Towards a typology of ASR errors via syntax-prosody mapping

Fabian Santiago, Camille Dutrey, Martine Adda-Decker

Santiago, F., M. Adda-Decker & C. Dutrey. (2015). Towards a Typology of ASR Errors via Syntax-Prosody Mapping. In G. Adda, V. Barbu Mititelu, J. Mariani, D. Tufiș & I. Vasilescu (eds). Errors by Humans and Machines in Multimedia, Multimodal and Multilingual Data Processing. Proceedings of ERRARE 2015. Satellite event of Interspeech 2015. Bucharest: Editura Academiei Române. 175-192.

TOWARDS A TYPOLOGY OF ASR ERRORS VIA SYNTAX-PROSODY MAPPING

Fabián SANTIAGO [1], Camille DUTREY [1, 2,] Martine ADDA-DECKER [1, 2]

[1] LPP (UMR 7018) – Université Sorbonne Nouvelle & CNRS
[2] LIMSI – CNRS
{fabian.santiago-vargas, martine.adda-decker}@univ-paris3.fr, {madda, camille.dutrey}@limsi.fr

Abstract. This study explores automatic speech recognition (ASR) errors from a syntax-prosody mapping perspective. Our contribution is threefold: we propose (i) an ASR error study according to French syntactic structures, (ii) a quantitative evaluation of the syntax-prosody mapping in reference transcriptions (iii) a qualitative analysis of syntax-prosody mapping violations in ASR transcriptions. Results show that some morphosyntactic and syntactic components are particularly prone to transcription errors such as proper names or verbal nuclei. In addition, we found that transcription errors in the ASR hypothesis may violate the syntactic-prosodic mapping rules. Such conflicting patterns may be used as clues to automatically detect ASR errors.

1. INTRODUCTION

Automatic speech recognition (ASR) has made tremendous progress over the last decades and error rates typically fall below 10% in many types of broadcast speech. In this contribution, we investigate relatively difficult French broadcast speech corresponding to various types of conversations with error rates above 20%. Analyzing ASR errors allows us to get a better idea of the relative strength and weaknesses of the speech model used to decode the speech acoustics. Studying ASR errors then aims at sorting out the relative contributions due to production variation from those due to intrinsic language ambiguities such as homophones or relatively infrequent wordings, or whether they.

Besides environmental characteristics (noisy background or channel), ASR errors are generally related to the following factors: speaking style, speaking rate

and articulation, speaker gender, word frequency… ASR systems are known to have difficulties with short words (although frequent) and infrequent words, typically proper names [1, 2]. More importantly, it could be observed that ASR errors significantly increase in spontaneous speech as compared to read speech. Different explanations can be proposed: spoken language wordings happen to be very different from written language ones. Thus, word frequencies (and language models) can be more reliably estimated from written sources for read speech than for spontaneous speech. Next, spontaneous speech may be less carefully articulated with different types of reduction phenomena [3]. Moreover, spontaneous speech tends to include more speaker hesitations and restarts [4, 5, 6, 7]. Some studies have also investigated ASR errors with respect to speaker gender [8]. Despite a considerable amount of research on ASR errors (*e.g.* [9]), studies analyzing ASR errors beyond the word-level are still under-represented. In particular, only few studies have addressed ASR errors taking a morphosyntactic perspective [10] or even combining both morphosyntactic and prosodic levels [11, 12].

In this contribution, our focus goes to syntax and to the syntax-prosody interface. The files selected for our study are part of the French Spoken Treebank compiled by [13]. They benefit from thorough syntactic annotations, which were manually checked and corrected if necessary. We first examine ASR output for error distributions with respect to part of speech (POS) classes and syntactic constituents. We then compute prosodic features corresponding to the syntactic groupings and ask whether some ASR errors could be detected in regions of weak syntax-prosody match.

The explored idea is the following: when grouping ASR output into syntactic constituents, the prosodic phrasing of the corresponding speech is expected to follow a typical French prosodic pattern with phrase-final rises and lengthening. This hypothesis is particularly expected to be true in correctly transcribed speech. However, ASR transcription errors may result in syntactic groupings conflicting with the actually produced prosodic pattern. Such conflicting patterns may then become clues as to the presence of ASR errors. As far as we know, there are no studies evaluating ASR errors using syntax-prosody mapping rules that account for prosodic phrasing patterns (at least in French). This paper deals with the latter topic. We first analyze ASR errors as a function of different syntactic structures before investigating syntax-prosody mapping rules in manual and automatic speech transcriptions.

2. MATERIALS

This study makes use of the French ETAPE corpus in [14] which was collected for the 2012 evaluation of spoken multimedia content processing. The

ETAPE corpus, distributed by ELDA (www.elda.fr), is composed of TV and radio broadcasts and covers a large variety of speaking styles, emphasizing spontaneous conversational speech. We selected the subset of files, which are also part of the French Spoken Treebank in [13]. For these files, we thus have both syntactic annotations of the manual reference transcripts as well as ASR transcripts. Our subset is exclusively composed of radio shows from the French public *France Inter* radio. It includes 25 speakers for a total of 1.5 hours of speech corresponding to debates on the latest cultural developments about films, literature and theatre (*Le Masque et la Plume*) and stories involving non-professional speakers collected in the wild (*Un temps de Pauchon*). The manual transcriptions (henceforth called "reference") contain a total of 16,377 words. The automatic transcriptions (henceforth called "hypothesis") were provided by the LIUM system, which achieved an average word error rate (WER) of 22.6% (reported in [15]) on the full ETAPE test data set. Details of our particularly difficult subset are shown in Table 1. As can be seen, the WER is quite high (from 27.4% up to 52.9%) due to highly spontaneous and very interactive speech, with up to six interacting speakers per show. This is the case of *Le Masque et la Plume (MP)* with multiple culture columnists of Radio France. For *Le Temps de Pauchon (TdP)*, the reporter goes to meet people at their places (exhibitions, streets, stores, farms...). These shows address a wide panel of different topics which further contributes to make the data most challenging for automatic speech recognition and downstream natural language processing.

