A Benchmark of French ASR Systems Based on Error Severity

Antoine Tholly, Jane Wottawa, Mickaël Rouvier, Richard Dufour

A Benchmark of French ASR Systems Based on Error Severity

Antoine Tholly[1], **Jane Wottawa**[2], **Mickael Rouvier**[3], **Richard Dufour**[1]
[1]LS2N, Nantes Université, France
[2]LIUM, Le Mans Université, France
[3]LIA, Avignon Université, France
antoine.tholly@univ-nantes.fr, jane.wottawa@univ-lemans.fr,
mickael.rouvier@univ-avignon.fr, richard.dufour@univ-nantes.fr

Abstract

Automatic Speech Recognition (ASR) transcription errors are commonly assessed using metrics that compare them with a reference transcription, such as Word Error Rate (WER), which measures spelling deviations from the reference, or semantic score-based metrics. However, these approaches often overlook what is understandable to humans when interpreting transcription errors. To address this limitation, a new evaluation is proposed that categorizes errors into four levels of severity, further divided into subtypes, based on objective linguistic criteria, contextual patterns, and the use of content words as the unit of analysis. This metric is applied to a benchmark of 10 state-of-the-art ASR systems on French language, encompassing both HMM-based and end-to-end models. Our findings reveal the strengths and weaknesses of each system, identifying those that provide the most comfortable reading experience for users.

1 Introduction

Every linguistic description is based on theoretical assumptions, whether acknowledged or not. This study is influenced by contextual linguistics (Rastier and Riemer, 2015; Col et al., 2012), in which the meaning of a word is interpreted not only by its form (spelling/morphology, *e.g.*, the five letters of "plane"), but also by the other words in its semantic network (*i.e.*, words belonging to the same categories and/or the same sequence). The human recognition process considers the surrounding grammatical and lexical/thematic *context* in which the word – whether correct or erroneous – is used; even more so if the word is erroneous.

In the context of Automatic Speech Recognition (ASR), there are two aspects to the interpretation of transcription errors: *detection* (*e.g.*, "pla" is an incorrect spelling) and *resolution* (*e.g.*, the reference "plane" might be reconstructed from "pla"). Consequently, we propose that ASR error detection and resolution should be analyzed *within the context of their sentences*. This relationship is bidirectional: an error can *import its solution* from the context (*e.g.*, "pla" understood as "plane" within the aeronautical theme of a sentence), and it can *export its problem* to the context (e.g., blurring or altering the overall meaning of the sentence).

Traditional error metrics such as Word Error Rate (WER) tend to ignore context. Objectively, a transcription error can be defined as a deviation from the reference. However, this perspective does not account for how users interpret errors solely through the transcription. In the *detection phase*, spelling issues (*e.g.*, "a pla" instead of "a plane") are seen as errors in all contexts, unlike context-dependent semantic errors (*e.g.*, "the plate lands on the tarmac"). The same applies to the *resolution phase* of errors: context may or may not be needed to retrieve the reference, and if needed, it may be adequate or insufficient. Furthermore, errors also have various *consequences* for the understanding of the contextual sequence (*e.g.*, an easily detectable and resolvable error like "choocolate" does not significantly affect other parts of the sentence).

In this paper, we propose degrees of *error severity* from an interpretative perspective in French, alongside an original study of textual automatic transcription errors from various ASR architectures. Section 2 introduces four classes of errors that reflect the severity of errors from the user's perspective. Errors can have minimal impact when the word is immediately recognizable, such as minor spelling issues without semantic consequences (Section 2.1). Grammatical errors on content words, while not impeding comprehension, contravene established social norms (Section 2.2). Moderate to significant errors require additional effort, as they can only be understood through contextual processing, which may be challenging (Sec-

tion 2.3). The final type of error is critical, as it completely disrupts understanding due to ambiguity, contextually uninterpretable words or deletions, or undetectable errors when the mistake matches the sentence meaning and syntactic structure (Section 2.4). Boundary cases and individual variations are also addressed (Section 2.5). Section 3 outlines the protocol using a French corpus and 10 ASR systems differing in modularity, self-supervised learning (SSL) audio models, tokenizers, and training data volume. These systems are benchmarked in Section 4 based on the four error categories, leading to a discussion of the system's capabilities. Section 5 concludes and suggests future perspectives.

