Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6514–6522
Marseille, 11–16 May 2020

A Manually Annotated Resource for the Investigation of Nasal Grunts

Aurélie Chlébowski, Nicolas Ballier
Université de Paris
CLILLAC-ARP, F-75013 Paris, France
aurelie.chlebowski@hotmail.fr, nicolas.ballier@u-paris.fr

Abstract

This paper presents an annotation framework for nasal grunts of the whole French CID corpus (Bertrand et al., 2008). The acoustic components under scrutiny are justified and the annotation guidelines are described. We carefully characterise the acoustic cues and visual cues followed by the annotator, especially for non-modal phonation types. The conventions followed for the annotation of interactional and positional properties of grunts are explained. The resulting datasets after data extraction with Praat scripts (Boersma and Weenink, 2019) are analysed with R (R Core Team, 2017), focusing on duration. We analyse the effect of non-modal phonation (especially ingressive phonation) on duration and discuss a specialisation of grunts observed in the CID for grunts with ingressive phonation. The more general aim of this research is to establish putative core and additive properties of grunts and a tentative typology of grunts in spoken interactions.

Keywords: grunts, paralinguistic features, disfluencies, manual annotation

1. Introduction

Although they might be found as entries in dictionaries, "*non-lexical conversational sounds*" (Ward, 2006), also known as *interjections*, have for a long time been deemed to be negligible and were discarded from linguistics analysis. The idea that those linguistic phenomena might in fact convey some information about the speaker arose later in various domains such as discourse analysis. Nonetheless, for their linguistic status is still debated today (Tottie, 2019), there is no consensus as to the way those sounds should be transcribed or analysed.

Non-lexical conversational sounds being *sounds* in the first place, Chlébowski and Ballier (2015) hypothesised that part of the information conveyed by *non-lexical conversational sounds* must be related to their acoustic form. Their semasiological approach contrast with previous work considered as onomasiological (*e.g.* Delomier, 1999; Clark and Tree, 2002; Tottie, 2019), in the sense that they wished to characterise the various acoustic signs under scrutiny to possibly capture and document putative meanings for some types of grunts. Chlébowski and Ballier (2015) proposed to follow the "compositional model" established by Ward (2006) and analysed a sub-category of *non-lexical conversational sounds* that they called "*nasal grunts*". Despite being preliminary, their study revealed the need for a homogenisation of the transcription of *nasal grunts* and indicated that a bottom-up analysis of such sounds was required in order to 1) elaborate a sound-meaning mapping of their acoustic components and 2) study their behaviour in interactional situation.

In that respect, this paper proposes to focus on the acoustic analysis of *nasal grunts* in the Corpus of Interactional Data (hereafter CID; Bertrand et al., 2008). We propose a procedure for the annotation of the acoustic categorisation of *nasal grunts* based on observable features, in order to study their distribution in interactional context.

The remainder of the paper is organised as follows: section 2 lists the prerequisites for our analysis, defines grunts and the annotation procedure. Section 3 details the resulting datasets from this annotation of the TextGrids. Section 4 discusses the robustness of the acoustic cues used in the analysis of the various acoustic components of grunts and presents our multi-layered representation of grunts. Section 5 details our procedure for the acoustic analysis. Section 6 presents the guidelines for the investigation of grunts in interaction. Section 7 presents some of our results. Section 8 discusses the endeavour and concludes.

2. *Nasal Grunts*: Definition and Prerequisites

This section presents our definition of "nasal grunts", how we retrieved them from the orthographic tokens of the CID corpus and explain why we need manual annotation for the features we want to investigate.

2.1 Definition of the *Nasal Grunt* Category

Non-lexical conversational sound is a too wide category of non-lexical items to be analysed altogether. Chlébowski and Ballier (2015) established a sub-category of *non-lexical conversation sounds* based on a distinctive acoustic feature: nasality. They defined *nasal grunts* as being "words which have no "clear meaning" (Ward, 2000) but possess a nasal feature" (p.54). Orthographic representation of tokens that might fit their *nasal grunt* category would be : *hein, mh, hm, hum, han*, *uh, huh, erm, hun, ehm*… with the proviso that such an inventory is but a very small sample of the tokens that stands for *nasal grunts* in varieties of English and French.

2.2 The Need for an "Integrative" Analysis

Ward (2006) showed the significance of "considering sound, meaning, and function together" when it comes to the analysis of *non-lexical conversational sounds*. Chlébowski and Ballier (2015) proposed to investigate the "sound part" of *nasal grunts* of the *Newcastle Electronic Corpus of Tyneside English* corpus (hereafter NECTE; Corrigan et al., 2001). They performed an acoustic analysis on grunts from the Phonological Variation and Change in Contemporary Spoken English project (hereafter PVC; Milroy, Milroy and Docherty, 1997); the sound was investigated in all its dimensions (*i.e.* prosodic features, morpho-phonological components and paralinguistic features). The acoustic analysis revealed a set of acoustic components, namely: segmental features, syllable structure and syllable division, variations in amplitude, variations of the f0 curve, shift in register, variations in phonation types, variation of length and insertion of glottal stop. The

analysis of the distribution of grunts as regard interactions was not considered.

2.3 The Need for a Homogenisation of Transcriptions and Annotations of *Nasal Grunts*

The study of the grunts from the NECTE corpus gave us hints as to what acoustic components should be analysed. Nevertheless, this study was restricted to a perceptual analysis of the acoustic components of grunts, which complexifies replication for other annotators. The need to homogenise conventions for annotations and transcriptions to build comparable data is underlined, among others, in Bigi et al. (2012).

We aim to describe a procedure for the annotation of the acoustic components of *nasal grunts* that would not be based on perception so as to homogenise the transcription and annotation of these *non-lexical conversational sounds*. Our procedure for the annotation of the acoustic components of *nasal grunts* was carried out on the sound files from the CID corpus. The annotation guideline illustrated with spectrograms and TextGrid screen captures[1], followed by one annotator, specified how "obvious" features needed to be interpreted from the spectrogram (*i.e.* visual acoustic cues that do not need deeper analysis of acoustic correlates). This annotation is intended as the first stage of a more ambitious analysis of these phenomena allowing to streamline annotated data to be analysed more specifically by specialists according to the annotated variables (*e.g.* creaks, glottal stops, ingressive phonation…).

