Morpho-Syntactic Study of Errors from Speech Recognition System

Maria Goryainova[1,2] **Cyril Grouin**[1] **Sophie Rosset**[1] **Ioana Vasilescu**[1]
[1]CNRS, UPR 3251, LIMSI [2]INaLCO, EA 2520, ERTIM
91403 Orsay, France 75007 Paris, France
{grouin,rosset,ioana}@limsi.fr mariagrnv@gmail.com

Abstract

The study provides an original standpoint of the speech transcription errors by focusing on the morpho-syntactic features of the erroneous chunks and of the surrounding left and right context. The typology concerns the forms, the lemmas and the POS involved in erroneous chunks, and in the surrounding contexts. Comparison with error free contexts are also provided. The study is conducted on French. Morpho-syntactic analysis underlines that three main classes are particularly represented in the erroneous chunks: (i) grammatical words (to, of, the), (ii) auxiliary verbs (has, is), and (iii) modal verbs (should, must). Such items are widely encountered in the ASR outputs as frequent candidates to transcription errors. The analysis of the context points out that some left 3-grams contexts (e.g., repetitions, that is disfluencies, bracketing formulas such as "c'est", etc.) may be better predictors than others. Finally, the surface analysis conducted through a Levensthein distance analysis, highlighted that the most common distance is of 2 characters and mainly involves differences between inflected forms of a unique item.

Keywords: Automatic Speech Recognition; Error Analysis; Morpho-Syntactic Analysis

1. Introduction

Automatic speech recognition (ASR) systems take as input oral signal and produce as output a text version of the signal: the transcription of the signal in natural language. They make use of a speech model composed of acoustic, pronunciation and lexical n-gram models to decode the incoming speech stream. However the transcription process entails a number of errors, the speech model being able to handle at various level the ambiguity characterizing the spoken signal. Such ambiguity is due to various factors:

- Quality of the signal: low quality acquired from the telephone vs. high quality from radio news,
- Type of speech: prepared speech vs. spontaneous speech,
- Quality of speech: overlaps, fast speech due to stress or emotions, etc.,
- Out-Of-Vocabulary (OOV) items—especially foreign names—, etc.

In-depth description of the observing ASR errors are essential to characterize the variability intrinsic to the spoken language and to consider improved speech models (Adda-Decker and Lamel, 1999; Adda-Decker, 2006). Herein, instead of studying ASR errors causes, the current study focus on the surface features of the errors so as to produce classes of errors. Our long run objective is to anticipate the impact of error classes on further NLP processes. The study described here is based on the **work hypothesis** that both speech transcription errors and their surrounding contexts are predictive of the regions likely to be problematic for ASR. An in-depth characterization of such regions may help to efficiently adapt language models. Whereas most of the studies on ASR errors considered the phenomenon from lexical or phonetic standpoints, we focus here on the morpho-syntactic structure of the erroneous regions. In this purpose, we provide an original **morpho-syntactic** taxonomy of the ASR speech errors, so as to categorize the unrecognized chunks as well as their larger contexts. Comparisons with error free contexts are also provided for an in-depth comprehension of local conditions inducing speech ambiguity and penalizing the ASR system. The following taxonomy would be also of valuable for various domains linked to automatic speech recognition such as speech understanding, named-entity recognition, question/answering systems, etc.

2. Related work

ASR transcription errors highlight speech regions which are problematic with respect to the ASR system's decoding capacities. ASR errors have been mainly investigated in the framework of comparisons between automatic vs. human decoding of speech (Scharenborg, 2007; Lippmann, 1997). They pointed out that although today best ASR speech models are quite efficient, they have not yet reached the status of being able to perfectly take into account all observed acoustic variation, human listeners still outperforming them 5 to 6 times better (Vasilescu et al., 2012). The taxonomy of errors pointed out that some words are frequently subjects to ASR errors: in particular short, acoustically poor and frequent items lead to local ambiguity (Adda-Decker, 2006).

The homophony is particularly challenging for ASR systems, as underlined in (Vasilescu et al., 2012): such lexical items are both problematic for ASR systems and human listeners. Although a rich literature analyzed errors from the perspective of the ASR vs. human (in)capacities in decoding spoken signal, there is a lack of studies which consider the morpho-syntactic patterns of erroneous contexts.