2 A User-based Metric

The metric aims to rank systems based on their ability to prevent different types of errors, defined according to their perceived severity for humans. This human-centered interpretative criterion contrasts with the purely formal, spelling-deviation criteria commonly employed by metrics such as WER and CER (Character Error Rate). These formal metrics are limited to detecting errors and offer no insight into their semantic distance from the reference—that is, the interpretative effort required to resolve them, if resolution is possible.

Expanding on these limitations, semantic metrics measure the semantic distance between the hypothesis and the reference (Zhang et al., 2019; Kim et al., 2021), but do not examine the underlying causes of this distance. Considering these causes, we argue that this distance is smallest when context is *superfluous* for error interpretation, greater when context is *necessary*, and greatest when context is *insufficient* or *misleading*.

This semantic distinction also guided the decision to use lexical/content words (nouns, adjectives, verbs, adverbs) as the unit of measurement within the metric. These words provide a simple, homogeneous criterion for objectivity in measurement, while carrying significant information at the word level, essential for interpretative relevance. Although other linguistic units, such as grammatical/function words and higher-level structures like phrases and clauses, are important, combining all linguistic levels where errors occur would not result in a comprehensive metric or clear interpretative criteria. Additionally, grammatical constructions are partially reflected in lexical word errors—first in their inflectional parts, and second, when phrases or clauses are severely distorted, as the core lexical words are interpreted through the distortions they reflect.

The proposed four-category error system is designed to account for all lexical word errors in the corpus, ensuring comprehensive, broad, and objective coverage while minimizing ambiguity during categorization (cf. Section 2.5). These primary categories are further divided into detailed subcategories, which are outlined and exemplified in their respective sections.

To illustrate distinctions between error types and subtypes, variations of a single example, 'gorilla,' are used as a pedagogical tool to facilitate focused comparison. Appendix A provides a classified sample of errors from the French corpus, accompanied by their English translations.

The four error types forming the metric will be revisited in Section 4 (Results) to highlight their insights into the performance of the benchmarked ASR systems.

2.1 Minimal Impact: Immediate Word Recognition (Lex)

The least severe category (comprehension-based) regroups cases where the identity of the word is successfully recognized without relying on context (*gorila*, *patato*, *advennture*).

In the 'Lexical' category, error occurs in the stem part of the word, i.e. without additional issues in the inflectional part (e.g., "two gorila" is to be classified in the 'Grammatical' category). This is similar to how speakers may recognize words they don't know from another related language, such as an English speaker understanding the French lexeme 'compétitivité'.

A variation of this error involves internal segmentation (or splitting) issues, where the word remains recognizable in isolation, such as "a gor illa" or "a go rilla".

As a clarification, although the context is not required for the solution, it still plays a role in word interpretation (as it always does), particularly because the immediate solution, out-of-context, could still be incorrect (in the case of a 'gorila', the reference would generally be 'gorilla' but could still be another word, like 'guerilla'. Cf. Section 2.4).

2.2 Special Disruption: Grammatical Botheration (Gram)

As with the previous type of error, the identity of the word is easily recognized without relying on

context, but it contains an error in the inflection, e.g., 'one gorillas.' In the French corpus, this pertains to gender and plural markers on nouns and adjectives, as well as tense and person markers on verbs.

This category reflects the specific role of grammar in both ASR processing and the 'botheration' reactions of end-users. 'Botheration' is a concept from psycholinguistics, referring to errors that do not affect comprehension but violate established norms and conventions. (Boettger and Emory Moore, 2018; Smith, 2015).

2.3 Moderate to Significant Difficulty: Effort Requirements from Context Processing (Cotx)

Erroneous words can be resolved fully or partially with the help of contextual words or structures, but this process requires greater cognitive effort, often at the expense of the end-user's reading comfort. Three subtypes of solution recognition are outlined below.

Local group recognition involves identifying multiword expressions and named entities (e.g., "magilla gor" for the "Magilla Gorilla" cartoon character), as well as contiguous collocations (e.g., "a goril sanctuary").