2.4 The Need for a Qualitative Annotation

In other words, we try to lay the foundations of a principled way to analyse a sample of *non-lexical conversation sounds*. To this end, we claim that, for now, the annotations of the acoustic components of *nasal grunts* remain to be done manually, on a lighter level, based on acoustic cues we hold to be obvious in order to obtain enough manually annotated material for potential automation.

3. Description of the CID Corpus

The CID corpus (Bertrand et al., 2008) "is a 8 hours (110K tokens) corpus composed of 8 conversations of 1 hour. It features a nearly free conversational style with only a single theme proposed to the participants at the beginning of the experiment. This corpus is fully transcribed and forced-aligned at phone level with signal" (Prevot and Bertrand, 2012).

The CID procedure for selecting participants was meant to elicit ecological conversations, *i.e.* the level of intimacy was controlled. This type of conversations is said to be more likely to trigger the use of disfluencies (Audhkhasi et al., 2009). The conditions and procedure of recording the sound files (*i.e.* on separated channels in laboratory conditions) makes it possible not only to analyse *nasal grunts* uttered in overlapping speech, which represent a non-negligible number of grunts (Chlébowski & Ballier, 2015), but also to observe and describe acoustic cues for the components of *nasal grunts* more precisely.

At the time we began our analysis of the grunts of the CID corpus, we only had access to phonetic, syntactic and (parts of the) discursive transcriptions of the CID corpus. A more complete annotation of multimodality is under way at the LPL lab (Prévot and Bertrand, 2012).

4. Definitions of the Acoustic Components of *Nasal Grunts* in the CID Corpus and Description of their Acoustic Cues

All acoustic analysis and annotations were made with Praat software (Boersma & Weenink, 2019). Provided they were tokenised orthographically in the CID transcriptions, we analysed all the grunts whatever their interactional distribution (*e.g.* utterance final or isolated). We did not analyse the grunts we deemed to be clitic[2]. This subsection details all the features we annotated in the CID corpus.[3]

4.1 Syllabification

Syllable division and syllable perception are complex issues (Meynadier, 2001). We do not mean to provide another definition of syllables; we wish to point out the psychoacoustic component in the perception of syllable counts. This has to be borne in mind, because our initial tokenisation of grunts is based on transcripts from the CID corpus, *i.e.* graphic syllables. We potentially risk a contradiction with the prosodic hierarchy, in the sense that our grunts may be phonological words with several phonetic syllables that may not be coded by orthographic tokens. Sequences of orthographic tokens (like four *mh* in a row in less than one second, *e.g.* grunt #700, NH) may correspond to a debatable number of (polysyllabic) phonetic grunts. This paper does not deal with this complicated issue. Suffice it to say that we believe there are only disyllabic or monosyllabic grunts as candidates for phonological words for grunts of C and $\tilde{V}$ types (see section 4.2.4.). For examples like Figure 1, which illustrates what we hold to be dissyllabic grunts for /m/ (Praat screenshot of the spectrogram and waveform of *nasal grunt* number 438 from the MB file of the CID corpus), the analysis is more complex. To prove our point, formant and amplitude tracking are activated on the left and f0 tracking on the right.

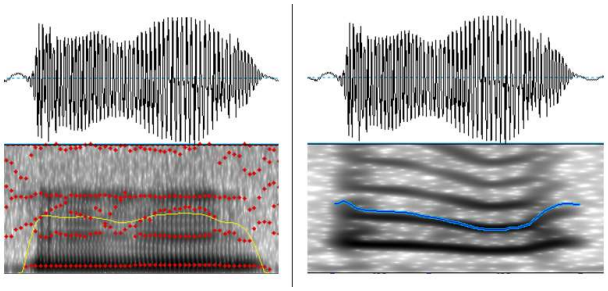

Figure 1: conflicting cues for grunt syllabification #438 (MB file)

As evidenced in Figure 1, some visual criteria are potentially contradictory. If we rely on the variations of the f0 curve (*i.e.* level, fall, rise), three syllables can be

[1] Available from https://github.com/achleb.

[2] Grunts that led to a resyllabification (Clark & Tree, 2002) of nearby words, *e.g. génial hein* /ʒe.nja.lɛ̃/.

[3] We did not investigate the variations of the amplitude on *nasal grunts* in the CID corpus for this phenomenon tackles too many variables.

construed with the image on the right[4]. However, two syllables are suggested by the other visual acoustic cues on the left (*i.e.* significant variations among the two "blocks" of formants, marks on the waveform and spectrogram corresponding to a drop in the amplitude).

Such a contradiction may be used as evidence of either trisyllabic or disyllabic representation of *nasal grunts*. More perception tests would be needed to assert that the 3-grams of perceived f0 variations (right) could be analysed as a level tone followed by a complex tone[5].

Due to space limitations, we only provide one supporting observation in favour of the disyllabic interpretation of grunts and an additional criterion that may be taken into account for this kind of analysis: the insertion of a glottal stop or of a /h/ component splits grunts into two phonetic syllables which we hold to be part of the same phonological word (*e.g.* /m.hm/ or /mʔm/). We nevertheless acknowledge that our syllabic count (especially as it is limited to disyllabic realisations) can be questioned for successive grunts, so that most of our data analysis relies on monosyllabic grunts and on features we deem to be more robust such as duration.

4.2 Segments and Syllable Patterns

Our annotation scheme provides a classification of the syllable patterns (*e.g.* CV(C) forms) of perceived "syllables" of *nasal grunts* in the CID corpus. For vocalic and consonantal segments of *nasal grunts* share a nasal feature, this level of annotation is influenced by *nasalisity* and adapted according to the distinction between *nasalised* (as in *hum*) and *nasal* (as in *hein* and *han*) vowels (Vaissière, 1995). This section proposes to define and describe consonantal and vocalic segments of *nasal grunts* in the CID corpus and their impact on our annotation of the syllable patterns.