In the next sections, we propose a preliminary analysis of the morpho-syntactic characteristics of the errors in French compared with the error free contexts. Most largely we aim at investigating the global morpho-syntactic characteristics of a corpus of spoken data used in the French ANR ETAPE project (Gravier et al., 2012) and of the ambiguous regions which lead to erroneous ASR transcriptions.

3. Material and methods

3.1. Corpora

The study is based on a textual corpus in French consisting of manual and automatic transcriptions. The data were gathered in the framework of the ETAPE project (Gravier et al., 2012) and correspond to different audio sources manually and automatically transcribed by an ASR system (Bougares et al., 2013). A general description of the corpus is given in Table 1.

Genre	Source	# words	# distinct words	# Error spans
TV News	BS	64318	6946	3851
	TQ	21896	3644	834
TV Debates	CVR	49033	5187	4374
	ELL	54267	5670	4081
	PF	43990	4481	2302
TV Amusement	PDV	2049	2645	3489
Radio shows	FrDeb	142417	12508	13752

Table 1: General description of the corpus depending on the source: BS ("BFM Story"), TQ ("Top Question"), CVR ("Ça Vous Regarde"), ELL ("Entre les Lignes"), PF ("Pile ou Face"), PDV ("La Place du Village"), French Debate (FrDeb)

Each sentence from each transcription has been aligned with the manual reference. The alignment highlights the *error spans* (Luzzati et al., 2014). An *error span* is defined as all the consecutive words in the hypothesis which are different from the reference. The error span level has been adopted in the current study (e.g., in contrast to the word level).

Figure 1 illustrates an extract from the data. For a given sentence, three levels of information are provided and considered in the analysis: the reference transcription (REF), the automatic transcription made by the system (HYP) and the description of the types of errors within each span (that is D=deletion, I=insertion, S=substitution).

```
REF:  <IL> y a <IL> y a quatre <VINGT > mille <******>
HYP:  <**> y a <**> y a quatre <VINGTS> mille <CHIENS>
EVAL: <D >     <D >            <S     >       <I     >

REF:  <BONSOIR À TOUTES ET TOUS MERCI> d' être avec
HYP:  <******* * ****** ** **** *****> d' être avec
EVAL: <D       D D      D  D    D    >
```

Figure 1: Extract from the aligned corpus

Global statistics on the corpus show that 21% of tokens are involved in an error span. The errors consist either in a substitution (49%), a deletion (35%) or an insertion (16%).

3.2. Methods

3.2.1. Presentation

This study is based on the hypothesis that erroneous spoken regions and their surrounding contexts convey some salient and predictive information about potentially ambiguous chunks for ASR. To gather as much as possible information about such chunks we are conducting an in-depth morpho-syntactic analysis. Three levels are then considered: (i) basic morpho-syntactic analysis (token, lemma, POS), (ii) contextual analysis (within the error span, on the left or on the right of the error span), and (iii) analysis based upon the edition distance of characters (distance frequency and POS involved).

The morpho-syntactic tagging has been made with the Tree Tagger (Schmid, 1994). This tool is known to be poorly adapted to process speech transcriptions which may involve repetitions and erroneous solutions. Nevertheless, for this preliminary study we make use of tagged data without a post-processing phase, in order to avoid potential new errors.

3.2.2. Issues

Morpho-syntactic analysis. The morpho-syntactic analysis is aimed to provide insights about the forms, lemmas and POS occurring in an erroneous chunk. The following questions have been addressed.

- Which forms, lemmas and POS are the most frequent in error spans?
- Which forms, lemmas and POS obtain the higher error percentage out of the whole corpus?
- Does the most frequent POS in the error spans represent the most frequent lemmas and forms?
- Does the form that achieve the highest error percentage belongs to the POS and the lemma that also achieve the highest error percentage?

Contextual analysis. Alongside with the error span analysis a contextual similar investigation have been also conducted to answer to the following points:

- Which n-gram and POS sequences are the most frequent on the left and right context of the error spans?
- Does the most frequent n-gram sequences correspond to the most frequent POS?
- Which kind of sequence precedes/follows an error span?
- Which kind of semantic information can we infer from an error span?

Surface analysis. At last, the edition distance is conducted to estimate the mean number of modifications to process from an erroneous string to a correct string, and the most frequent POS concerned with high distance editions.