Broader context recognition is triggered by lexical/thematic cues, such as 'my favorite animal is the gori', or by grammatical triggers like anaphora, as in 'there's a gorilla, and this gril [...]'.

Partial recognition occurs when context provides limited information, such as identifying an animal ('my favorite animal is the gr') or an agent-like entity ('I talked to this gr'). Partial understanding often edges toward the critical error category.

2.4 Critical Miscommunication: Non-Understandable Errors (Fail)

The inability to retrieve the solution represents the most severe class of errors.

Hesitation or ambiguity arises when multiple solutions could apply, as in "[*no additional context*] a gril is an animal" where it's unclear whether the term refers to 'grizzly' or 'gorilla'. A second case occurs when a unique contemplated solution seems uncertain: the user recognizes 'gorilla' from a spelling error, but is not sure if he can rely on his interpretation.

The user can also acknowledge an impossible interpretation. The recognition of an impasse in the resolution process can be triggered by several factors: i) uninterpretable word-forms, even with context, such as "there is a gorallo", ii) certain ungrammatical sentences, e.g. "there is a * that" (a missing word is inferred from the incomplete noun phrase), and iii) certain meaningless associations between the context and a (seemingly) correctly spelled word (e.g., "the polarization of semiconductor gorilla"): while this error is detected through lexical semantics incompatibility, no solution is identified.

Finally, misleading interpretations (or false positives) occur when the error is undetectable, such as i) a lexical substitution that fits the context, like "guerilla" instead of "gorilla" in "a guerilla in the forest", or ii) a deletion of a syntactically optional element, as in "I'm in the forest", where the optional adjunct "with a gorilla" is omitted and cannot be perceived.

2.5 Remarks: Boundary Cases and Individual Variations

Our categories are clearly delineated for the purposes of the upcoming quantification, grouping the variety of subtypes around a central criterion. However, there is undoubtedly a continuum between errors that can be resolved through context ('Cotx') and those that cannot ('Fail'), and we have already mentioned that partial resolutions often lean toward critical errors. In practice, we do not believe that errors at the edges of categories are fundamentally a problem, for two reasons: firstly, in the distribution of errors into their respective categories (intra-system quantification of errors), they represent infrequent occurrences; and secondly, in the task of benchmarking (inter-system quantification of errors), they primarily need to be addressed consistently, i.e., in the same way across all systems.

Additionally, there are certainly individual variations in resolution abilities, which this typology and the forthcoming quantification appear to overlook. Here again, a clear response can be provided by refining the definitions. Context-resolvable errors ('Cotx') are expected to be resolved by most readers, whereas context-irresolvable errors ('Fail') are expected to remain unresolved by most readers (with the rarer boundary cases likely dividing interpretations). As our categories ground human interpretation on objective linguistic cues, we anticipate that a clear majority of human interpretations will align with these objective linguistic cues. Furthermore, variations between individuals are expected to be similar across systems, rendering them

neutral with respect to the benchmark.

3 Experimental protocol

The transcription corpus is derived from the REPERE corpus, which consists of French-language French television programs on news events, such as politics and culture. (Giraudel et al., 2012). The phonostyle (de Mareüil, 2014) can be described as public speaking by communication professionals, in prepared or semi-prepared formats, such as studio interviews and scripted delivery by journalists, and recorded under excellent audio conditions. REPERE had already segmented each broadcast into short audio sequences and expertly transcribed the audio of each sequence (i.e., the reference). We processed these audio segments with 10 different ASR systems to create a corpus of transcriptions. For this study, we used a subset of this corpus, specifically transcriptions of four different broadcasts.

All lexical words in this transcription corpus were classified either as correct (i.e., corresponding to the reference) or as one of the four distinct types of errors described in Section 2. These analyses were conducted by a linguistic expert through Glozz, an annotation tool developed for expert annotation in text corpora (Widlöcher and Mathet, 2012). The annotation of the 10 ASR transcriptions from the four broadcasts resulted in a total of 10,007 lexical words (1,125 annotated errors), with overall lexical word accuracy rates ranging from 79.6% to 93.7% across the broadcasts (averaged across the 10 systems).

