had first been defined on a small subset of readings, and then tested on audiobook recordings completed by other readers. It was found to be generic enough to render a global perceptive description of the speech. As shown in description table 1 , 38% of the whole corpus have been processed manually to provide characters annotation, and 15% - included in those 38% - to describe emotional and prosodic patterns contents.

Some "2DO" labels can appear in a very few number of files which have not been entirely annotated.

1.3.1 Characters annotation process

Character annotation files have a ".perslab" extension.

The speaker, who is the same for the whole corpus, can personify the different characters of the book by changing her voice or her way of speaking. The character tags were identified from the text and turns of speech have been labeled according to the following annotation scheme:

- CHARACTER ID: indicates the talking character according to the text, or otherwise if it is the narrator.
    - either [_] (underscore): narrator's voice, which has the natural timbre of the speaker
    - or a character's identifier which is, depending on the book:
        - either directly the character's name
        - or the prefix [pers] followed by a number [x], referring to a character referenced in the corpus-description file.
- VOCAL PERSONALITY ID: indicates the talking character according to the vocal personality. Indeed, even if the speaker is very talented and coherent along the books, she can do mistakes or forget to change her voice when acting a character who should be acted in his personified voice. Therefore, for such speech intervals, the timbre remains the own speaker's timbre or corresponds to another character.

  structure: separator [/t] followed by the voice identifier.

  Example: If ever the timbre of a character y is used for the character x, the label is pers[x]/t[y].
- Convention: The number 0 designs the narrator, then [pers0] = the narrator speaking in direct style, and pers[x]/t0 = the character x is speaking with the natural voice of the narrator.
- Note: Dialogue tags were reported as parts of the narrator's speech.

[3] http://www.speech.kth.se/wavesurfer/

Each VOCAL PERSONALITY is summarily described with meta-data gathered in the corpus description file, including the name of the character and the description of his timbre as shown in table 3: 4 letters are used for name, age, gender, prosody and timbre features, each one can be [I] or [?] if undetermined, or [O] for the natural voice of the speaker. A capital letter is the equivalent of the non-capital one but stronger.

letter 1	letter 2	letter 3	letter 4
age	genre	perceived f0 range	main characteristic of the timbre
A: adult E: young V: old	F: woman H: man	A: high G: low	C: veiled D: soft E: muffled I: internalized K: broken M: melodic P: precious R: husky, rough, hoarse S: dry timbre T: tense (authoritative) U: murmur V: vulgar W: week Z: nasalized

Table 3: Timbre classification

1.3.2 Prosody annotation process

After considering the whole speech data, eight prosodic descriptors were defined, then encoded and assigned to audio tracks corresponding to chapters. Prosody annotation files have a ".proslab" extension.

As far as possible, labels were assigned according to the perceived prosody, without taking into account the linguistic content. They characterize units which could range in length from a word to several sentences. Seven of them correspond to speech showing the following types of prosodic patterns: QUESTION (interrogative), NOTE, NUANCE, SUSPENSE, RESOLUTION (authority, or imperative), SINGING, and IDLE (no particular prosodic pattern, or declarative). The eighth label, EMOTION, was used to report - but without describing it - the presence of any perceived emotional content.

Let's notice as of now that the tag EXCLAMATION is not listed above. This is because this information can be simply deduced from another level of description: in this corpus, the *Exclamation* pattern was found strictly correlated with the emotional content of *surprise*, which is reported in the emotion labeling level (presented in Section 1.3.3). Manual annotation is costly in time and redundancy is not desirable in its process. In the following analysis of the prosodic manual labeling, each emotion labels *surprise* will therefore be assimilated to an hidden prosodic labels EXCLAMATION.

Another important point is that, when needed for a more precise description, labels were combined (e.g. Emotion+Question+Nuance), using the separator "/". Hence, simply summing the label durations for each type of prosodic pattern gives a value which exceeds the duration of the sub-corpus.

Among the prosodic parameters, the perceived pitch-curve during voice production takes an important role in assigning the labels. For instance, the NUANCE pattern, which is one of the reading strategy of the speaker, maintains listener's attention. This pattern is characterized melodically by a high pitch at the beginning, then a decrease with modulations, and finally a slight increase when it doesn't end the sentence (see Figure 1).

