The Added Value of Metadata and Annotations: Evidence from Two Large-Scale, Naturalistic Corpus Studies

Anisia Popescu, Johanna Cronenberg, Ioana Vasilescu, Ioana Chitoran, Lori Lamel, Martine Adda-Decker
TransCrit, Université Paris 8;
LPP, Sorbonne Nouvelle & CNRS;
LISN, Université Paris-Saclay & CNRS;
LLF, Université Paris Cité & CNRS;
LISN, Université Paris-Saclay & CNRS & Vocapia Research;
LPP, Sorbonne Nouvelle & CNRS
anisia.popescu@univ-paris8.fr, johanna.cronenberg@uni-koeln.de, ioana.vasilescu@lisn.fr,
ioana.chitoran@u-paris.fr, lamel@lisn.fr, martine.adda-decker@sorbonne-nouvelle.fr

Abstract

This paper presents two case studies that highlight both the challenges and benefits of working with large-scale, naturalistic phonetic data. Our aim is to encourage researchers not to shy away from phonetic data found "in the wild", even when such data are messy, noisy, or incomplete – because they can yield robust, novel insights beyond the reach of controlled laboratory studies. We focus on challenges that are endemic to large corpora, including degraded audio quality, sparse or inconsistent annotations, and missing speaker metadata. By comparing two corpus-based studies that diverge in methodology and statistical design, we show how different approaches can mitigate these limitations while still extracting meaningful patterns.

Keywords: Large-Scale Speech Corpora, Importance of Metadata, Explained Variance Evaluation

1. Introduction

The phonetic sciences are an inherently data-driven discipline, drawing evidence from both small, tightly controlled laboratory studies and large, heterogeneous datasets. Historically, phonetics has relied heavily on acoustic data collected in laboratory experiments designed for specific research questions, where recording conditions, speaker samples, and speech materials are carefully regulated. This approach offers the clear advantage of allowing precisely phrased hypotheses to be tested under optimal conditions in terms of the phonetic viability of the data. Well-described experimental designs also allow for the replication of studies which is an essential scientific endeavour. At the same time, laboratory settings and corpora tailored to specific research questions often yield data that are limited in size and speaker diversity – participant pools, for instance, often consist of university students – and may not fully capture the variability of natural speech. The overwhelming focus on laboratory phonetic studies, however, was at least also partially due to limits in computational power and requirements such as perfectly balanced data for certain statistical tests.

What laboratory studies lack in data quantity, naturalness and heterogeneity, large corpora make up for by capturing a wide range of natural speech patterns across diverse contexts and speakers. Large corpora in language sciences are a combination of (1) carefully designed and documented collections such as those featured in the "Mining a Year of Speech" project (Coleman et al., 2011), (2) data collected for speech technology purposes (e.g., automatic speech recognition (ASR)) and (3) the increasingly common use of web and social media data. These collections provide a large volume and a large variety of speech data. Curated collections provide rich metadata that is of great value for linguistic research. For instance, the British National Corpus (BNC) (Consortium, 2007), the German Oral Corpus (Corpus Gesprochenes Deutsch (CGD)), and the Switchboard Corpus (Godfrey and Holliman, 1993) include demographics such as age, gender, social class, and/or regional dialects. Similarly, the CHILDES corpus (MacWhinney, 2000) provides rich metadata on children and their caregivers, making it invaluable for language acquisition studies. These curated corpora enable nuanced sociolinguistic analyses that are not feasible with most ASR training datasets and data from media platforms because they lack detailed speaker information.