This study evaluated 10 ASR systems from the Kaldi (Povey et al., 2011) and SpeechBrain (Ravanelli et al., 2024) toolkits, using various speech recognition methodologies. Two systems are DNN-HMM systems based on Kaldi, while 8 are end-to-end systems from SpeechBrain using various techniques such as SSL (self-supervised learning) audio models or tokenizers (see Table 1).

Systems based on Kaldi are prefixed with KD. The KD_wR and KD_woR systems used a 3-gram language model for decoding, but KD_wR also includes an additional posterior rescoring step based on the RNNLM deep neural network language model.

Systems based on SpeechBrain are prefixed with SB. The SB_no_char, SB_XLSR_char, SB_XLSRFR_char, SB_LB1k_char, SB_LB3k_char, and SB_LB7k_char systems use a character tokenizer, while SB_LB3k_bpe750 and SB_LB3k_bpe1000 a Byte Pair Encoder (BPE) tokenizer. All the systems, except SB_no_char, used an SSL audio model. SB_XLSR_char and SB_XLSRFR_char used the XLS-R model (cross-lingual speech representation based on wav2vec 2.0) whereas the others used the LeBenchmark models (French speech representation based on wav2vec 2.0) (Parcollet et al., 2024). We note that LeBenchmark 1k, 3k, and 7k are pre-trained on 1k, 3k, and 7k hours of unlabeled data, respectively.

All ASR systems were trained on French data using various corpora (ESTER 1 and 2 (Galliano et al., 2006, 2009), EPAC, ETAPE (Gravier et al., 2012), REPERE (Giraudel et al., 2012), and internal data), totaling about 940 hours of audio.

Systems	SSL Audio	Tokenizer
SB_no_char	No	Character
SB_XLSR_char	XLS-R	Character
SB_XLSRFR_char	XLS-R FR	Character
SB_LB1k_char	LeBenchmark 1k	Character
SB_LB3k_char	LeBenchmark 3k	Character
SB_LB7k_char	LeBenchmark 7k	Character
SB_LB3k_bpe750	LeBenchmark 3k	BPE 750
SB_LB3k_bpe1000	LeBenchmark 3k	BPE 1000

Table 1: Systems overview with different SSL audio models and tokenizers.

4 Results

In total, 1,125 lexical words errors are classified and quantified across 10 systems, along with 8,882 correct lexical words. Each system transcribed an identical corpus. The statistics are organized into 4 categories of errors, described in the metric section (Section 2). 'All' refers to the total of errors (in percentage compared to the total of lexical words).

Percentages of errors for each system regarding each category are presented in Table 2. The systems are ranked from best to worst on the vertical axis, taking into account the total rate of errors and giving greater weight to Fail errors.
In Table 2, we observe that:

Immediate Lexical Recognition errors (Lex). Kaldi systems achieve fewer 'Lex' errors compared to SpeechBrain systems. This can be attributed to the use of a language model, which helps avoid hallucinations of words, unlike the SpeechBrain systems.

Grammatical errors (Gram). The Kaldi system without rescoring achieves the second-lowest score, while the Kaldi system with rescoring

achieves one of the best scores. This improvement is due to the RNNLM used as an additional posterior rescoring step, which considers a broader context, allowing for better error correction. Interestingly, LeBenchmark models achieve results equivalent to Kaldi with rescoring, better than the XLS-R model. This indicates that training an audio SSL model in the target language effectively captures this type of information.

Systems	(WER)	*All*	Lex	Gram	Cotx	Fail
Total	18.77	11.8	2.1	2.1	2.2	5.3
KD_wR	13.21	5.4	0.5	1.5	0.2	3.2
SB_LB7k_char	16.56	7.0	2.0	1.3	1.6	2.2
SB_LB3k_bpe750	15.33	8.4	2.6	1.6	1.8	2.5
SB_LB3k_bpe1000	15.98	8.5	2.4	1.5	2.2	2.5
SB_LB3k_char	17.16	9.5	3.0	2.2	2.0	2.4
KD_woR	15.43	7.6	0.3	3.1	0.3	3.9
SB_LB1k_char	18.94	10.8	2.2	1.9	2.1	4.6
SB_XLSR_char	22.69	14.9	3.0	2.0	4.6	5.3
SB_XLSRFR_char	21.48	16.6	3.6	2.6	4.0	6.4
SB_no_char	30.94	23.4	2.2	3.4	3.9	14.0

Table 2: Error rates for each ASR system across categories, color-coded: white for lowest errors, light grey for moderate errors, and dark grey for highest errors.