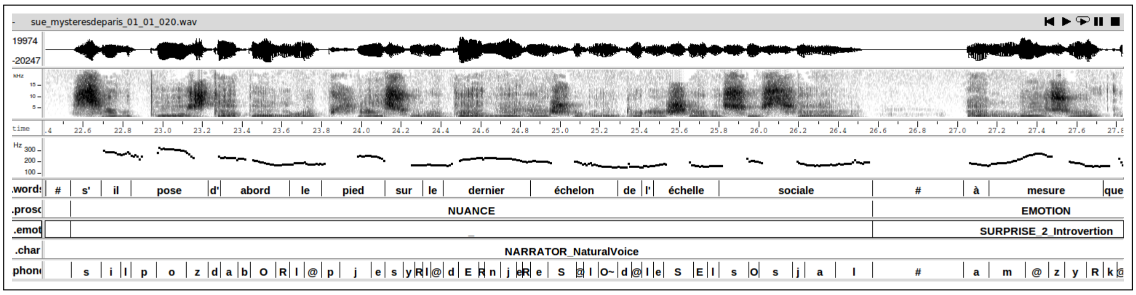
sue_mysteresdeparis_01_01_020.wav
19974
-20247
kHz
15
10
5
time
.4
22.6
22.8
23.0
23.2
23.4
23.6
23.8
24.0
24.2
24.4
24.6
24.8
25.0
25.2
25.4
25.6
25.8
26.0
26.2
26.4
26.6
26.8
27.0
27.2
27.4
27.6
27.8
Hz
300
200
100
.word:
# s' il pose d' abord le pied sur le dernier échelon de l' échelle sociale # à mesure que
.prosc
NUANCE
EMOTION
.emot
_
SURPRISE_2_Introvertion
.char
NARRATOR_NaturalVoice
phon
s i l p o z d a b O R l @ p j e s y R l @ d E R n j eR e S @ l O~ d @ l e S E l s O s j a l # a m @ z y R k @

Figure 1: Manual annotations - NUANCE prosodic pattern example

1.3.3 Emotion annotation process

Emolab annotation files have a ".emolab" extension.

Different theoretical backgrounds are classically used to identify emotional states, principally based on either distinct emotion categories or affective dimensions [3]. Usually, choosing the emotion categories and their number, or the emotion dimensions is an issue.

In the present study, the basic scheme used to manually encode emotions has three items:

- *Emotion category*: Six categories are available, those selected by the Basic Emotions theory [4]: SADNESS, ANGER, FEAR, HAPPINESS, SURPRISE, DISGUST. Two other categories were added to better represent the content of the different books: IRONY and THREAT.

- *Intensity level*: This item, on a scale from 1 to 3, is meant to give a measurement of the experienced emotion intensity according to the speech. For instance, one can interpret its values as follows: SLIGHTLY ANGRY (1), ANGRY (2) , and STRONGLY ANGRY (3).

- *Introversion/Extroversion*: This binary item reflects the way the emotion is rendered through the speech (discreetly, prudently / obtrusively, ostentatiously)

The second and third items may have strong correlations with some of the widely used affective dimensions, as activation and arousal.

Furthermore, an important feature of the manual emotion annotation used for the corpus is that the three items labels can be mixed together (using the separator "/") to provide a more precise description of the perceived emotion. For instance, speech can continuously convey strong and very expressive SADNESS as well as FEAR through some words, which could be tagged as [sadness-3-E + fear-1-E].

1.3.4 Other events

Besides acoustic indications of loud noises or music, different unexpected speech events were also reported:

– Linguistic events: for example, the use of foreign languages, e.g.: *PAce* introduction and conclusion in English ; *POga* introduction and conclusion in French and in English ; *AScа* - foreign words (Hispanic, Rom, ...)

– Phonetic events which are not written in the text: phoneme substitutions, elisions and insertions, high elongations, breaks and pauses, specific voice quality (e.g. whispered voice). These features can be of high interest for rendering a more human synthetic voice [2].

Phenomenon annotation files have a ".phlab" extension. The list of labels are presented in table 4. Labels can be combined using the separator "/", to describe a same interval of speech.