It would, however, be of great value to the phonetic sciences to create pathways that render these data usable for research for three main reasons: first because the amount of such data has exponentially increased in the last decades, documenting the respective current state of spoken language (Puggaard-Rode et al., 2022; Harrington et al., 2000; Sonderegger et al., 2023); second because these are rich sources of speech recordings capturing extensive speech variation across lan-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5767–5775
11–16 May 2026.

guages, speakers, and linguistic conditions (Ahn and Chodroff, 2022; Black, 2019; Popescu et al., 2023; Chodroff and Zhang, preprint; Cronenberg et al., 2024); and third, because recordings found "in the wild" (Cohn and Renwick, 2019) represent speech in its most unaltered, natural form which is impossible to recreate in the lab (Liberman, 2019; Cohn and Renwick, 2019; Labov, 1972). In this paper, we will carefully weigh the challenges and benefits of working with large, naturalistic speech corpora, but we aim to ultimately encourage researchers to consider using such data when appropriate and we provide insights into possible strategies for metadata enrichment.

We draw from two case studies both of which used large quantities of data from media archives, but differ in terms of the investigated phenomenon, the methodological approach, and the language. The first study uses acoustic measurements to investigate the duration reduction of the hiatus /ia/ in naturalistic European Portuguese as a result of lexical stress and position within the word. The second study is a sociophonetic investigation of a well-established phonetic reduction phenomenon in French: the deletion of post-obstruent /ʁ/ in word-final position (e.g., *livre* [livʁ] 'book' produced as [liv]). Unlike the first study, this analysis relies exclusively on the output of an ASR system, which distinguished between canonical ([livʁ]) and non-canonical ([liv]) pronunciation variants. The detailed results of the two studies are beyond the scope of this paper and can be found in Popescu et al. (2025); Cronenberg et al. (In prep). Here, we focus instead on the processing required to prepare the data for analysis, and examine how each processing step influenced the goodness-of-fit and adequacy of regression models. We argue that scaling up phonetic research requires automatic and manual metadata augmentation as well as the strategic use of resources.

2. Data Resources and Processing

Media archives hold an abundance of audio data from thousands of speakers across diverse communicative settings. However, these recordings are typically not collected for linguistic research purposes, resulting in several challenges: variable audio quality, sparse or inconsistent annotations, and missing speaker metadata. The two datasets analyzed in this study – Portuguese and French broadcast corpora – illustrate these issues clearly.

The Portuguese data consists of 114 hours of television and radio recordings from both news and interview shows (Fraga-Silva et al., 2013, 2011), while the French data combines 100 hours from the ESTER corpus (Galliano et al., 2006) and 50 hours from the ETAPE corpus (Gravier et al., 2012). Both were originally extracted and processed by speech technologists to train ASR systems (Vasilescu et al., 2020; Lamel and Gauvain, 1992). Whereas ASR development prioritizes quantity and diversity of data, linguistic analyses require higher control, consistency, and detailed metadata.

In order to render these corpora suitable for acoustic-phonetic and sociophonetic analyses, we had to implement a series of data filtering and augmentation steps. These included (i) the identification and exclusion of noisy recordings as well as a manual check of segment boundaries, (ii) data normalization, (iii) speaker diarization, and (iv) the linguistic annotation and augmentation of speaker metadata. These time- and labor-intensive processing steps will be described in the next sections before we evaluate their effectiveness in terms of how much they improved the fit of regression models to the respective data.

2.1. Data Cleaning and Optimization

Word and phone-level segmentations (Lamel and Gauvain, 1992; Adda-Decker and Lamel, 1999) were available for the corpora used in both studies. For the French corpora, previous studies reported a Cohen's Kappa coefficient (Cohen, 1960) of 0.832 (Wu, 2018), indicating a very high level of agreement between manual and automatic segmentations. In addition, the first author manually inspected a small subset of the ASR output consisting of 300 tokens that had been automatically classified as non-canonical (deleted /r/). The manual verification showed that 98% of the tokens were correctly identified as deleted. The high accuracy level most likely results from the fact that the baseline dictionaries used to train the French acoustic models included the phonologically motivated variants that captured the well-known phenomenon of word-final /r/ deletion.