4.2.1 Consonantal Segments

Our definition for nasal consonants would be: "Nasals are made with a complete closure in the oral tract, but with the velum lowered so that air escapes through the nose" (Ogden, 2017). The consonantal segment involved in *nasal grunts* seems to be a bilabial nasal consonantal (*i.e.* /m/) according to its orthographic transcriptions in English and French (*e.g. mh, hm…*) as well as its auditory acoustic form (as suggested in Ward, 2004, 2006). However, it would be interesting to investigate consonantal components of grunts from the CID corpus to find strong acoustic correlates that would give credit to a /m/ type realisation of the grunts.

4.2.2 Vocalic Segments

The definition of nasal(ised) vowels varies according to the language under scrutiny. Ogden (2017) and Ladefoged and Disner (2012) explain that nasal vowels are derived from nasal context in English and used as distinctive phonemes in languages such as French. Vaissière (1995) established a clear difference between what she classified as *nasalised vowels* (*voyelles nasalisées*) and *nasal vowels* (*voyelles nasales*). There are vowels *nasalised* by a nasal context, and vowels that can be *nasal* without the need for a nasal context, the velum being lowered in both cases. For example, the pronunciation of the orthographic vowel <i> is not the same in words such as *lapine* (female rabbit) and *lapin* (rabbit) in French. In the first case, the vowel is *nasalised* by the phonemic environment when pronounced, and the coarticulation with the /n/ consonant, but is not fully *nasal* (/la.pin/ or /la.pi.nə/). In the second case, this vowel is fully *nasal* by itself (/la.pɛ̃/).
Nasal vowels in the CID corpus are to be found in *hein* and *han* orthographic tokens while *nasalised vowels* are to be be found in *euh+mh* or *hum*. As stated by Vaissière (2011), nasality is an issue for the distinction of *nasal vowel* qualities in French (p.23). Our phonemic transcription of *nasal* and *nasalised vowels* of grunts from the CID corpus is based on that of their entry in French reference dictionaries (CNRTL). For grunts, vowels and consonants are transcribed as: /ɛ̃/ for *hein*, /ɑ̃/ for *han* and /œm/ for *euh+mh* or *hum* tokens. Then again, this transcription of the "expected" phonemic forms for *nasal grunts* could be subjected to further validation procedures.

4.2.3 Vocalic *vs.* Consonantal segments

The distinction between consonantal and vocalic components is clear in a *nasalised vowel + nasal consonant* contexts where we can see that amplitude drops significantly and a fading spectrogram can be observed between the *nasalised vowel* and the nasal consonant (Ogden, 2017). However, the acoustic cues for *nasality* being a wider and less intense F1 bandwidth and the presence of antiformants (Ogden, 2017; Vaissière, 2011), it is harder to differentiate between nasal consonants and *nasal vowels* when they are not associated. The distinction between nasal consonants and *nasal vowels* in those cases was auditory and needs further investigation to find strong acoustic correlates to clearly distinguish between the two.

4.2.4 Syllable Patterns

Our annotation scheme privileges syllable patterns where we maintain a distinction between *nasalised* vowels and *nasal* vowels. In this sense, we distinguish between "V" which stands for *nasalised* vowels (such as in /œ/ in /œm/) and "Ṽ" which stands for *nasal* vowels (such as / ɛ̃/ and /ɑ̃/). C stands for consonants (/m/ in our case).

4.3 Variations of the Fundamental Frequency (f0) and Register :

The analysis of f0 that we propose is this paper differs from the one made for grunts in the NECTE corpus where Chlébowski and Ballier (2015) tried to assign prosodic contours to variations of the f0 curve. We consider this approach to be somewhat too hasty and we propose a simple visual analysis of the variations of the f0 curve. Acoustic correlates under scrutiny are the variations of the f0 curve. Annotating the variations of f0 might be difficult in cases of non-modal voicing; perception is required in such cases and might be supported by the analysis of the variations of the harmonics.
Register was annotated perceptively and confronted with an analysis that derives from the observation of the

[4] Because this visual representation is ambiguous, we also relied on auditory perception. In this case, it correlated with the three-syllable interpretation.

[5] From a psychoacoustic perspective, it should be noted that the same conflicting cues can be observed between the /œ/ and /m/ parts of *hum* (/œm/) grunts. But our annotator was never tempted to split /œm/ grunts into two syllables.

variations of the f0 curve. This acoustic component "refers to pitch height, that is, whether the intonation contour is relatively low or high in the fundamental frequency scale that a speaker uses" (Snow and Balog, 2002, p. 1027). Usual vocal range of a speaker would correspond to a low register as opposed to a high register when the speaker gets out of his/her vocal comfort zone (Snow and Balog, 2002). Grunts are therefore associated to a register value (either low or high).

4.4 Voice quality and glottal stops:

We claim that variations in phonatory modes as well as the insertion of medial glottal stop can be observed during the production of grunts. We would like to make the point that they are neither artefacts of respiratory features, nor permanent features (Crystal 1985, p.80). Although these instances of non-modal voice are usually characterised as stylistic, they are also likely to convey meaning (Vaissière, 2011).