.phlab label	description
phI	phonetic insertion (phItxt : insertion written in the text)
phS	phonetic substitution (phStxt : substitution written in the text)
phE	phonetic deletion (phEtxt : deletion written in the text)
phB	noise
phCe	caesura (short breaks of silence) in the speech, between words (e.g. focus strategy) or inside words (e.g. emotional effect)
phCh	whispered or partially voiced speech *Note: this label is not written when it is a character feature (then reported in the character manual annotation).* *Note: A more complete or precise notation would have needed a special label to report breath in the speech.*
PhD	lengthening (elongation) of a phonetic unit
phL	foreign language
pv	Paraverbal (pvtxt : Paraverbal written in the text) either absolutely non-verbal or sever deformation of a word (e.g. : repetition of a syllable)
F0	pitch event *Note: this pitch notation is not exhaustive at all, it was more or less abandoned during the annotation*
Musique	music

Table 4: Manual annotation labels used to describe unexpected speech events

1.4 Manual validation of the automatic phonetic segmentation

A few validations of the automatic phonetic segmentation were done using the Praat [1] Software.

2 Analysis

The manual data-sets could provide valuable guidance for experimentation, especially by linking with linguistic information, acoustic measurements, and other descriptions. Examining how manual labels are distributed among literary genres can also be of great interest.

2.1 Estimation of amounts of narration and direct style for each book reading

A first estimation of proportions of direct style by book has been calculated from the typographic indications. It is an approximation because ignoring dialogue tags (then included into direct style) and direct styles appearing in a paragraph without beginning a new line, but results points reliable enough estimations to mark the tremendous differences, between the books reading data, for proportion of narration and direct speech, as shown on figure 2.

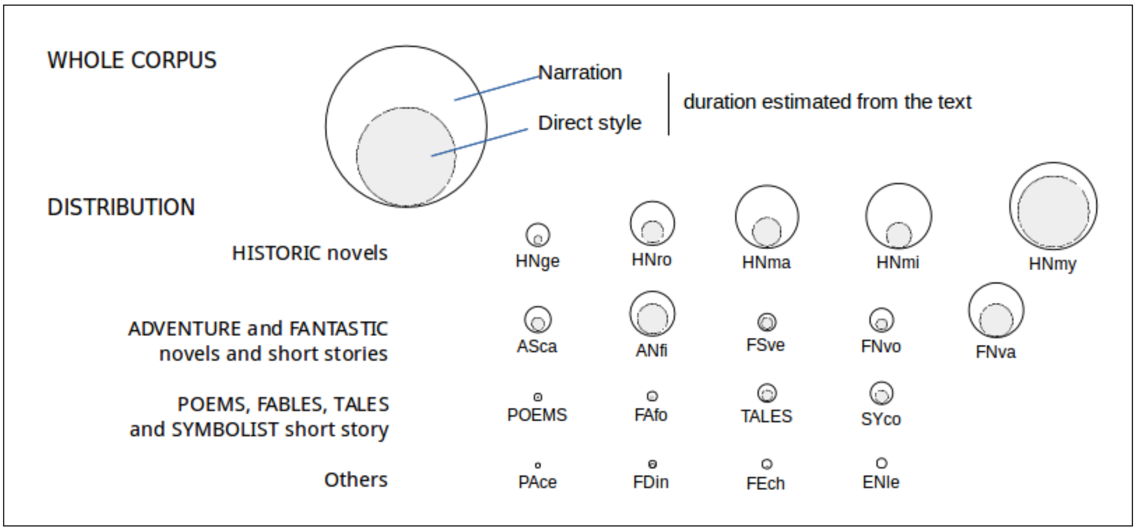

Figure 2: Duration representation for each literary work reading according to estimated style of discourse

Mystères de Paris, which is one third of the whole corpus, is mainly direct speech (67%). Among the other novels is found big proportion of direct style in *La fille du pirate* (45 %), La Venus d'Ille (41 %), *La vampire* (36 %), *Le roman de la momie* (30 %). The symbolism short stories *Contes cruels* also show around 30 % of direct style, the other main texts go from 13% (*Germinal*) to 22 % (*Carmen*). We must separate *Infernaliana*, *Lettres persanes* and *Les chants de Maldoror*, because the direct style is performed for the narration.

As far as small pieces are concerned, tales can have almost half of direct style, the mean for the fables is 24 %, and poems can have some direct style too.

Other indications for discourse style are given by the books characters manual labeling of the speech (see section 1.3.1), which includes a particular tag for direct style that is be the author tag (here called the narrator tag). Its reading has the particularity of having necessarily been acted with the natural voice of the speaker.

2.2 Narration speech subjective classification

According to perception, the speaker adopts specific prosodic attitudes to perform narration, depending on the book, and sounding constant through each literary work or book. By listening to narration, degrees can be assigned for the two following parameters: narrative stylization, and more natural speech (leading to a narration that may be referred to as direct speech or characters acting).