Contextual errors (Cotx). Kaldi systems achieve better results due to their language model. We observe that the LeBenchmark models, trained on a large amount of data, can manage to correct these errors. Additionally, BPE tokenizers yield better results than character tokenizers.

Failure errors (Fail). The best results are obtained with LeBenchmark models using BPE tokenizers, as well as LeBenchmark models with character tokenizers and ample training data. Among these LeBenchmark models with character tokenizers, a notable reduction in the 'Fail' error rate is observed as the training data increases from 1K (4.6% error rate) to 3K (2.4% error rate). Kaldi models demonstrate moderate performance. Regarding SLL Audio, all LeBenchmark models outperform XLR-S models. The Seq2Seq model without SLL Audio (i.e., SB_no_char) performs the worst, as it also did with Gram and Cotx errors and overall: this is consistent with previous observations, as this system does not integrate a language model (and is limited to characters) while having no rich acoustic information.

Closing Analysis. The Kaldi system with rescoring achieves the best overall performance, as reflected in the 'All' error rates. However, the LeBenchmark model with character tokenizers and 7K of training data, while second overall, ranks slightly stronger than the former in addressing the most critical errors ('Fail' error rates). It is worth noting that LeBenchmark models with BPE tokenizers also demonstrate commendable performance overall, despite being limited to 3K of training data.

WER comparison. To assess these figures, we compare them with the WER results from the entire REPERE corpus, providing a broader evaluation of system error performance. The key comparison is not the absolute error rate (WER rate is higher): grammatical words are excluded from our experiment, and it only suggests that the audio subset used for the experiment was easier to process. The critical metric lies in the deviation among systems: our results and the WER results follow similar trends, offering strong evidence of our method's reliability. More importantly, our study introduces finer-grained dimensions for benchmarking these systems, extending beyond what the automatic metric alone can measure, as outlined above.

Statistical Relevance. For this corpus, the percentage difference that is statistically significant between two error rates has been calculated to be approximately 1.7%. Between previously discussed system differences, this significance threshold is usually exceeded.

5 Conclusions and Perspectives

A new typology of transcription error severity in ASR systems is proposed, based on a model of their reception by end users. This approach allows for both qualitative and quantitative analysis of errors from 10 French ASR systems, highlighting varying capabilities depending on the used architectures. Through this analysis, the method itself proves effective in the ASR benchmarking task.

This original study is of interest for improving ASR performances: the integration of a user's interpretative model provides valuable feedback, helping align ASR systems more closely with user expectations. Furthermore, the theoretical explanations based on English errors, combined with the performance analysis of French errors, demonstrate the applicability of this benchmarking method to multiple languages.

Looking ahead, severity criteria will be refined through a perception test evaluating how participants perceive ASR errors. This will help correlate audience assessments of error severity with error frequency across systems, leading to a more detailed and comprehensive performance evaluation.

6 Limitations

Single Expert. Only one linguistic expert was involved in the annotation process, which may introduce a bias. The interpretation quality was prioritized at this stage, focusing on consistency in human interpretation. The methodology aimed to mirror end-user perspectives to some extent, but further work could involve additional linguistic experts and user-centered validation to enhance objectivity and reliability.

Benchmarking and Data Scope. Although the corpus contained about 10,000 lexical words and 1,200 errors, only about 120 errors per system were ultimately categorized and benchmarked. This could limit the breadth of the comparison.

Additionally, finer-grained errors that had been annotated were finally included in larger categories of quantification, due to their scarcity, further constraining the scope of the benchmarking. Future evaluations will benefit from expanding the size of the corpus to increase the number of quantified categories and the statistical relevance of the results.