Table 1

Main characteristics of the selected ETAPE subset: 1 cultural debate MP show (*Le Masque et la Plume*); 3 field reporting TdP shows (*Un Temps de Pauchon*)

Radio Show	**Duration**	**# Sentences**	**# Words**	**Word Error Rate**
MP	1:06:02	788	11,829	36.6%
TdP_1	0:10:57	153	1,636	52.9%
TdP_2	0:10:57	132	1,640	41.9%
TdP_3	0:10:57	93	1,272	27.4%
Total	1:38:53	1,166	16,377	39.7%

To study ASR errors from a syntactic-prosodic perspective, we first made a semi-automatic extraction of morphosyntactic annotations provided by the French Spoken Treebank compiled by [13]. In the framework of the French ANR ETAPE project, the French Spoken Treebank (FST) has been developed as an extension of the French Treebank [16], a large resource of French written press articles. The FST rench Treebank is enriched with three types of grammatical annotations:

morphosyntactic annotations (POS labels), constituent annotations and function annotations. These three types of annotations were also added to the FST. Furthermore, [13] introduced speech-specific labels into FST to account for events such as speech disfluencies, discourse markers and overlapping speech. The French Treebank and the FST aim at being usable by researchers from various theoretical backgrounds as well as easily understandable by human annotators. The adopted grammatical annotations should be as far as possible "theory neutral", compatible with various syntactic frameworks (see [16] for an in depth discussion). There are neither discontinuous constituents nor empty categories but headless phrases are allowed. As an illustration, Figure 1 shows the clause *il s'est vendu dans le monde entier* ("it has been sold all around the world") in the FST annotation framework. Each word is associated with a POS label and various syntactic nodes (constituent labels) to represent the grammatical structure of the clause. When necessary, function labels are provided as well. All nodes derive more or less directly from the root node (main clause or SENT). These annotations represent the grammatical structure of the clause according to the FST.

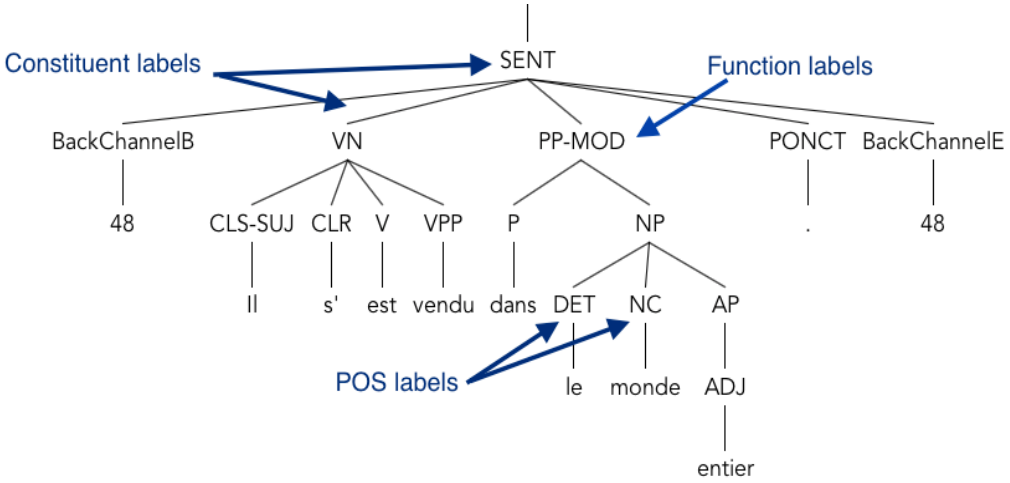


Figure 1. Grammatical annotations of the clause *il s'est vendu dans le monde entier* ('it has been sold all around the world') as provided by the French Spoken Treebank.

To link syntactic information with prosody, prosodic annotations were automatically created at word level. Transcripts (both reference and hypothesis) were automatically time-aligned using the LIMSI system described in [17] in forced alignment mode. With the help of a *Praat* script, different measurements including f_0 and formant values were automatically extracted. As forced alignment provides both word and phone boundaries, acoustic-prosodic features words such

as durations and lengthening, f_0 and f_0 changes could be automatically computed on phone and word levels. These prosodic features will be detailed in Section 3.2.

3. METHODS

This section describes the method towards a prosody-syntax mapping, starting with morphosyntactic and syntactic units of analysis in Section 3.1. Section 3.2 describes the prosodic unit level of this study: the accentual phrase. Working at the interface of syntax and prosody, we face conflicting terminology conventions: AP may be used both to refer to adjectival phrases in syntax and to accentual phrases in prosody [18, 19]. In the remainder of the paper, we keep AP for adjectival phrase and we will term accentual phrases as prosodic words adopting the PW notation. In Section 3.3, we present how both syntactic and prosodic features were used in order to set out a syntactic-prosodic mapping accounting for the prosodic phrasing in French and how this mapping may highlight regions of ASR errors.

3.1. Morphosyntactic features

From the different labels that are provided by the FST, only POS labels and constituent annotations were used in this study. Our goal was to examine how ASR errors distribute accross POS categories and syntactic constituents. All POS and constituent labels were automatically extracted from the FST (French Spoken Treebank) together with the corresponding reference words *via* a Perl script. We manually discarded 4,241 words (25.8% of the original amount of words) produced in overlaps or in sequences immediately preceding/following overlaps. The extracted information was also converted into different *Praat* TextGrid files in order to facilitate qualitative cross-comparisons.

Concerning the **POS labels,** we grouped the 25 FST original POS categories into a new tag set of 11 labels for lexical words, grammatical words and interjections. The reduced POS tag set illustrated in Table 2 enables us to get a more word tokens per POS tag by discarding less relevant distinctions of the original tags.