- Narration prosodic modulation: Recurrent prosodic patterns into every few to several word, appear in intonation and cadence, creating a propitious atmosphere for listening attentively and for a long time. They are anchored into limits of syntactic intervals, but accentuation don't necessarily reflects semantic focus. Those particular figures were annotated manually as described in section 1.3.2). Expressiveness is the narrator's one, acted by the speaker.
- Natural expressiveness: Also narration can be sounding close to spontaneous speech. Expressiveness is the speaker's one.

Illustration 3 is a global mapping of narration audio data, with duration estimated from the text as specified section 2.1. Considering narration style was constant during each literary work reading, only a few minutes of each book audio data was examined, the degrees of expressiveness were assigned on the two axes narration stylization and natural speaking.

We can see that narration expressiveness vary in accordance either with degrees of lively natural, or with narration stylization, except *Les mystères de Paris* and the panflet, which have high scores for both. Historic novel narration prosody is stylized, except in the case of *Les misérables*, which is rather on spontaneous side. Fantastic and, more, Adventure novels and short stories narration prosody is little stylized but, as tales, epistolary novels and most of poems, are naturally expressive. Narration is very slightly expressive during the historic novel *Germinal* chapters recordings, same goes for the symbolism short story CONTES CRUELS and the *fables*.

2.3 Characters annotation analysis

The characters manual labeling was applied on more than one third of the whole corpus (33h39m) mined from 18 different books.

Results from the annotated sub-corpus indicate that one third of the speech is in direct speech style. The average duration for speech turns being of 7s, against 29s for the narrator. In some chapters, direct speech segments can also be very long, typically when a character becomes a narrator who tells his own story. 370 characters were identified, and the full data of their vocal personality labeling indicates a not negligible amount of prosody and vocal tone personification. Covering a wide range and typology, the speaker's voice is thus more or less radically far from her natural style (males, children and elderly people embodiments, psychological singularization, imaginary figures). These vocal personality changes often happen: around 20% of the whole speech (included narration) is concerned and, for two thirds, in stark contrast with the natural speaker's voice.

Figure 4 is a visualization of voice personification in direct style according to the literary works, for the whole SynPaflex corpus, estimated from the characters annotated sub-corpus data. Doted lines indicate which data is estimation or direct count when the whole book had been annotated.