4.4.1 /h/

The complex distinction between *breathiness* or /h/ as regards the analysis of *non-lexical conversational sounds* is defined by Ward (2006): "/h/ or breathiness is also present in items such as *hm* (versus *mm*), and in the back-channel *uh-huh*. Some such items involve breathiness throughout, others involve a consonantal /h/, while others are ambiguous between these two realizations" (p.11). Cues for both components are similar, but differ as to their distribution. While [h], being a fricative, would look like noise on the spectrogram that can be isolated from other speech events (Ladefoged and Disner, 2012), breathiness would result in *additive noise* over several segments (Hillenbrand, Cleveland, and Erickson, 1994; Wayland and Jongman, 2003). During the annotation process of the CID corpus, we noticed that this acoustic component, either in onset or coda position of perceived syllables, displays a noisy part of a certain duration which seems to spread over the adjacent segment when the component is in onset position. Examples of these phenomena are evidenced on the waveforms and intensity curves on Figure 2. Ogden (2017) reports that: "One commonly suggested analysis of [h] is that it is a period of voicelessness superimposed on a vowel" (p.130). This definition might correspond to our findings, thus making the component under scrutiny a /h/ consonant and not just breathiness. However, this would apply in the case of nasal and nasalised vowels, where the mouth is open. As explained in section (4.2.1), consonantal components of nasal grunts are uttered with the mouth closed throughout. Therefore, /h/ in those cases would also be uttered with the mouth closed and the air will escape through the noise only. Further investigations are needed to fully disambiguate the status of this component, but we will call this component "/h/" in this paper for convenience.

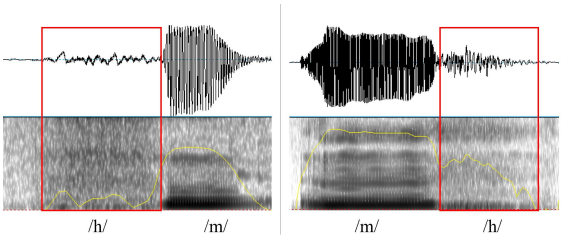

Figure 2: Examples of /h/ as onset (left; #522, AC) and coda (right; #537, LL) of /m/ grunts

4.4.2 Ingressive Phonation

Ingressive phonation, a phenomenon known under many names (DeBoer, 2012), is found in pulmonic ingressive speech. As shown in Eklund (2007), pulmonic ingressive speech is neither restricted to Swedish language nor to words. Interestingly, it seems that this type of speech might also be found in *nasal grunts*. "Pulmonic ingressive speech is the technical term for a phenomenon that is very frequent in Swedish speech: words produced on inhalation airstream" (Eklund, 2007, p.21). As regards our investigation of the CID corpus, the acoustic cue for ingressive phonation in *nasal grunts* would be noise on the spectrogram that is apparently continuous and not necessarily congruent with the whole segment of the grunt[6]. Noisy parts begin before the segment, spread over it and ends after the segment. Most of the time, ingressive phonation seems evenly distributed around the segment, shown by a thicker first formant, of the grunt (as evidenced by Figure 3, on the left), though this is not always the case (on the right).

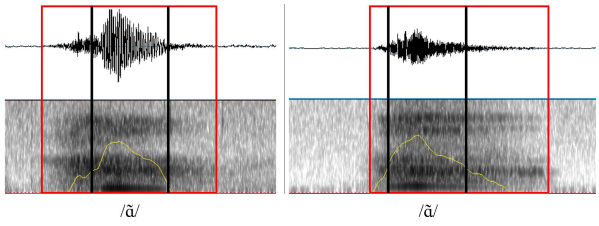

Figure 3: Distribution of ingressive phonation on /ɑ̃/ grunts (left; #226, IM and right; #264, IM).

It is interesting to note that ingressive phonation was found on consonantal as well as vocalic components. However, this feature was found only in /ɑ̃/ vocalic components and not in /ɛ̃/. This might come from the fact that ingressive phonation does not allow certain realisations (Eklund, 2008). Conversely, we did not find any /ɑ̃/ uttered without ingressive phonation.

4.4.3 Creaky Voice

Creaky voice "involves closure along the vocal folds leaving an opening at the front end" (Ogden, 2017, p.50). Occurrences of creaky voice can clearly be seen from "the way vertical striations change from being rather regularly spaced to being irregularly spaced, and further apart from one another (Ogden, 2017, p51). This phenomenon can clearly be seen when creaky voice is used in *nasal grunts*.

[6] Acoustic cue for /h/ and ingressive phonation is therefore the same: noise on the signal. The only difference we made was auditory: /h/ is similar to exhaled breath, while ingressive phonation resembles inhaled breath. Ideally, our analysis on ingressive phonation and /h/ should be supported by other data channel, such as supraglottal measurement (DeBoer, 2012; Gick, Wilson and Derrick, 2013).

We also noticed that the distribution of creaks in *nasal grunts* was not congruous with the entirety of the segment and creaks may occur sporadically (see Figure 4 that displays a Praat screenshot of a monosyllabic /œm/ grunt). Moreover, a certain amount of work showed that there are several types of creakiness (Keating, Garellek, and Kreiman, 2015). Further analyses of this type of phonation have yet to be performed.[7]

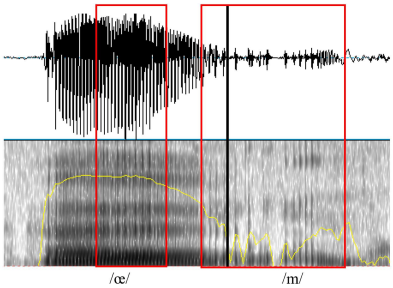

Figure 4: Example of creaky voice (grunt 148, EB).

4.4.4 Glottal Stops

We annotated glottal stops only when they were in medial position of two acoustic syllables. As regard acoustic cues for glottal stops, the "silence-burst bar" couple is not the only cue. Indeed, we sometimes "cannot see the single burst that identifies the presence of a glottal stop in a spectrogram" (Roengpitya, 1997, p.26) in *nasal grunts*. However, glottal stops being a phenomenon of glottalization we can see the sharp drop of fundamental frequency and amplitude (Gerfen 1999)[8]. Those phenomena are illustrated on Figure 5.

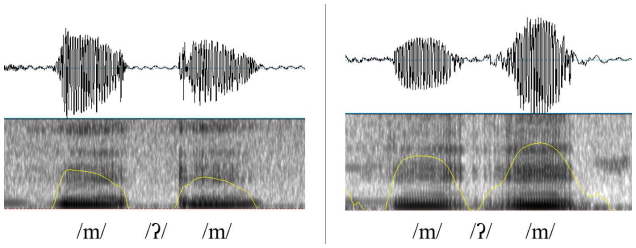

Figure 5: Examples of medial glottal stops in /m.m/ grunts (left; #696, NH and right; #734, NH).