4. Results

4.1. Error span analysis

In this section an overview of the main results is provided. Table 2 underlines the frequencies inside an error span in comparison with the frequencies in the whole corpus for some of the most frequent forms found in an error span.

Form	Corpus	Error span	Ratio
y (there)	3,016	891	29.5%
il (he)	6,526	1,871	28.7%
c' (it)	7,307	1,737	23.8%
qu' (that)	3,588	844	23.5%
est (is)	11,288	2,395	21.2%
on (one/we)	5,800	1,215	20.9%
a (has)	5,550	1,162	20.9%
je (I)	4,138	866	20.9%
ça (this/that)	3,070	630	20.5%
et (and)	7,965	1,615	20.3%
à (to)	6,628	1,041	15.7%
le (the, masculine)	8,264	1,210	14.6%
en (in)	4,778	637	13.3%
pas (no)	4,567	593	13.0%
de (of)	15,237	1,814	11.9%
les (the, plural)	6,142	667	10.9%

Table 2: Frequencies in the whole corpus and inside an error span for some of the most frequent forms in an error span

Table 3 highlights the frequencies inside an error span in comparison with the frequencies in the whole corpus for some of the most common POS.[1]

POS	Corpus	Error span	Ratio
Pro:per	37,442	8,858	23.7%
Ver:pres	39,996	8,027	20.1%
Conj	22,581	4,418	19.6%
Pro:dem	14,209	2,739	19.3%
Subst	84,873	13,490	15.9%
Pro:rel	9,547	1,347	14.1%
Adj	23,162	3,235	14.0%
Adv	27,852	3,827	13.7%
Ver:infi	10,603	1,340	12.6%
Det:art	35,242	3,778	10.7%
Prep	37,089	3,787	10.2%
Name[2]	108	10	9.3%

Table 3: Frequencies in the whole corpus and inside an error span for some of the most frequent POS in an error span

4.2. Contextual analysis

4.2.1. Left context of an error span

Sequences of forms. In this sections we focus on the contextual analysis. The context is viewed here as the forms, lemmas and POS at left and right sides of an erroneous span. It is analyzed increasingly (from one item left/right to 3 items left/right) as to evaluate the impact of the increasing surrounding information in erroneous chunks prediction. Table 4 provides the frequencies of sequences of one, two or three forms in the whole corpus and in the left context of an error span.

[1]We used the following POS abbreviations: Adj (adjective), Adv (adverb), Conj (conjunction), Det:art (article), Name (proper name), Pro:dem (demonstrative pronoun), Pro:per (personal pronoun), Pro:rel (relative pronoun), Subst (substantive), Ver:info (verb at infinitive), Ver:pres (verb at present tense).

Sequence (1, 2, 3 forms)	Corpus	Context	Ratio
bien ... (good/well)	1,153	152	13.2%
très ... (very)	1,165	104	8.9%
de ... (of)	13,423	859	6.4%
le le ... (the the)	224	41	18.3%
très bien ... (very good)	119	19	16.0%
c' est ... (it is)	5,086	467	9.2%
tous les départements ... (all the departements)	6	3	50.0%
Roche sur Foron ...	19	7	36.8%
est à dire ... (is to say)	207	20	9.7%
c' est pas ... (it is not)	352	33	9.4%

Table 4: Frequencies of sequences of 1, 2 or 3 forms in the whole corpus and in the left context of an error span

Sequences of POS. Table 5 underlines the frequencies of sequences of one, two or three Part-of-Speech in the whole corpus and in the left context of an error span.

Sequence (1, 2, 3 POS)	Corpus	Context	Ratio
Adv ...	27,852	2,957	10.6%
Pro:rel ...	9,547	678	7.1%
Adv Adv ...	2,546	297	11.7%
Det:art Adj ...	3,188	250	9.1%
Subst Pro:rel ...	3,660	245	6.7%
Pro:rel Ver:pres Adv ...	297	33	11.1%
Det:art Subst Adj ...	3,431	378	11.0%
Pro:per Pro:per Ver:pres ...	3,747	294	7.8%

Table 5: Frequencies of sequences of 1, 2 or 3 POS in the whole corpus and in the left context of an error span

4.2.2. Right context of an error span

Sequences of forms. Table 6 underlines the frequencies of sequences of one, two or three forms in the whole corpus and in the right context of an error span.