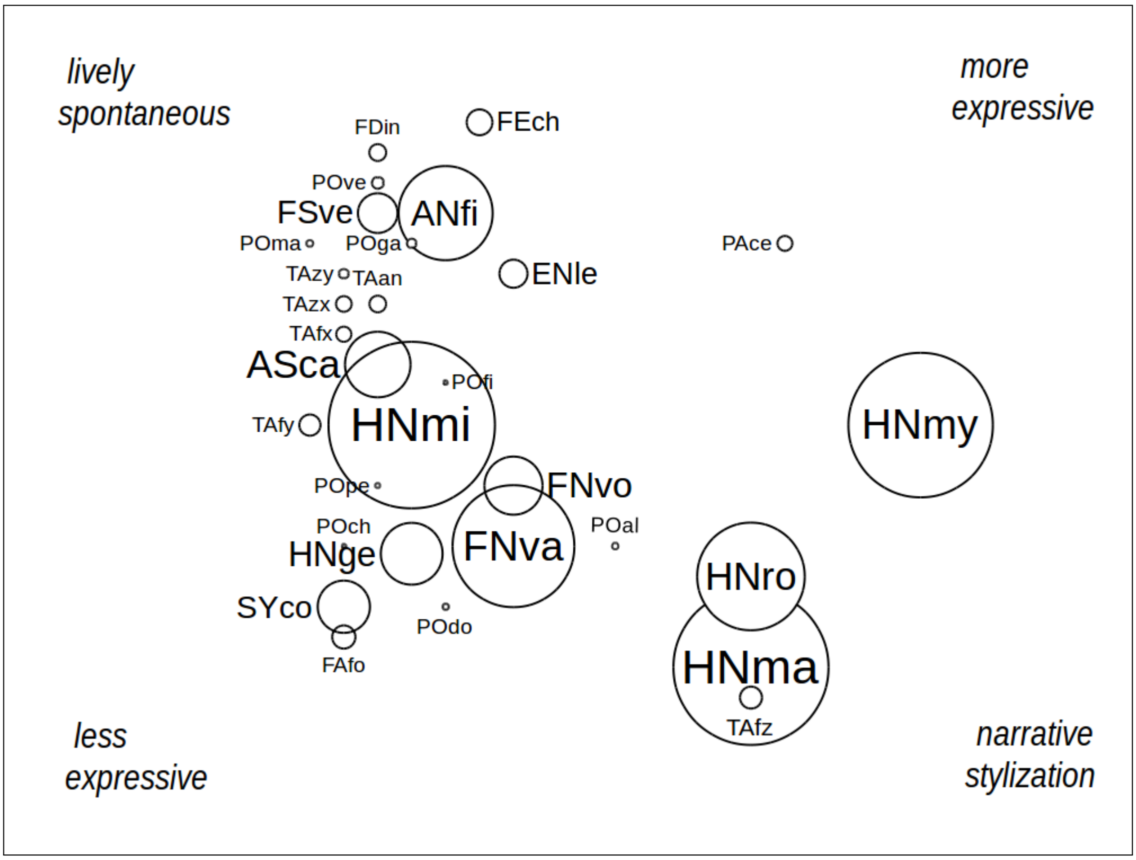

Figure 3: Narration data mapping for each book, according to degrees and types of expressiveness reported from listening (disks are superimposed but without any common data)

This simulation shows changes of voice personality in direct style most of the time (60 %) , particularly by impersonating men (30 %). The following main represented data concerned slight changes, then impersonation in other women. Young and old people have representation as well, but more often the voice changes according to other particularities.

2.4 Prosody annotation analysis

The prosody manual labeling described in section 1.3.2 was applied on 13h25 sub-corpus included in the character manual sub-corpus, and concerns 8 different books. Table 5 shows total duration for each label.

This analysis doesn't yet benefit from a linking with the character manual labeling, that could enable to interpret prosodic label as assigned to either narrator or characters speech. It should be done in priority, because it would lighten and precise any other observation made on the speech data files (e.g. Narration style and voice personifications analysis illustrated figures 3 and 4.

However the following comments can already be done.

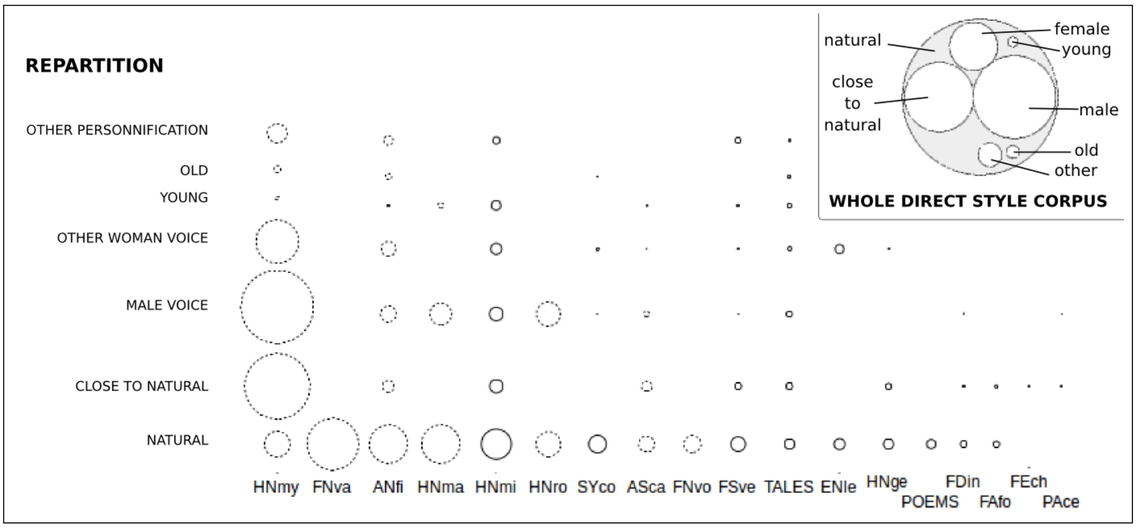

Figure 4: Voice personification in direct style - estimation from the characters annotated sub-corpus data

Prosodic label in the 13h 25 sub-corpus	EXCLA-MATION (hidden label)	IDLE	NUANCE	RESO-LUTION	SUSPENSE	QUESTION	NOTE
Duration	4h 42m	4h 21m	3h 58m	45m	41m	38m	39m
Sub-corpus %	34.8 %	32.2 %	29.5 %	5.6 %	5.1 %	4.7 %	4.8 %

Table 5: Total duration of manual prosodic labels (including combinations)

A non-IDLE prosodic tag has been assigned to 68% of the speech. As shown in Table 5, the hidden EXCLAMATION tag is very largely represented (more than 4h42), before the IDLE one (4h21m). The first particular prosodic pattern that comes after is NUANCE (3h58m), then come all the other prosodic patterns that are relatively well represented and evenly distributed (around 40m): RESOLUTION, SUSPENSE, QUESTION and NOTE. SINGING was found to be exceptional and is not reported here.