4.5 Duration

Using Praat time stamps for the boundaries of all our annotated features, total durations of grunts as well as duration of specific features can be extracted (duration of the vowel, following consonant for /oem/, glottal stop, creaks, breathy and ingressive phonation). We summarise our endeavour for the distribution of the grunts according to this less disputable acoustic cue in section 7.

5. Investigation of the Acoustic Components of *Nasal Grunts* in the CID Corpus

This section details our annotation scheme using Praat TextGrids.

5.1 Investigating the Orthographic Transcriptions of the CID Corpus

We investigated the TextGrid files that display the phone level of transcription of the CID corpus. We entered simple queries in AntConc software (Anthony, 2004) to get an approximation of the number of *nasal grunts* in the CID corpus (Bertrand et al., 2008) that might be eligible for our analysis. There are 2,159 occurrences of orthographic tokens in the CID corpus that might fit the criteria for our *nasal grunt* category (*i.e.* 424 *hein,* 22 *han*, 166 *hum,* 1,545 *mh* and 2 *hm*). Those number must be taken carefully for the tokenisation of these type of *non-lexical conversational sounds* is disputed (Chlébowski and Ballier, 2015)[9].

5.2 Procedure for the Annotation of the Acoustic Components of Nasal Grunts, a Three-step Approach

This subsection lists the remaining manual annotations and the resulting features in our final dataset.

5.2.1 First listening to the grunts: data selection

The first step of our annotation aims at getting a comprehensive vision of the issues we might be facing. As to what concerns the grunts from the CID corpus (Bertrand et al., 2008), we realised that: 1) the .TextGrid transcriptions were often not aligned with the sounds, 2) some of the grunts were, as expected from previous researches (*e.g.* Chlébowski and Ballier, 2015), linked to the word preceding or following them, 3) laughter misinterpreted for grunts and, 4) surprisingly, there was a huge amount of audible cases of overlapping speech.
We kept the three tiers of annotation from the CID corpus and realigned theirs tokens for selected *nasal grunts* in three other tiers. We also numbered selected grunts and we acknowledged discarded grunts (*i.e.* linked grunts and other problematic grunts such as laughter mistaken for overlapping grunts) on another TextGrid.[10]

5.2.2 Second Listening to the Grunts: Basic Annotations

The second step of our annotation consists in a basic labelling of some acoustic components, namely, syllables, syllable structure and phonemes. The counting of syllables was performed perceptively by our annotator. Subsequent annotations are based on this very tier. The syllabic structure of grunts was aligned, as well as an expected phonemic transcription of the consonantal and vocalic segments of the grunts.

[7] A participant from the CID corpus spoke in a "by-default" creaky voice. Although all occurrences of creaks were annotated, some creaks were still perceived as more intense by our annotator when they seemed to be used for communicational purposes.
[8] It is interesting to note that glottal stops might affect the quality of adjacent components, resulting into phenomenon of creaky voice (Roengpitya, 1997).
[9] We focused our analysis on the results from AntConc software. Thus, "missed items" (Le Grézause, 2017), that could increase the number of analysable *nasal grunts* were not included in our analysis.
[10] We will not discuss the annotations of these discarded sounds.

5.2.3 Third Listening: Refined Annotations

The last step of our annotation aims at getting a less perception-based annotation of the acoustic components of the grunts. Syllabification suggested by acoustic cues (other than f0 curve), f0 variations[11], voice quality and medial glottal stops are annotated and their distribution inside a given grunt is specified[12]. Duration was not annotated since it can be extracted from any tier with a simple Praat script. A tier is dedicated to various comments our annotator made during the annotations.

5.3 Labels for the Annotation of the Acoustic Components and their Distribution

A restricted number of labels per acoustic component was constituted in order to facilitate the annotation process and further analysis of the data.

Syllabification was annotated both perceptively and visually (according to acoustic cues other than the variations of the f0 curve). Syllables were aligned manually accordingly and numbered inside each grunt (*e.g.* "1", "2"…). Syllable structure of the grunt was aligned on the signal manually. Labels for syllabic structure are either "C", "V" or "V̂". Segments were aligned manually. Labels for their phonemic transcription are either /œ/, /œ̃/, /ɛ̃/ or /m/. Labels for the annotation of the variations of the f0 curve are "level", "fall" and "rise". The distribution of the variations of the f0 curve was aligned on the signal and seems to be strongly related to that of the number of perceived syllables[13]. Labels for the perception of register are either "low" or "high".

Labels for the annotation of voice qualities and glottal stop are either "yes" or "no". The distribution of ingressive phonation was not investigated for reasons explained in section 4.4.2. Therefore, annotation of the distribution of ingressive phonation corresponds to that of the whole grunt. Distribution of glottal stops was aligned manually and labelled as "s1_s2" (*i.e.* between first and second perceived segment) or "s2_s3" (*i.e.* between second and third perceived segment). For the annotation of the distribution of creaky voice, we adopted the B - 'beginning', I - 'inside', L - 'last', O - 'outside', U - 'unit' (BILOU) format, adapted from the IOB (Inside, Outside, Beginning) format (Ramshaw and Marcus, 1995). Figure 6 gives examples of this annotation on a monosyllabic and a disyllabic grunt of "C" type. For instance, "B-s1" stands for creakiness being at the beginning of the first perceived segment.

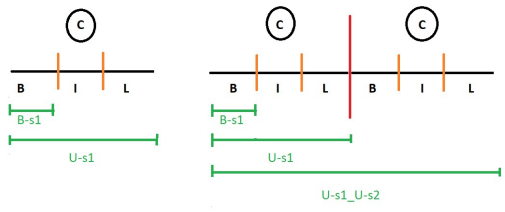

Figure 6: BILOU annotation scheme for monosyllabic (left) and disyllabic (right) grunts.