For Portuguese, the duration – based on the automatic segmentations – was extracted for all 9941 /ia/ sequences that did not occur in absolute word-initial or word-final position. Upon listening to the Portuguese data for the first time, it became apparent that the audio quality varied widely. The second author of this study listened to all audio files and labelled them for background noise (1406 files), background music (1138 files), and sub-optimal audio quality, e.g. a lot of reverb or multiple speakers talking over each other (565 files). Without these noisy audio files, the dataset contained 6789 audio files which corresponded to 6828 tokens, or 68.7% of the original tokens. The next step was to manually correct the boundaries of the /ia/ sequences to make sure that the duration measurements were accurate. The manually and automatically determined start and end of the /ia/ sequences differed by less than 15 ms for 79% of the non-noisy tokens.

The disagreement between manual and automatic alignment was more than 30 ms in less than 5% of the tokens. Similarly to the French case, the forced-aligner used on the Portuguese data was trained on parts of said data (Fraga-Silva et al., 2013, 2011), which explains its high accuracy. During this manual review of the alignment, another 43 tokens were found to be of sub-optimal audio quality, and they were hence excluded. The final, clean dataset consisted of 6785 tokens of /ia/ which occurred in 1170 distinct words.

2.2. Data Normalization

The automatic alignment used on the Portuguese data assigned each phonetic segment a duration in number of frames, where each frame was equal to 10 ms. That is, all duration values based on the automatically aligned segment boundaries were multiples of 10, thus creating a distribution of discrete rather than continuous values. This would have presented an additional challenge for the statistical analysis, but duration values are typically normalized for speech rate differences which, as a side effect, also solved this issue. To make sure that the duration was not biased by varying speech rates, we calculated the articulation rate as the number of phonemes per second (excluding silence, hesitations, as well as the duration of the vowel sequence itself) for each utterance that contained an /ia/ token. The duration measurements were divided by the respective articulation rate for each token. Finally, to get rid of the right-leaning skew that is typical for duration distributions, we transformed the normalized duration values with the logarithm. The resulting distributions of the log. normalized durations are shown in Figure 1 for the noisy (9941 tokens, uncorrected segment boundaries) and clean datasets (6785 tokens, corrected segment boundaries).

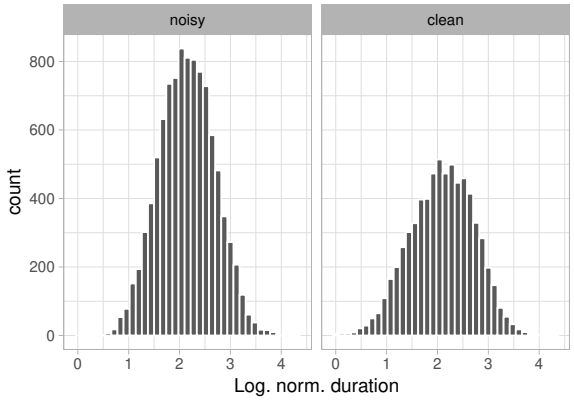

Figure 1: Distribution of log. normalized duration of /ia/ sequences in the noisy and clean datasets.

2.3. Speaker Diarization

One of the main challenges in working with both corpora was the lack of reliable speaker IDs. While some information on the speaker ID was available, most of the data only contained generic identifiers, where a single ID could refer to multiple speakers and, conversely, a single speaker could appear under multiple IDs. The two case studies addressed this issue in different ways. For the Portuguese study, the second author listened to the audio files in chronological order and performed an auditory speaker ID assignment which resulted in 3,658 new speaker IDs in the noisy dataset (2,422 new IDs for the clean dataset) compared to the original 201 generic ones. This is likely an overestimation of the number of distinct speakers, but it reduced the risk of conflating different speakers under the same ID, thereby preserving the independence assumption of linear regressions. In the French study, by contrast, extending previous work on the corpus by Wu et al. (2022), the first author constructed a subcorpus limited to speakers with verifiable identities, such as public figures (e.g., politicians, journalists, actors, athletes). This filtering process reduced the dataset from 1,802 speaker IDs and over 26,000 /r/ tokens to 391 verified speakers producing approximately 14,000 tokens, as illustrated in Figure 2.