Acknowledgements

This research received funding from the Agence Nationale de la Recherche (ANR), France, as part of the DIETS project (Automatic Diagnosis of Errors in End-to-End Speech Transcription Systems from the User's Perspective).

We also wish to thank Thibault Bañeras-Roux for his informal contribution.

References

Ryan K Boettger and Lindsay Emory Moore. 2018. Analyzing error perception and recognition among professional communication practitioners and academics. *Business and Professional Communication Quarterly*, 81(4):462–484.

Gilles Col, Jeanne Aptekman, Stéphanie Girault, and Thierry Poibeau. 2012. Gestalt compositionality and instruction-based meaning construction. *Cognitive Processing*, 13:151–170.

Philippe Boula de Mareüil. 2014. Qu'est-ce qu'un (phono) style. *Cahiers de linguistique française*, 31:9–19.

Sylvain Galliano, Edouard Geoffrois, Guillaume Gravier, Jean-François Bonastre, Djamel Mostefa, and Khalid Choukri. 2006. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*, pages 139–142.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. 2012. The repere corpus: a multimodal corpus for person recognition. In *LREC*, pages 1102–1107.

Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier Galibert. 2012. The etape corpus for the evaluation of speech-based tv content processing in the french language. In *LREC-Eighth international conference on Language Resources and Evaluation*, page na.

Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L Seltzer. 2021. Semantic distance: A new metric for asr performance analysis towards spoken language understanding. *arXiv preprint arXiv:2104.02138*.

Titouan Parcollet, Ha Nguyen, Solène Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, et al. 2024. Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, 86:101622.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

François Rastier and Nick Riemer. 2015. Interpretative semantics. In *The routledge handbook of semantics*, pages 491–506. Routledge.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaelle Laperriere, Mickael Rouvier, Renato De Mori, and Yannick Esteve. 2024. Open-source conversational ai with speechbrain 1.0. *Preprint*, arXiv:2407.00463.

Sara D Smith. 2015. *Botheration and Recognition of Prescriptive Rules*. Brigham Young University.

Antoine Widlöcher and Yann Mathet. 2012. The glozz platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering*, pages 171–180.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix: Sample of Error Types and Subtypes.

Formatting of examples. Errors are highlighted in **bold**, followed by their reference in [brackets]. Where applicable, context words serving as resolution cues are underlined. Missing words or morphemes are indicated by an asterisk (*).

A.1 (LEX) Minimal Impact: Immediate Word Recognition.

Subtype 1.1: Minor distortion of word stem, preserving comprehension.

"syndic**t**ats" [syndicats] — (*unions*).

"comp**a**titivité" [comp**é**titivité] — (*competitiveness*).

Subtype 1.2: Word segmentation error, preserving comprehension.

"la pât**ie n**oire" [la patinoire] — (*the ice skating rink*).

"une l**e ç**on" [une leçon] — (*a lesson*).

A.2 (GRAM) Special Disruption: Grammatical Botheration.

Subtype 2.1: Verbal inflection error.

Tense. "le comité qui a organis**ai**" [organis**é**] — (*the committee that organized*).

Person. "qu'il renonçai* à" [renonçai**t**] — (*that he gave up on)*.

Subtype 2.2: Nominal and adjectival inflection error.

Gender. "au palmarès importante" [important] — (*with an important track record*).

Number. "les papys rockeur*" [rockeur**s**] — (*the rocking grandpa**[s]).

A.3 (CTX) Moderate to Significant Difficulty: Effort Requirements from Context Processing.

Subtype 3.1: Local context resolution: multiword expressions.

Compounds. "une **nombre** [onde] de choc" — (*a shock**whale*** [shockwave]).

Contiguous collocations. "la viande **a lal** [halal]" — (***a lal*** [halal] *meat*).

Named entities. "valérie **pécrese**" [Pécres**s**e] .

Subtype 3.2: Broader context resolution: lexical/thematic or syntactic comprehension cues.

Lexical relations. "ressembler à la **gresse** [Grèce] et à l'espagne" — (*to resemble **gresse*** [Greece] *and spain*). — Lexical relation: "Spain"

and "Greece" ("gresse") are *coordinate terms* (European countries).