6. Annotation of the Distribution of *Nasal Grunts* in the CID Corpus

Investigating the distribution of *nasal grunts* as regards interactional context was proposed by Chlébowski and Ballier (2015). Confronting the annotation of the acoustic components of grunts and the distribution of grunts in interaction might tell us more about the information conveyed. To cater for this need, we have duplicated the annotations distinguishing between positional properties for the speaker and the interactional status of grunts. The most common annotation for grunt position distinguishes three positions (initial, final, medial) in the speaker's speech. This disregards the fact that a grunt might be medial but surrounded by silences. We have added a supplementary category for the position in the speaker's speech: the isolated grunts (*i.e.* grunts surrounded by silences). As to the distribution of grunts in terms of interaction, we follow standard practice of distinguishing self *vs.* others (Fruehwald, 2016)

6.1 Annotation Procedure and Labels for the Position of Grunts in the Speaker's Own Speech

We annotated the distribution of *nasal grunts* in the speaker's speech according to four possibilities, namely whether the grunt is: 1) surrounded by silences, 2) at the beginning of an utterance, 3) at the end of an utterance, or 4) medial. A tier was created for each one of those possibilities and our annotator was asked to fill them with either "yes"[16] or "no" labels.

6.2 Annotation Procedure and Labels for Interactional Features

The annotation of the distribution of *nasal grunts* as to interactions is based on four possibilities, namely whether: 1) Speaker is speaking before the grunt, 2) Speaker is speaking after the grunt, 3) Interlocutor is speaking before the grunt, 4) Interlocutor is speaking after the grunt. A tier was created for each one of those possibilities and our annotator was asked to fill them with either "yes"[17] or "no" labels in a +/- 1 second window around the grunt. This categorical multitier system for 6.1. and 6.2. was deemed to be much faster by the annotator.

7. Preliminary Results

This section presents a typological sample of some of the exploitations of our resulting dataset and a putative conceptualisation of *nasal grunts*. For quantitative results, we propose to focus on the acoustic component *duration*. First, we propose our findings on monosyllabic grunts uttered in modal voice (the less objectionable acoustic cue we have collected, for reasons explained in section 4). Then, we discuss the interactional properties of the acoustic components of *nasal grunts* by showing the impacts of syllabification and voice quality on the duration of grunts. Only the tokens for which the perceptual and acoustic

[11] Settings of the Praat software were adapted to men and women and the tracking of f0 was set in semitones re 1 Hz. Harmonics were controlled with a narrow band spectrogram.

[12] The perception of the register was also acknowledged at this stage.

[13] For they are composed of two different segments, the annotation of the variations of the f0 curve was performed on the component "V" and "C" of "VC" grunts.

[16] Overlaps are considered with the "yes_overlap" label.

[17] Overlaps are considered with the "yes_overlap" label.

analysis agreed (kappa= 0.8) for syllabification were considered (see section 4.1).

7.1 Overall Duration

Figure 7 shows the whole duration in milliseconds of the 515 monosyllabic grunts uttered in modal voice. It distinguishes between men and women and does not distinguish between the various positions of grunts.

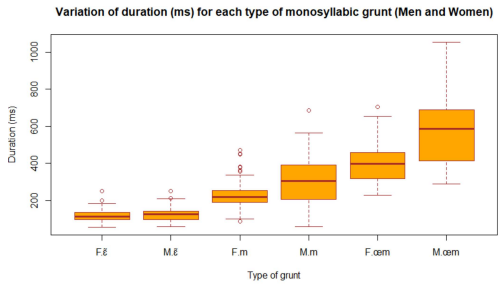

Figure 7: Duration (ms) for men (M) and women (F)

Even if the duration of /œm/ accounts for two segments, the difference in duration between men and women is significant. The data was tested for heteroscedasticity with a Levene test ($p<0.001$) and an ANOVA showed a strong interaction of duration with sex and type of grunt, but no real interaction with utterance final position.

	ɛ̃	m	œm	Total
F	124	253	27	404
M	58	42	11	111
Total	182	295	38	515

Table 1: distribution of monosyllabic grunts

It should be noted that in the whole dataset a great variety of grunts can be observed among the 947 grunts analysed.

7.2 Influence of voice quality on duration

To analyse the main determinants of grunt duration, we drew a conditional inference tree using the R package {party} (Hothorn et al., 2006). Figure 8 shows the output of the conditional inference tree analysis. In this kind of representation, out of the six variables included in the Classification Tree analysis, the main criterion for the classification is ranker higher in the classification tree. For instance, the type of grunt accounts for a first subdivision (ɑ̃, ɑ̃.ɑ̃, œm vs. ɛ̃, ɛ̃.ɛ̃, m, m.m, m.m.m, see node 1) and creaky voice is then what mostly distinguishes longer and shorter ɑ̃, ɑ̃.ɑ̃ and œm. As can be observed by its position in the classification tree, the number of perceived syllables is less important than the use of creaky voice which is higher in the hierarchy. As was predictable, phonation types play a role in duration, and the regression tree shows at nodes 2, 6, 10 and 12 that non-modal phonation types have longer realisations than their modal counterpart for each type of grunts considered. Our last presentation of results zooms in nodes 11 and 12 to consider the duration of grunts realised with ingressive phonation.

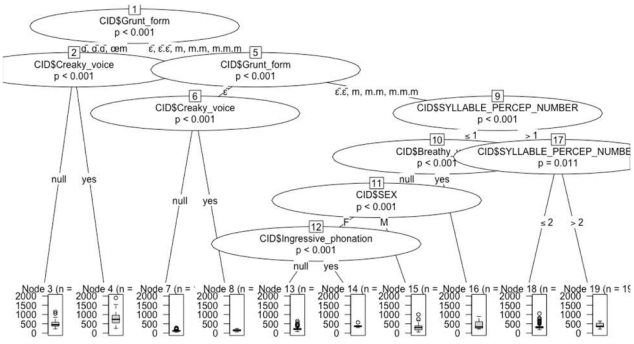

Figure 8: Classification Tree visualizing Duration ~ Number of syllables + Sex + Creaky voice + Breathy voice +Ingressive phonation

7.3 Ingressive phonation

When investigating ingressive grunts, a certain specialisation of the grunt forms seems to be at work. Men only produce / ɑ̃ / with ingressive phonation – at least in the CID (Figure 9).