As for **constituent labels,** we extracted the grammatical tags of the POS immediate parent nodes in the FST. For instance, for the clause *il s'est vendu dans le monde entier* (cf. Figure 1), the extraction produces the words-labels association as follows: [il s'est vendu]$_{VN}$ [dans]$_{PP}$ [le monde]$_{NP}$ [entier]$_{AP}$. The FST provides a 12 label tag set for constituent sequences, comprising in our data from one to eight words. The constituent label set is summarized in Table 3. Constituent heads are in

bold except for embedded finite and root clauses. In the following, we refer to generic constituents as XP (X for "generic" and P for "Phrase") without specifying categories such as noun phrase, verbal nucleus, etc.

Table 2
Reduced (11) part-of-speech tag set derived from the 25 morphosyntactic labels in FST

Lexical Words	ADJ	adjectives	*entier* 'whole' / *anciens* 'old'
	ADV	adverbs	*très* 'very' / *doucement* 'slowly'
	NC	common names	*monde* 'world' / *films* 'movies'
	NPP	proper names	*Clint Eastwood* / *Morlaix*
	V	verbs	*vendu* 'sold' / (*ils*) *sont* '(they) are'
Grammatical Words	CL	weak clitics	*il* 'he' / *je* 'I'
	C	conjunctions	*et* 'and' / *comme* 'since'
	DET	determiners	*la* 'the' / *mon* 'my'
	PRO	strong pronoun	*que* 'that' / *qui* 'who'
	P	prepositions	*dans* 'in' / *de* 'from' or 'of'
Interjections	I	interjections	*euh* 'hum' / *ah* 'eh'

Table 3
Constituent tags as extracted from the French Spoken Treebank (FST)

AdP	adverbial phrases	*Il parle* ***peu*** 'he speaks little'
AP	adjectival phrases	***très très*** *bon* 'very very good'
COORD	coordinate phrases	***et*** *toi?* 'and you?'
NP	noun phrases	*certains* ***films*** 'some movies'
VN	verbal nucleus	***j'ai râlé*** 'I complained'
PP	prepositional phrases	***dans*** *le monde entier* 'all around the world'
VPinf	infinitive clauses	***être*** *boucher* 'being a butcher'
VPpart	nonfinite clauses	*en* ***entrant*** 'getting in'
Sint/Srel/Ssub	finite clauses	*quand j'étais gamin* 'when I was a little boy'
SENT	sentences	*Claude est au téléphone* 'Claude is on the phone'

3.2. Prosodic features

Basic acoustic information was automatically extracted from the signal. The acoustical analysis was conducted using *Praat* [20] *via* Perl scripts. Extracted

measurements were then combined to derive different prosodic features using either reference or hypothesis forced alignments. Differences in reference and hypothesis features are then due to transcription errors. For both reference and hypothesis streams, we generate various acoustic-prosodic features at the phone, syllable and word levels. In this study, we mainly focus on variations in pitch (f_0) and durations at right boundaries of prosodic words (PWs). Hence, two main prosodic features were set up from the acoustical analysis:

- Delta f_0 values: they consist in differences in semitones between the final and the initial vowel of each word. f_0 values were computed in the middle of each vowel. Positive delta values correspond rising melody movements. Delta f_0 > 2 st in final vowels was encoded as H* (an important rise was realized). When final vowels had delta values < 2 st. they were encoded as 0* (no important melody movement). Conversely, important negative delta values (Delta f_0 < -2 st) were encoded with L* label. The labels H*, 0* and L* stand for rising pitch accent, the absence of pitch accent and falling pitch accent respectively. We will come back to these terms in Section 3.3.
- Durational cues: durations of all words and constituents in the reference were computed and normalized. Normalizations were carried in both words and constituents as follows: normalized duration = duration of word or constituent / number of comprising phones in the word or constituent).

3.3. Syntax-Prosody mapping: derivation of Prosodic Words in French

In this section, we describe briefly the most common prosodic group described in French studies: the prosodic word (PW) and how we derive this prosodic unit on the basis of the syntactic-prosodic mapping. According to several authors, (cf. among others [19, 21]), French has been generally described as a rising language: utterances are parsed internally into prosodic units called accentual phrases or *prosodic words* (PWs) that tend to end with a rising f_0 shape in addition to a final-syllable lengthening. Authors agree that many information types should be considered to predict the places of the speech chain, where French speakers produce PWs, in particular, morpho-syntactic, rhythmic and extralinguistic information. According to [19] and [21], PWs may be predicted using the following syntactic-prosodic mapping:

- A PW groups together any lexical word and its dependent functional words at their left side within the maximal projection and with any other non-lexical item on the same side [19, 21].

We exemplify this syntactic-prosodic mapping using the same clause as in Figure 1: *il s'est vendu dans le monde entier* can be parsed into three PWs as shown below. The lexical heads (in bold) of each syntactic constituent (in square brakets) align at their right edge with three PW's (in braces) grouping together the lexical head and all the preceding items:

- [il s'est **vendu**]$_{XP}$ [dans le **monde**]$_{XP}$ [**entier**]$_{XP}$
- {il s'est vendu}$_{PW}$ {dans le monde}$_{PW}${entier}$_{PW}$

It is commonly accepted that the right edge of a PW is realized with a final pitch accent (if any). In French, prefinal pitch accents are generally realized with important f_0 rises and lengthening [19, 21]. Hence, prefinal pitch accents tend to take the form of H* and more rarely the form of L*. The absence of pitch accents (when speakers do not produce them), *i.e.* a static f_0, is represented here with the 0* label.

Using this mapping rule, we measured the prosodic implementations in the acoustic data of the right edges of constituents as provided by the syntactic information. In particular, the presence of H* labels in the signal occurring at the rightmost full syllable of constituents' content words was considered as evidence of the presence of the right edge of the predicted PWs. In other words, the presence of H* labels in the rightmost constituent positions shows that the syntactic-prosodic mapping rules apply in our data, whereas 0* labels indicate that the mapping fails. Figure 2 illustrates the mapping: if the rightmost full syllables of the words *vendu*, *monde* and *entier* were produced with H* (an f_0 increase of at least 2 semitones), we considered that the PW's right boundaries were prosodically produced.