Lexical properties. "il y a des **sars** [stars] mais alors après il y a aussi beaucoup de femmes qui sont beaucoup moins connues" — (*there are **sars*** [stars]*, but then after that there are also many women who are much less famous*). — Interpretation: "famous" is a *defining characteristics* of "stars" ("sars").

Lexical fields. "dites moi euh vous êtes spécialisé dans les **tracs** [tracts] de l'ump parce qu'il y a aussi des **tracs** [tracts] du ps qui expliquent l'inverse." — *(tell me uh you are specialized in **tracs*** [tracts/leaflets] *of the ump because there are also **tracs*** [tracts/leaflets] *of the ps that explain the opposite)*, UMP and PS being political parties — Lexical field: *politics*.

Syntactic structure. "monsieur bayrou **je vous** [joue] les prophètes" — (*monsieur bayrou **I you*** [plays] *the prophet*) — Syntactic cues: the sequence is detected as ungrammatical (1 noun phrase + 2 pronouns + 1 noun phrase). The word order suggests the following resolution: 1 noun phrase subject + 1 verb + 1 noun phrase object, all the more so as the two pronouns "je" (/ʒə/) + "vous" (/vu/) bear a phonetic resemblance to the verb "joue" (/ʒu/).

Anaphora. "pendant cinq ans la droite monsieur fillon et monsieur sarkozy monsieur **skozy** [Sarkozy] d'abord monsieur fillon ensuite" — (*for five years the right party monsieur fillon and monsieur sarkozy monsieur **skozy*** [Sarkozy] *first then monsieur fillon*).

Subtype 3.3: Partial context resolution: limited comprehension cues.

"à l'occasion de la sortie en salle de **france canouni** [Frankenweenie], le réalisateur propose une exposition autour de la création de ce film d'animation" — (*on the occasion of the theatrical release of **france canouni*** [Frankenweenie]*, the director is offering an exhibition about the creation of this animated film*) — Partial interpretation: the error refers to the name of an animated film just released, but its identity remains unclear.

A.4 (FAIL) Critical Miscommunication: Non-Understandable Errors.

Subtype 4.1: Ambiguity.

Uncertainty between competing solutions. "on vient de nous expliquer que trois milliards serait un **denjeu** [enjeu] national non ça n'est pas sérieux" — (*we were just told that three billion would be a national **danjeu** no that's not serious*) — Ambiguous interpretation: does "denjeu" [dãjø] refer to "danger" [dãje] (*danger*) or to "enjeu" [ãjø] (*stake*)?

Hesitation to accept a considered solution. "je **re guemarque** [remarque] là-dedans que pour revenir à l'équilibre [...]" — *I **nog otice*** [notice] *in this that to return to balance* [...]) — Ambiguous interpretation: could "I nog otice" mean "I notice"?

Subtype 4.2: Acknowledged impossible interpretation.

A detected but unsolvable form distortion. "un univers **wason** [foisonnant] que les Parisiens pourront découvrir" — *(a **tymnge*** [teeming] *universe that Parisians will be able to discover).*

A detected but unsolvable lexical incompatibility with contextual meaning. "ceux qui ont créé le désastre brandissent l'épouvantail **lagos** [de la gauche]" — (those who created the disaster are brandishing the **lagos** looming threat *[the looming threat of the Left]*).

A detected but unsolvable syntactic deletion. "qu'un premier ministre * [dise] si la gauche passe la zone euro va s'effondrer" — (*that a prime minister ** [would say] *that if the left comes to power the eurozone will collapse*), "dise"/"would say" being a missing word.

Subtype 4.3: Misleading interpretation (false positives).

Undetectable lexical substitution. "on va nous réunir à treize heures trente de la sauvette pour présenter le **problème** [programme] de stabilité de la France" — (*we'll be meeting at one thirty in a rush to present France's stability **problem*** [program]).

Undetectable deletion of an optional syntactic element. " * [merci] Dominique de Montvalon et alors France Soir peut-être pourrait renaître de ses cendres on verra" — (* [thank you] *Dominique de Montvalon and then France Soir might rise from its ashes we'll see*).