Sequence (1, 2, 3 forms)	Corpus	Context	Ratio
... bien (good/well)	1,153	200	17.3%
... par (by)	1,358	162	11.9%
... je (I)	3,272	438	13.4%
... des des (the the)	210	24	11.4%
... par le (by the)	128	12	9.4%
... parce que (because)	710	66	9.3%
... de la semaine (of the week)	46	8	17.4%
... qui est un (who is a)	46	8	17.4%
... c' est c' (it is it)	240	41	17.1%
... il y a (there is)	1,095	151	13.8%

Table 6: Frequencies of sequences of 1, 2 or 3 forms in the whole corpus and in the right context of an error span

Sequences of POS. Table 7 underlines the frequencies of sequences of one, two or three Part-of-Speech in the whole corpus and in the right context of an error span.

Sequence (1, 2, 3 POS)	Corpus	Context	Ratio
... Prep	37,089	2,860	7.7%
... Prep Prep	1,169	138	11.8%
... Pro:per Pro:per	6,021	564	9.4%
... Pro:dem Pro:rel	1,884	177	9.4%
... Pro:dem Pro:rel Pro:per	921	97	10.5%
... Prep Det:art Subst	7,716	597	7.7%
... Pro:rel Pro:per Ver:pres	1,960	112	5.7%

Table 7: Frequencies of sequences of 1, 2 or 3 POS in the whole corpus and in the right context of an error span

4.3. Surface analysis

Figure 2 shows the edition distance according to Levenshtein's algorithm (Levenshtein, 1965), in terms of characters between the correct and the erroneous forms in an error span.

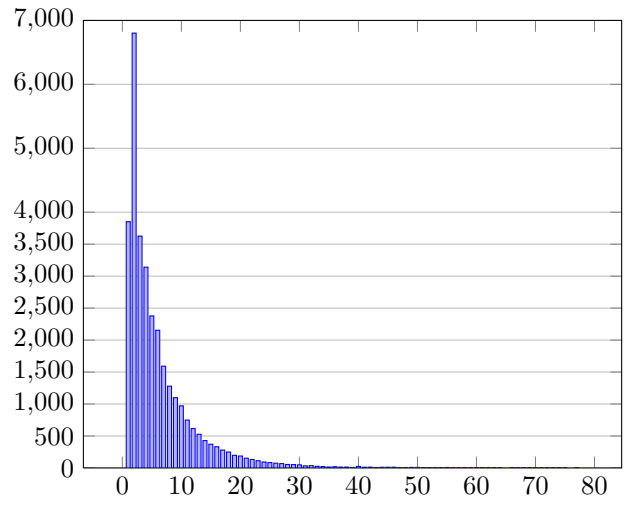

Figure 2: Edition distance in characters between erroneous and correct forms

5. Discussion

5.1. Forms in error spans

In this section some trends highlighted by the morpho-syntactic analysis are mentioned.

As shown in Table 2, the most frequent forms in the error spans correspond to grammatical words that is short, acoustically poor and subject to homophony items, which are good candidates to recognition errors as underlined in (Adda-Decker, 2006).

While the most frequent POS category in the corpus is substantives (20.74% of all categories[3]), followed by personal pronouns (9.15%) and prepositions (9.06%), three main classes are particularly represented in the erroneous chunks: (i) grammatical words *(to, of, the)*, (ii) auxiliary verbs *(has, is)*, and (iii) modal verbs *(should, must)*. The table confirms the trends highlighted by the analysis of word frequencies (see Table 2) and by previous studies both in ASR output analysis and comparison with humans (Vasilescu et al., 2012).

However, it is worth noticing that some word classes occur more frequently in an error span than others: for instance, the category of the proper names is frequently unrecognized by the system (9.3% of proper names are within an error span).