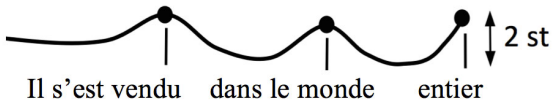


Figure 2. f_0 stylisation of the clause *il s'est vendu dans le monde entier* ('it has been sold all over the world') with circles indicating the presence of three H*/PWs.

We have to note that syntactic-prosodic mapping is not the only constraint to consider for predicting the PWs in speech. The absence of syntactically predicted pitch accents may be explained by various reasons. The first one concerns rhythmic effects [22]: syntactic constituents must be long enough (in terms of number of syllables) for calling the realization of PWs. For instance, the chain [il est]$_{XP}$ [grand]$_{XP}$ ('he is tall') will certainly be parsed into only one PW {il est grand}$_{AP}$ instead of two {il est}$_{AP}$ {grand}$_{AP}$: generally, PWs comprising only one single monosyllabic word like *grand* do not form an independent PW. A second reason

deals with speech rate: fewer pitch accents tend to be produced in spontaneous or faster speech as compared to slower or read speech (see, for instance, [22]). In Section 5, we will examine to what extent both reference and hypothesis transcripts respect the syntax-prosody mapping explained above. Our prediction is that reference words of the manual transcripts tend to better comply with the syntactic-prosodic mapping than the hypothesis, as the latter conveys errors.

4. RESULTS ON THE ASR ERRORS AND MORPHO-SYNTACTIC FEATURES

In this section, we first examine which POS and constituent categories are most prone to ASR errors (Section 4.1). We then turn to the distribution of errors across syntactic constituents (Section 4.2).

4.1. ASR errors at the word level: part-of-speech analysis

Extending the study of [11], the present investigations aim at getting a better understanding of which word types are hard to recognize. In Figure 3, we examine 6,287 lexical words (representing 51.7% of the words in our subset) and the distribution of the 5,438 grammatical words (44.7% of the words) according to their POS categories (cf. Table 2 for POS description). For the sake of simplicity, interjections (3.5% of the words) were excluded from this analysis. The abscissa groups the POS labels of lexical words followed by those of grammatical words. The left part of the figure gives absolute counts of erroneous (bad) and well recognized (good) words. The right figure shows the word error rates within POS classes.

One can observe that ASR errors in lexical words are particularly high (41% bad) for proper names (NPP) whereas common names (NC) tend to be well recognized (17% bad). This observation suggests that NPP tend to be particularly problematic for ASR systems. Note that the NPP class includes an important number of various foreign names – such as *Clint Eastwood* or *Safy Nebbou* – that may be either missing from the system's dictionary or be present with poor pronunciations, thus directly impacting ASR error rates. ADJ, ADV and NC categories seem to be similarly hard to be recognized, with respectively 20.9%, 18.7% and 16.9% word error rates.

Concerning grammatical words, weak clitics (CL), which tend to be very frequent but short words, are globally more problematic (30% bad) than the rest of grammatical words. In addition, when analysing the top 25 most frequent words in

our data (>100 occurrences), we find worse recognition for grammatical words (70.58%) than for lexical words (23.93%) and interjections (7.76%). This observation is consistent with previous results reported in [11, 12].

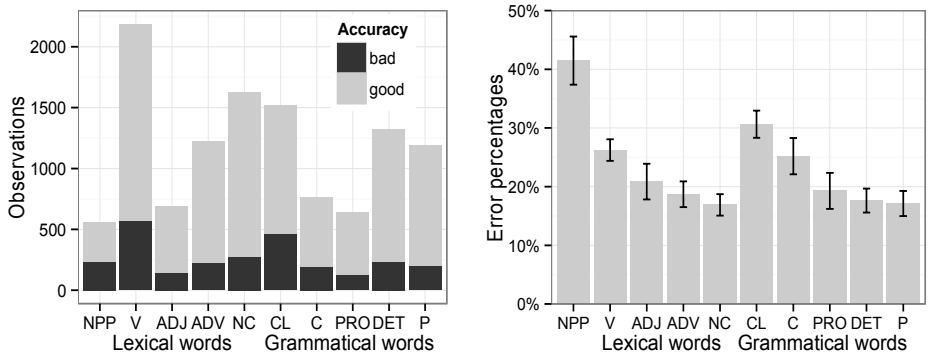


Figure 3. (a) Number of ASR errors according to POS categories. Lexical words: NPP = proper names, V= verbs, ADJ= adjectives, ADV = adverbs and NC = common names. Grammatical words: CL = week clitics, C = coordinates and conjunctions, PRO = pronouns, DET = determiners, P = prepositions. (b) Relative error rates as a function of POS category.

4.2. ASR errors at syntactic level: constituent analysis

Here, we investigate how errors appear across syntactic constituents. The 12,156 reference words retained for the current study are grouped into 7,567 syntactic constituents. Figure 4 shows the constituent distribution depending on the 12 distinct types as defined in the FST (see Table 3). Similarly to Figure 3, the abscissa of Figure 4 first gives the lexical headed constituents before the grammatical headed ones (constituents in which their head is not specified are also grouped under this label).

A high number of constituents include only one word, in particular in grammatical headed constituents. With respect to prosody, we selected syntactic structure types, which may contain several words. To form potential PWs, we further selected the following constituents in which the head is (i) a lexical word and (ii) is located at their right edges: Noun Phrases (NP), Verbal Nuclei (VN) and Adjectival Phrases (AP). Note that Infinitive Clauses (VPinf), Adverbial Phrases (AdP) and Nonfinite Clauses (VPpart) are also likely to form potential PWs. Yet, we discarded them due to their relatively low frequencies in our corpus. Note that the rest of syntactic constituents are not lexical headed and do not allow to form potential PWs (cf. mapping rules detailed in Section 3.3).