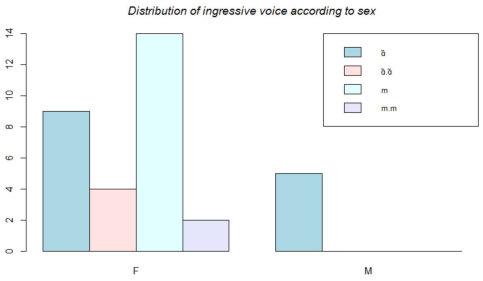

Figure 9: Distribution of ingressive voice according to sex

By contrast, in isolated form (between pauses), women realised more ingressive grunts in the CID and all their ingressive /m/ were found in this position.

7.4 A Proposal for the Anatomy of *Nasal Grunts* in the CID Corpus: the Core Components

We observe that grunts may have additional components such as insertions (glottal stops or /h/) or phonation types (creaky voice or ingressive phonation). We propose that these components are additional in relation to what we believe to be core components. The five layers of the box displayed on Figure 10 summarize the components we hold to be at the core of the production of a monosyllabic *nasal grunt* uttered in modal voice in the CID corpus. The symbols on the right recap the range of possible values for each subcomponent. For segments, vocalic and consonantal components were found, namely: the nasal vowel /ɛ̃/ and the nasal consonant /m/. The vowel /œ/ was also found in monosyllabic /œm/ grunts and is therefore nasalised by phonetic context. Three contours of pitch variations were used for the annotation : level, rise and fall, based on visual inspection of the f0 curve. We claim that there is a mean duration for monosyllabic grunts uttered in modal voice (see Figure 10) that differs for each segment type and it might be either decreased or increased for communicational purposes. We did not investigate the variations in amplitude in *nasal grunts* from the CID corpus. Nonetheless, we hypothesise that the variations in amplitude would behave like duration, that is with a baseline. Finally, two levels were identified for the

characterisation of the register in which grunts were uttered, namely: low and high.

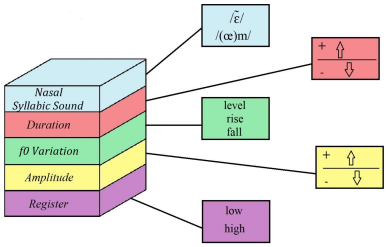

Figure 10: The core components of a monosyllabic grunt uttered in modal voice in the CID corpus.

The distribution of /œ/ and /m/ in /œm/ grunts seems to be fixed: /œ/ will always appear before /m/. This phenomenon not only shows that distinct *non-lexical conversational sounds* (*euh* and *mh* in this case) can combine but also points out the existence of a grammar specific to those sounds. The analysis of the distribution of components that may convey additional information (such as creaky voice, /h/, ingressive phonation and the insertion of a medial glottal stop) as well as that of syllabification in a given grunt is under way. This study would show whether those components appear as sequential or combinatory with the core components of grunts, thus reinforcing our idea of an intra-grunt grammar.

8. Discussion and Conclusion

In this paper, we have presented a procedure for the manual annotation of nasal grunts, to analyse the complexity of phonetic events associated to these productions. The corresponding TextGrids are to be associated to the CID TextGrids and could be made available from the Ortolang repository. The extracted dataset and the annotation guidelines will be available from https://github.com/achleb. The following subsection summarises the potential shortcomings of our method and our resulting precautions. We insist on three kinds of restrictions.

Due to the potential ambiguity when chunking polysyllabic grunts, we mostly reported results on monosyllabic grunts in this paper. Nevertheless, as to the ambiguity of our subdivision of complex grunts into phonetic syllables, it should be borne in mind that the orthographic tokens of the CID corpus convey similar ambiguity with the variety of forms they use for the inventory of tokens (*mh mh*, etc).

An annotation of the sound in its various dimensions requires phoneticians to be careful as to the analysis of the extracted data, for acoustic phenomena might interfere with one another. We here further elaborate on our precautions regarding f0 extraction. The use of non-modal voice might impact the variations of the f0 and their analysis (DeBoer, 2012; Keating et al., 2015). For instance, the use of creakiness and ingressive phonation interfered with our annotation of the f0 and had impact on the automatic extraction of mean, min and max f0, *i.e.* a certain amount of "--undefined--" were found with Praat. This led to our reluctance to automate the f0 extractions and to control the register in which grunts were uttered. We did not report the results of our pitch extraction, due to the complex interaction with breathy, creaky and ingressive phonations but we have collected datasets that could be used for mixed-effect modelling of the effect of voice quality on f0.

We acknowledge that only one annotator took part in the experiment. The only inter-annotator analysis we performed so far was against the original transcriber of the CID corpus (kappa=0.18 between the annotator phonemic transcription and the phone tokenisation of the CID)[18]. Replicating the annotation by another transcriber is obviously the next step, as well as a comparison with automatic detection of features such as creaks with Voicesauce (Shue et al., 2008) or syllable peak detection with an algorithm (de Jong and Wempe, 2009).

Another line of future investigation bears on the elaboration of visual perception tests intended to check the ability of trained phoneticians to replicate the type of judgement passed by the annotator, namely identifying the types of f0 variation among a closed set of three choices: level, fall or rise, on the basis of a screen capture from Praat displaying f0 curve. To test the robustness of our annotation judgements, a second perception test is to bear on perceived syllable patterns (like VC), type of vowel perceived, register, insertions and phonatory modes. These two tasks are designed for control in perception tests on specific features we hold to be distinguishable among annotators. The most daunting task for the annotator proved to be the syllabic count of these realisations.