More than a half of the speech showing particular prosodic figures is described with combined labels, pointing out where prosody may be more complex.

Most of all, it was found that the EXCLAMATION pattern happens very frequently.It will be of high interest to know if it is rather characteristic of the narration or direct style and could for example be associated to such style of narration of some book (as analyzed section 2.4), or of some characters (as analyzed section 2.5). But in any case, it can be deduced that EXCLAMATION is an inherent part of the speaker's prosodic style.

The generic EMOTION prosodic indicator is assigned to 39% of the whole sub-corpus (5h18m), showing a large amount of emotional data, then mainly under *surprise* (EXCLAMATION). Its manual description is presented in Section 1.3.3.

Emotion	Effects on the first syllable of focus word(s)	Pitch median	Pitch curve	Rate	Loudness	Timbre changes
SURPRISE	accentuation	high				
SADNESS	disappearance	low	flat	slow	low	breath during the speech
JOY		according to joy type	flat (suave joy)	according to joy type	loud (intense joy)	breath during the speech (suave joy)
ANGER	accentuation	low	flat or top-down	fast		
DISGUST	accentuation	low	flat or top-down	fast on focus words	low	yes
FEAR		low	flat	varying with fear intensity		yes

Table 6: Examples of perceived impacts of emotion on the speech

2.5 Emotion annotation analysis

Manual emotion labeling was done on the same sub-corpus as for prosody (13h25m). Duration of tagged speech for each category of emotion is given in Table 8, and the number and average duration of labels are indicated in Table 7.

Emotion	Idle	Anger	Joy	Sadness	Fear	Surprise	Disgust	Other	Comb.	Total
# Seg.manual	3 364	147	115	295	76	2 895	47	23	1 699	8751
Avg. dur (s)	8.76	2.62	2.99	2.67	2.20	3.83	2.26	2.30	3.45	5.55

Table 7: Number and duration of emotion manual segments. Other includes IRONY and THREAT.

Emotion	IDLE	SURPRISE	SADNESS	JOY	ANGER	DISGUST	FEAR	IRONY	THREAT
Duration	8h 11m	4h 42m	44m	32m	31m	15m	11m	10m	3m
Subcorpus (%)	61.0 %	34.8 %	5.4 %	3.9 %	3.9 %	1.9 %	1.3 %	1.2 %	0.4 %

Table 8: Total duration of emotion manual segments (including combinations)

As for the prosodic manual labeling, this analysis doesn't yet benefit from a linking with the character manual labeling, and it should be done in priority for the same reason as explained in section 2.4.

A large amount of emotional content was reported (39% of the speech, including 13% with combined tags). But we can see as well how much SURPRISE label takes part in the account 35 %.

Significant observations have emerged during the annotation. A challenging one is that two radically different types of JOY can be conveyed by the speech, whereas none of the three items could take over their differentiation: on the one hand suave joy, and on other hand elation or gladness. Also, it is suggested that labels should be interpreted in context, notably in conjunction with the discourse mode. In particular, the expressive strategy implemented in the corpus narration is very specific, conveying almost continuously positive valence but in a subtle way, through pitch modulation and with focus words. The SURPRISE label was widely assigned to those recurrent patterns showing (i) a sudden pitch shifting upwards (ii) at least one accentuation onto the first syllable of a focus word (iii) a phonetic elongation or a short silence before this first syllable. Thus, as introduced in section 1.3.2, SURPRISE describes a recurrent emotional attitude of the reader, attracting the listener attention by regularly emphasizing the text.

Other types of variation occur when the speech conveys emotion, some examples are gathered in table 6.

References

[1] Paul Boersma and David Weenink. Praat: Doing phonetics by computer.[computer program]. version 6.0. 19, 2016.

[2] Nick Campbell. Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1171–1178, 2006.

[3] Alan S. Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017.

[4] Paul Ekman. Basic emotions. In Dalgleish T. and Power M., editors, *Handbook of Cognition and Emotion, 1999*, 1999.