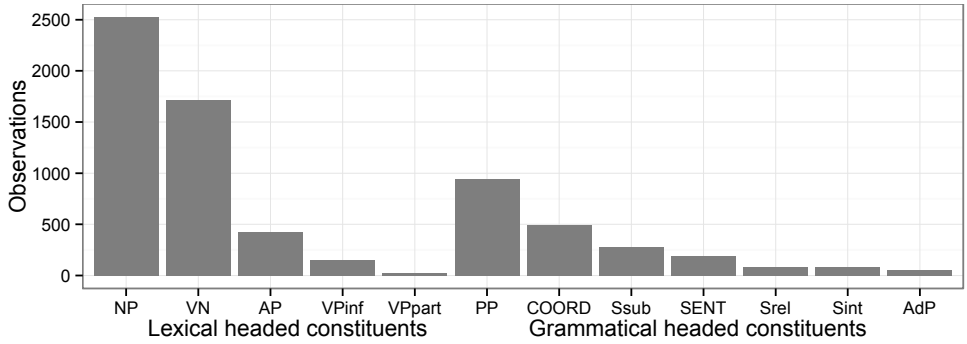
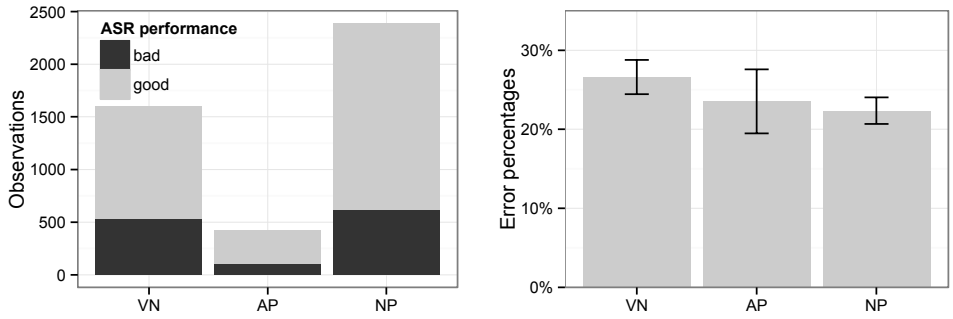


Figure 4. Frequencies of constituents extracted from the French Spoken Treebank. At the left, constituents having a lexical word as a head: NP = noun phrases, VN = verbal nuclei, AP = Adjectival Phrases, VPinf = infinitive clauses, VPpart = nonfinite clauses. At the right, constituents having a grammatical word or constituents in which the head is not specified: PP = prepositional phrases, COORD = coordinated phrases, Sint/Srel/Ssub = finite clauses, SENT = sentences.

Our data shows that 42% of these constituents include only 1 word (short XPs): de facto, in that case, this word is both the XP's syntactic head and is associated to one of the following POS categories: Verb, Adjective or Common Name/Proper Name. For instance, the clause *ça marche* ("it works") is parsed in two different XPs containing one word each being, at the same time, its own head: $[\text{ça}]_{\text{NP}}$ $[\text{marche}]_{\text{VN}}$. For XPs containing at least 2 words (long XPs), they are composed as follows: the XP's head, located at the end of the syntactic group, and all the material existing at their left side (essentially clitics). The longer XPs have almost 60% of its constituents with more than one word: 43% of them comprise 2 words, 11% 3 words, 3% 4 words and only 0.07% comprise more than 4 words.

First, we examined which syntactic groups are most prone to ASR errors. A contrast analysis between XPs with all words well recognized (XP_{good}) versus XPs with at least one word bad recognized (XP_{bad}) was carried out. Figure 5 shows the distribution of XP_{good} *vs*. XP_{bad} according to grammatical constituents: VN, AP and NP. In absolute terms (cf. Subfigure 5 (a)), there are about as many VN as NP constituents involved in ASR errors, APs being much less common. However, when looking at relative error rates within each syntactic group (Subfigure 5 (b)), it appears that VN are most prone errors: 26.6% of VN constituents include at least one ASR error, *versus* 23.5% for AP and 22.3% for NP.

Next, we explored a possible correlation between syntactic group durations and ASR errors. Previous studies have shown that ASR errors in spontaneous speech are related to faster produced speech and to speech reductions. Hence, we wanted to check whether durations of XP_{bad} were significantly different from those of XP_{good} and whether this pattern interacts with XP types.



(a) (b)

Figure 5: (a), number of occurrences of XPs_{bad} *vs*. XP_{good} per constituent type. (b), XP_{bad} rates per constituent type and 95% CI's. VN = Verbal Nuclei, AP = Adjective Phrases, NP = Noun Phrases.

Figure 6 shows the normalized durations of XP_{good} and XP_{bad} according to XP type. It appears that, generally speaking, VNs tend to be faster than NPs, which are faster than APs. Furthermore, VN_{bad} have significantly shorter durations than VN_{good}, but in AP and NP no significant duration differences can be observed. This result suggests that VP errors may be at least partially explained by temporal reduction.

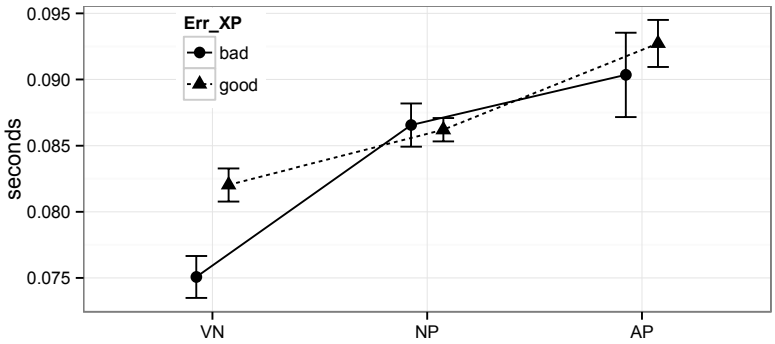


Figure 6. Normalized duration of XPs_{good} *vs*. XPs_{bad} according to their syntactic classification and 95% CI's. VN = Verbal Nuclei, AP = Adjective Phrases, NP = Noun Phrases.