We believe that a semasiological approach of phenomena such as *nasal grunts* is needed to fully understand their meanings and functions in interactional context. We acknowledge that the process of setting up the annotation guidelines was time-consuming and that our results must be submitted to validation procedures. Nonetheless, our procedures and annotations can benefit the research community of corpus linguists to analyse *er, um, uh* and the like in spoken data. Our findings as regards a potential intra-grunt grammar are promising and testify that our study was worth the workload. The replicability of our annotations to corpora in languages other than French was performed by our annotator[19] on *nasal grunts* of the Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000).

This study remains a preliminary step to a semantic evaluation of those phenomena which might hold for their distinctives functions. The distribution of the acoustic components in a given grunt, as well as their interactions and variability, could be at the core of the reasons why those sounds were assigned so many meanings and functions.

9. Acknowledgements

We thank Roxane Bertrand for her help with the speech language and data repository (SLDR) and advice about the CID. We also would like to thank her and Sophie Herment (as well as the reviewers) for their comments on an earlier version of this paper.

[18] We did not report missed or hallucinated grunts, (Le Grézause's terms, 2017) when investigating the misperceptions of *um* and *uh* in the Switchboard corpus.

[19] The whole annotation of around 200 *nasal grunts* can be made manually is less than a week, when trained.

10. Bibliographical References

Audhkhasi, K., Kandhway, K., Deshmukh, O. D. and Verma, A. (2009). Formant-based technique for automatic filled-pause detection in spontaneous spoken English. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4857–4860. IEEE.

Bigi, B., Péri, P. and Bertrand, R. (2012). *Influence de la transcription sur la phonétisation automatique de corpus oraux*.

Chlébowski, A. and Ballier, N. (2015). Nasal grunts" in the NECTE corpus, Meaningful interactional sounds. *EPIP4-4th International Conference on English Pronunciation: Issues & Practices*, 54–58.

Clark, H. H. and Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Crystal, D. (1985). *A Dictionary of Linguistics and Phonetics, 2nd edn updated and enlarged*. Oxford, Basil Blackwell in association with André Deutsch.

DeBoer, A. R. (2012). *Ingressive phonation in contemporary vocal music* (PhD Thesis). Bowling Green State University.

Delomier, D. (1999). Hein, particule désémantisée ou indice de consensualité? *Faits de Langues*, *7*(13), 137–149.

Eklund, R. (2007). Pulmonic ingressive speech: A neglected universal? *Fonetik 2007, 30 May–1 June 2007, Stockholm, Sweden*, 21–24. Universitetsservice.

Eklund, R. (2008). Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. *Journal of the International Phonetic Association*, *38*(3), 235–324.

Fruehwald, J. (2016). Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, *22*(2), 6.

Gerfen, C. (1999) Amplitude drop as the primary cue for glotatlization: evidence from production. Paper presented at the Linguistic Society of America Meeting, Los Angeles, California, 8-10 January.

Hillenbrand, J., Cleveland, R. A. and Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, *37*(4), 769–778.

Keating, P., Garellek, M. and Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. *Proceedings of ICPhS*.

Ladefoged, P. and Disner, S. F. (2012). *Vowels and consonants* (3rd ed). Malden, MA: Wiley-Blackwell.

Le Grézause, E. (2017). *Um and Uh, and the expression of stance in conversational speech* [PhD Thesis].

Meynadier, Y. (2001). *La syllabe phonétique et phonologique: Une introduction.*

Milroy, A. L., Milroy, J. R. D. and Docherty, G. J. (1997). *Phonological variation and change in contemporary spoken British English*. Economic and Social Research Council.

Ogden, R. (2017). *Introduction to English Phonetics*. Edinburgh University Press.

Prévot, L. and Bertrand, R. (2012). *Cofee-toward a multidimensional analysis of conversational feedback, the case of French language*.

Ramshaw, L. A. and Marcus, M. P. (1995). Text chunking using transformation-based learning. CoRR. *ArXiv Preprint Cmp-Lg/9505040*, *50*.

Roengpitya, R. (1997). Glottal stop and glottalization in Lai (connected speech). *Linguistics of the Tibeto-Burman Area*, *20*(2), 21–56.

Snow, D. and Balog, H. L. (2002). Do children produce the melody before the words? A review of developmental intonation research. *Lingua*, *112*(12), 1025–1058.

Tottie, G. (2019). From pause to word: *Uh, um* and *er* in written American English. *English Language & Linguistics*, *23*(1), 105–130.

Vaissière, J. (1995). Nasalité et Phonétique In Le voile du palais et la parole. *Colloque Sur Le Voile Pathologique*.

Vaissière, J. (2011). *La phonétique* (2e éd. mise à jour). Paris: Presses Universitaires de France.

Ward, N. (2000). Issues in the transcription of English conversational grunts. *1st SIGdial Workshop on Discourse and Dialogue*, 29–35.

Ward, N. (2004). Pragmatic functions of prosodic features in non-lexical utterances. *Speech Prosody 2004, International Conference*.

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, *14*(1), 129–182.

Wayland, R. and Jongman, A. (2003). Acoustic correlates of breathy and clear vowels: The case of Khmer. *Journal of Phonetics*, *31*(2), 181–201.

11. Language Resource References

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL*, 7–13.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B. and Rauzy, S. (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique Des Langues*, *49*(3), 105-134.

Boersma, P. & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.06, retrieved 8 November 2019 from http://www.praat.org/

Corrigan, K., Allen, W., Beal, J., Maguire, W., Moisl, H. and Rowe, C. (2001). *Newcastle Electronic Corpus of Tyneside English Corpus*.

CNRTL Centre National de Ressources Textuelles et Lexicales – http://www.cnrtl.fr/

De Jong, N. H. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., & Martey, N. (2000). Santa barbara corpus of spoken american english. CD-ROM. Philadelphia: Linguistic Data Consortium.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355-373.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Shue, Y.-L., P. Keating , C. Vicenik, K. Yu (2011) VoiceSauce: A program for voice analysis, *Proceedings of the ICPhS XVII*, 1846-1849.