Paronymes are also good candidates to ASR errors:

- "*il*" [il] (he) / "*y*" [i] (there)
- "*et*" [e] (and) / "*est*" [ɛ] (is)
- "*a*" [a] (has) / "*à*" [a] (to)
- "*un*" [ɛ̃] (a, one) / "*en*" [ɑ̃] (in) / "*on*" [ɔ̃] (one)

5.2. Contexts

One may notice that the salient information is provided by (at least) two items in particular at the left side of the error span. The left 3-grams suggest that some contexts (e.g. repetitions, that is disfluencies, bracketing formulas such as "c'est" etc.) may be better predictors than others (see Table 4). Among the most frequent sequences at the left side of an error span, several syntagms have a bracketing role, that is they introduce information (*c' est/c' est pas, it is/it is not, est à dire*, is to say). Speech disfluencies and in particular repetitions may also occur (*les les*, the the). Such phenomena are spontaneous speech proper.

At last, the contexts also involved (more frequent) substantives and proper names (*départements*, departments, *Roche sur Foron*[4]).

The analysis of the left and right contexts points out that grammatical words are the most frequent neighbors of an erroneous span ((Table 5 and 7). They also occur within disfluent regions suggesting that speakers' difficulties in building the verbal message may involve less accurate pronunciations and then errors.

The same contextual analysis conducted in terms of POS (Table 5 and 7) underline the high frequency of short words, potential candidates to transcription errors.

Finally, one may notice the similarity between erroneous and error free regions close to the erroneous spans: the most frequent items present in erroneous span are also present as surrounding context suggesting that a "fragile" chunk in terms of morpho-syntactic characteristics may anticipate an error.

5.3. Surface analysis

Finally, concerning the surface analysis conducted through a Levensthein distance analysis, the maximum distance is of 256 characters, which corresponds to a deletion of a whole sequence. The most common distance is a distance of 2 characters, which mainly involves inflection differences between an infinitive and a past participle in French or between singular and plural of names and adjectives. This finding suggest that the most frequent errors in French do not necessarily affect the content of the message.

[3]This percentage corresponds to the 84,873 substantives found in the corpus out of the total number of 409,185 tokens.

[4]*La Roche-sur-Foron* is the name of a town in Savoie, France.

6. Conclusion

Nowadays ASR systems reached high levels of accuracy, however speech transcription errors still occur. Several studies on the speech transcription errors have been conducted during the last decade, mainly focusing on the frequency of the lexical items concerned and on the acoustic patterns of the items likely to by unrecognized. In our paper we provide an original standpoint of the phenomenon by focusing on the **morpho-syntactic features** of the erroneous chunks and of the surrounding left and right contexts. The typology concerns the forms, the lemmas and the POS involved in erroneous chunks, and in the surrounding contexts. Comparison with error free contexts are also provided. The study is conducted on French. Findings comfort previous observations about the presence of grammatical words among the most frequent missrecognized items. Results also underline the presence of such items before and after an erroneous span as well as the presence of "fragile" contexts (e.g., disfluences) as predictors of erroneous regions. However, the analysis of surface forms (Levensthein analysis) points out the high frequency of errors of level 2 (2 characters difference) which correspond to inflection differences. The long run aim is to make use of such investigation as to improve language models. Similar work will also be conducted on different corpora and languages as to lead to an in-depth comprehension of the speech transcription errors.

7. Acknowledgements

This work was supported by the French National Agency for Research as part of the project VERA (adVanced ERrors Analysis for speech recognition) under grant ANR-2012-BS02-006-04.

8. References

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

Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Proc of JEP*, Dinard, France.

Bougares, F., Deléglise, P., Estève, Y., and Rouvier, M. (2013). LIUM ASR system for ETAPE French evaluation campaign: experiments on system combination using open-source recognizers. In *Sixteenth International Conference on TEXT, SPEECH and DIALOGUE (TSD 2013)*, Pilsen, Czech Republic.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A., and Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *Proc of LREC*, Istanbul, Turkey.

Levenshtein, V. (1965). Binary codes capable of correction deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.

Luzzati, D., Grouin, C., Vasilescu, I., Adda-Decker, M., Bilinski, E., Camelin, N., Kahn, J., Lailler, C., Lamel, L., and Rosset, S. (2014). Human annotation of asr error regions: Is "gravity" a sharable concept for human annotators? In *Proc of LREC*, Reykjavik, Iceland. ELRA.

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language*.

Vasilescu, I., Adda-Decker, M., and Lamel, L. (2012). Cross-lingual studies of ASR errors: paradigms for perceptual evaluations. In *Proc of LREC*, Istanbul, Turkey.