4.3. Discussion

The analysis of word errors as a function of POS highlighted that proper names are particularly challenging for ASR systems (>40% errors). This can be explained both by system limitations (proper name unknown or with poor pronunciations) and by poor speaker consistency, especially for foreign proper names. For both POS and grammatical constituent analyses, results showed that

verbs tend to be more problematic than either adjectives or common nouns. For verbal nuclei including ASR errors, an acoustic correlate (low average phone duration) could be identified: VPs tend to be produced with lower durations than other categories be they adjectival or noun phrases. A possible hypothesis supporting VP shortening may be related to their position in French utterances: verbal nuclei tend to be in sentence non-final positions. They are often followed by one or several XPs and may thus prosodically group with following XPs.

5. TOWARDS AN ASR ERROR TYPOLOGY *VIA* SYNTACTIC-PROSODIC MAPPING

In this section, we examine to what extent both reference and hypothesis transcripts respect the syntax-prosody mapping explained in section 3. Our guess is that reference words of the manual transcripts tend to better comply with the syntactic-prosodic mapping than the hypothesis, as the latter conveys transcription errors. We check right constituent edges for presence/absence of PWs in the signal. We first evaluate to what extent predicted PWs are actually found in the signal using our reference transcriptions. We then investigate whether this mapping is violated when using hypothesis transcriptions, in particular in ASR error regions.

5.1. Evaluating the syntactic-prosodic mapping using reference transcriptions

The syntactic-prosodic mapping described in Section 3.3 allowed us to identify 4,415 potential PWs in the reference words (the total of VN, AP and NP constituents of this study). We then examined the acoustic signal for the presence of final pitch accents (positive delta f_0 values encoded as H*) in the rightmost full syllables of these constituents. Our acoustic measurements allowed us to label only 1,901 PWs. In other words, only 43% of the expected PWs were actually detected in the signal by the presence of a H*. Several explanations may be proposed: the H* criterion (2 st rise) may be too selective, XPs may be too short to consistently map WPs, spontaneous speech may include long stretches with flat pitch contours. We manually checked parts of our data. An example of the presence/absence of PWs in our data is given with the clause *certains films sont anciens* (‘some movies are old’). According to our mapping rules, the three constituents [certains **films**]$_{XP}$, [**sont**]$_{XP}$ [**anciens**]$_{XP}$ call for three PWs {certains films}$_{PW}$ {sont}$_{PW}$ {anciens}$_{PW}$. However, Figure 7 shows that only the third PW {anciens}$_{PW}$ was prosodically marked and a H* was detected (see the tier named PWs). However, at the end of

the expected PWs $\{\text{sont}\}_{\text{PW}}$ and $\{\text{films}\}_{\text{PW}}$ no increase of pitch could be observed indicating the absence of the first two PW's right boundaries (encoded as 0*).

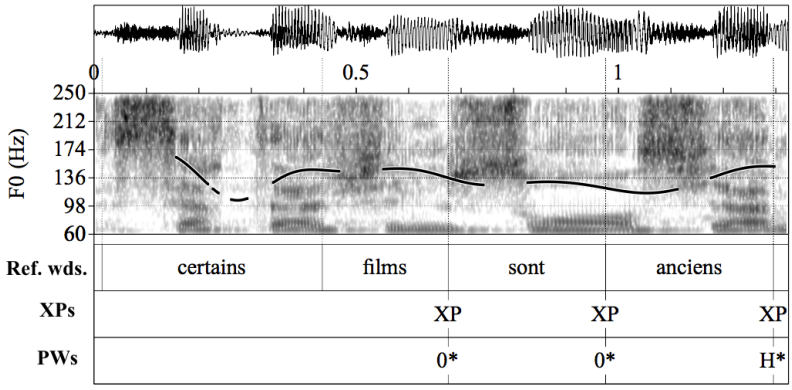


Figure 7. Spectrogram and f_0 curve of the clause *certains films sont anciens* ('some movies are old') parsed into 3 XPs and mapped to only one prosodic word.

As the rate of detected vs potential PWs is very low, we further develop possible reasons to explain the poor effectiveness of the mapping rules. The first one concerns speech rate: as stated in Section 3.3, fewer pitch accents tend to be produced in fast speech than in slow or read speech (Post, 2010). Since our data deal with spontaneous speech, the realization of pitch accents could decrease. A second explanation concerns rhythmic effects. Monosyllabic constituents do not tend to be produced as independent PWs in the signal. This may explain the absence of the predicted pitch accent in the word *sont* in the example above. We have to recall that (see Section 4.2), 43% of the constituents contain only one word (with a large share of monosyllabic words). It is thus not surprising that an important number of expected PWs was not prosodically marked. The relatively poor effectiveness of the mapping rules thus suggest that other constrains such as speech rate and XP length (in number of syllables) enter into play in PW groupings in spontaneous speech.

5.2. Do ASR errors respect syntax-prosody mapping?

What happens to syntax-prosody mapping rules when replacing reference transcriptions with ASR hypotheses? In particular, what happens in error regions where erroneous words entail different XP types with different XP boundaries? Since the hypothesis words were not grammatically tagged, the present analysis remains rather qualitative and requires a quantitative validation in the future.

We manually tagged a subset of the ASR hypotheses into FST like constituents. When applying the same syntactic-prosodic mapping rules to the hypothesis, we could observe that the right edges of XP constituents of the ASR hypothesis do not always align with H* labels, similarly to the above observations for reference transcripts. Can we find H* in positions violating the mapping rules? Such a mismatch example highlighting a H* in a wrong XP position of the ASR hypothesis is illustrated in Figure 8. First, concerning the reference sequence *j'ai râlé* ('I complained') a H* was produced at the end of the head *râlé* as predicted by the mapping rules. The syntactic group [j'ai râlé]$_{XP}$ was actually parsed into one PW {j'ai râlé}$_{PW}$, since the final syllable of the word *râlé* carries a H*. Now, when considering the corresponding hypothesis *jura les* ('(he/she/) swore the'), a mismatch arises between the syntactic constituents of the hypothesis with respect the prosodic pattern. The hypothesis words [(il/elle) jura]$_{XP}$ [les…]$_{XP}$ call for the following PWs: (il/elle) jura}$_{PW}$ {les…}$_{PW}$. The H* is wrongly aligned with the hypothesis word *les* violating the mapping rules (a wrong alignment is labelled as '!' in the PWs tier). In French, clitics like *les* are too rarely produced with a H* in this position.

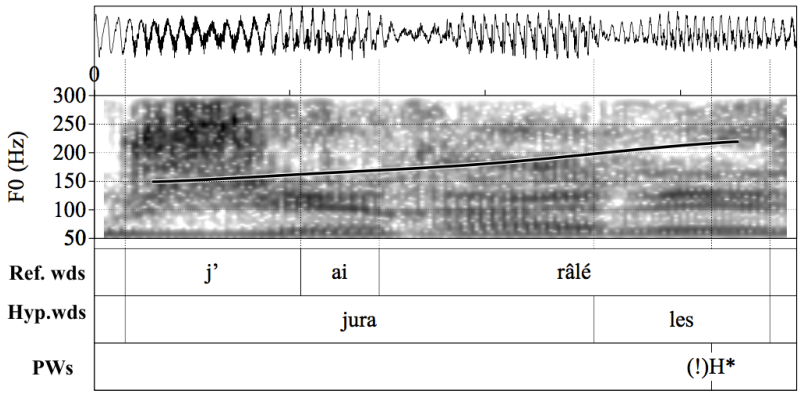


Figure 8. Spectrogram and f_0 curve of the clause *j'ai râlé* ("I complained") and the corresponding ASR hypothesis *jura les* ("(he/she) swore the").

Another example of syntax-prosody mapping violation in ASR hypotheses is illustrated in the sequence *(certains films sont peut-être) de fils et là* ('(some movies are probably) by sons and then'). According to the reference transcription, the right edges of the syntactic structures [de fils]$_{XP}$ [et là]$_{XP}$ ('[by sons]$_{XP}$ [and then]$_{XP}$') call for two PWs:{de fils}$_{AP}$ and {et là}$_{AP}$. In line with the mapping rules two H* can be found and align with the final syllables of the words *fils* and *là* (see Figure 9). However, when switching to the corresponding ASR hypothesis, the

syntactic structures [de ficelle]$_{XP}$ [à...]$_{XP}$ (‘[of string]$_{XP}$ [in…]$_{XP}$’) do no longer correctly match the same H* events. Such syntax-prosody mismatches in the hypothesis give clues to detect ASR transcription errors.

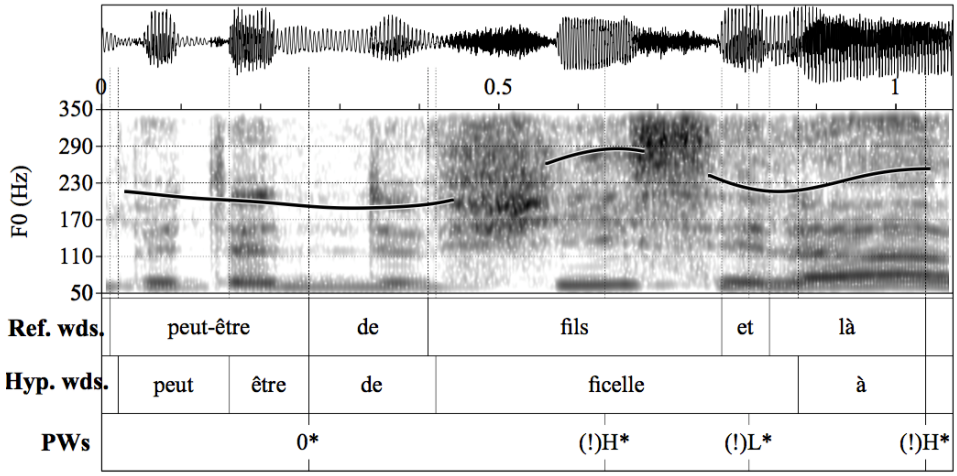


Figure 9. Spectrogram and f0 trace of the clause *de fils et là* (‘about children and then’) and the corresponding hypothesis *de ficelle à* (‘about string in’).

5.3. Discussion

The achieved results allow us to at least partially answer the question whether both reference and hypothesis transcripts respect the syntax-prosody mapping rules. Using reference transcriptions in Section 5.1, we found a relatively low percentage (43%) of XPs' right edges aligning with H* events of PWs in speech. In other words, there are far less pitch accents in the signal than there are predictions from the syntactic level. Possible explanations include a too fine-grained XP grouping (with too many mono-syllabic phrases) as well as speech rate and rhythmic factors in spontaneous speech. It is important to highlight that the absence of pitch accents is not considered to contradict the syntactic-prosodic mapping rules – the found pitch accents do align with the syntactic right edges of the reference words. This contrasts with the presults reported in Section 5.2 using ASR hypotheses. Hypothesis words (errors) may violate the syntactic-prosodic mapping, since wrong alignments emerge when mapping the H* pitch accents of the signal with wrong syntactic material. Such alignments may result in unnatural prosodic patterns with respect to French prosodic phrasing: H* are generally not admissible in grammatical words and/or in phrases’ prefinal positions.

6. CONCLUSION

In this study we made use of 1.5 hours of highly challenging France Inter spontaneous speech radio shows (*Le Masque et la Plume, Un Temps de Pauchon*) which were automatically transcribed by the LIUM speech recognition system. As part of the French Speech Treebank, the radio shows were also syntactically annotated. These data are hence extremely valuable to study the link between ASR errors and syntactic information and to further investigate syntax-prosody mapping rules and their effectiveness in spontaneous speech. A first line of investigations aimed at studying ASR errors with respect to syntactic information. The analysis of word errors as a function of POS highlighted that proper names are particularly challenging for ASR systems (> 40% errors). This can be explained both by system limitations (proper names may be unknown or with poor pronunciations) and by poor speaker consistency, especially for foreign proper names. Results showed that verbs tend to be more problematic for ASR systems than either adjectives or common nouns. For verbal nuclei, we could observe that ASR errors in verbs correlate with short durations, which was not the case for other categories such as adjectival or noun phrases.

With respect to syntax-prosody mapping, we first worked with manual reference transcriptions. We found a relatively low percentage of about 40% of XPs' right edges (nominal phrases, adjectival phrases and verbal nuclei) aligning with H* events in speech. In other words, there are far less pitch accents in the signal than there are predictions from the syntactic level. Possible explanations include a too fine-grained XP grouping (with too many mono-syllabic phrases) as well as speech rate and rhythmic factors in spontaneous speech. We then switched to ASR hypotheses including transcription errors, which may result in different syntactic groupings. The idea is to obtain clues towards a syntax-prosody error typology and towards automatically locating ASR errors from syntactic and prosodic points of view. Hypothesis words (errors) may violate the syntactic prosodic mapping rules, as wrong alignments may arise when mapping H* pitch accents with wrong syntactic material. They may result in prosodic patterns with violate French prosodic phrasing rules: H* are generally not admissible in clitics (grammatical words) located in the left periphery of the lexical words on which they depend.

In future studies, we will further investigate the various constraints ruling the mapping between pitch accents and syntactic constituents in spontaneous speech and investigate how prosodic contours in ASR errors may contribute to automatic error detection.

Acknowledgments. This work was funded by the French National Agency for Research (ANR) as part of the VERA project (adVanced ERrors Analysis for speech recognition) and supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083).

REFERENCES

[1] Shinozaki T. and Furui S., *Error analysis using decision trees in spontaneous presentation speech recognition*, in Proceedings of ASRU, 2001.

[2] Adda-Decker M., *De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux*, in Proceedings of JEP, 2006, pp. 389–400.

[3] Ernestus M. and Warner N. (eds), *Speech reduction* [Special Issue], Journal of Phonetics, 39, 2011.

[4] Holmes J. and Holmes W., *Speech Synthesis and Recognition* (2nd edition), London: Taylor & Francis, 2001.

[5] Fosler-Lussier E. and Morgan N., *Effects of speaking rate and word frequency on pronunciations in conversational speech*. Speech Communication, n° 29, pp. 137–158, 1999.

[6] Shriberg E.E., *Phonetic consequences of speech disfluency*", in Proceedings of ICPhS, 1999, pp. 619–622.

[7] Adda-Decker M., Habert B., Barras C., Adda G., Boula de Mareuil Ph. and Paroubek P., *A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models*, in Proceedings of DiSS workshop, 2003.

[8] Adda-Decker M. and Lamel L., *Do speech recognizers prefer female speakers?*, in *Proceedings of Interspeech*, 2005, pp. 100–104.

[9] Telaar D., Weiner J. and Schultz, T., *Error signatures to identify Errors in ASR in an unsupervised fashion*, in Proceedings of ERRARE, 2015.

[10] Huet S., Gravier G. and Sébillot P., *Morphosyntactic resources for automatic speech recognition*, in Proceedings of LREC, 2007.

[11] Goryainova M., Grouin C., Rosset S. and Vasilescu I., *Morpho-Syntactic Study of Errors from Speech Recognition System*, in Proceedings of LREC, 2014, pp. 3050–3056.

[12] Goldwater S., Jurafsky D. and Manning Ch.D., *Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates*, Speech Communication, n° 52, 2010.

[13] Abeillé A. and Crabbé B., *Vers un treebank du français parlé*, in Proceedings of TALN, 2013, pp. 174–187.

[14] Gravier G., Adda G., Paulsson N., Carré M., Giraudel A. and Galibert O., *The ETAPE corpus for the evaluation of speech-based TV content processing in the french language*, in Proceedings of LREC, 2012, pp. 114–118.

[15] Bougares F., Deléglise P., Estève Y., Rouvier M., *LIUM ASR system for ETAPE French evaluation campaign: experiments on system combination using open-source recognizers*, in Proceedings of TSD, 2013.

[16] Abeillé A., Clément L. and Toussenel F., *Building a treebank for French*. In Abeillé Anne (ed), Treebanks, Kluwer, Dordrecht, 2003.

[17] Gauvain J.-L., Lamel L., Schwenk H., Adda G., Chen L., and Lefèvre F., *Conversational telephone speech recognition*, in Proceedings of ICASSP, 2003, pp. 212–215.

[18] Jun S.-A. and Fougeron C., *The Realizations of the Accentual Phrase in French Intonation*, in Probus, n°14, pp 147–172, 2002.

[19] Post B., *Tonal and phrasal structures in French intonation*, Holland Academic Graphics, 2000.

[20] Boersma P. and Weenink D.D. *Praat, a system for doing phonetics by computer* [computer program]. V. 5.3.51, retrieved from http://www.praat.org/.

[21] Di Cristo A., *Intonation in French*, in Hirst Daniel and Di Cristo Albert, Intonation Systems: A survey of twenty Languages, Cambridge University Press, pp. 195–218, 1998.

[22] Post B., *The multi-faceted relation between phrasing and intonation in French*, In Gabriel Christoph & Lléo Conxita (eds), Intonational Phrasing at Interfaces: Cross-Linguistic and Bilingual Studies in Romance and Germanic, Amsterdam: John Benjamins, 2011, pp. 43–74.