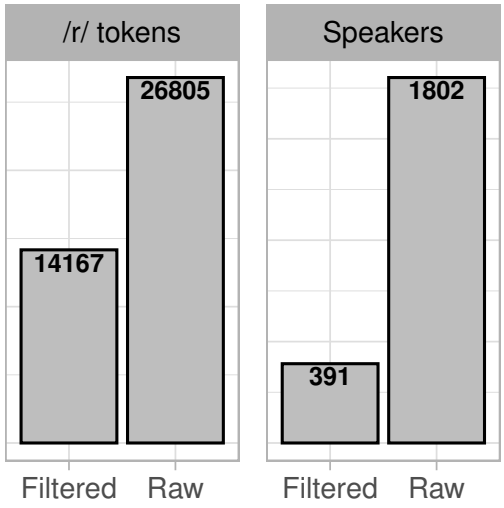

Figure 2: Comparison of filtered and raw counts of speakers and /r/ tokens in the French dataset.

Unreliable speaker identification in large-scale corpora presents a key methodological challenge for linguistic studies. Researchers must choose between labor-intensive perceptual verification which preserves more data but introduces subjectivity, or restricting analyses to recordings with confirmed

identities, which ensures reliability but reduces corpus size. At the same time, adding speaker information to large corpora can present an ethical issue if the speakers did not give their written consent to being identified in their speech recordings. In the cases presented here, we took great care to minimize this ethical concern. The true identities of the Portuguese speakers are still unknown to us; the auditory speaker diarization was based only on the perceived similarity of voices. The 391 verified French speakers, on the other hand, are adult public figures who were aware of being recorded for a broadcast (which often introduces them by name; also cf. other phonetic studies of public figures, e.g. (Harrington et al., 2000; Quené, 2013; Hay et al., 1999)).

2.4. Linguistic Annotation and Metadata Augmentation

The two studies addressed distinct research questions, requiring different linguistic annotations and additional metadata. In the Portuguese study, the proximity of /ia/ to the stressed syllable was an important analytical factor. All 1,487 unique word types containing this hiatus were therefore manually labelled for stress position, with partial reference to the Portuguese Stress Lexicon by Garcia (2017). The position of /ia/ within the word, also referred to as initiality, was determined algorithmically based on the canonical phonetic transcription: /ia/ was labelled as "initial" when preceded by a maximum of two consonants, and as "medial" otherwise.

In the French study, the sociophonetic focus required additional speaker metadata, including age, gender and socioeconomic status (SES). As the retained speakers were public figures, this information was sourced online. Profession, grouped into 18 different categories, was used as a proxy for SES. Since post-obstruent /r/ deletion can vary by speech style, part of speech (POS) and phonetic context, POS tags, speaking style and right-phone contexts were also added for all 14,167 tokens. The latter two parameters were derived directly from the corpus information, using automated extraction methods, without relying on external information.

Together, these annotations and metadata enabled consistent, fine-grained analyses for both studies. In order to demonstrate the effects of each processing step and additional information on the statistical analyses of the data, we ran incremental, nested regression models and evaluated how well these models fit the data. Section 3 explains the construction of the models for both case studies and section 4 presents the goodness-of-fit results.

3. Statistical Analysis

Differences in research focus motivated distinct statistical approaches: linear mixed-effects regressions were used for Portuguese to determine to what extent the /ia/ duration was influenced by proximity-to-stress and initiality, and logistic mixed-effects models were applied to French to identify factors affecting the realization of canonical versus non-canonical pronunciation variants. For each study, we have run multiple different models (in R, using the `lmerTest` package, version 3.1.3, Kuznetsova et al. (2017)), as summarized in Tables 1 (Portuguese) and 2 (French).

For Portuguese, the LMERs were run either on the noisy or clean data, with or without proximity-to-stress as a fixed factor, with or without a random intercept by speaker, and durations based on the automatic or manually corrected segmentation. All LMERs had the log. normalized duration as the dependent variable, initiality as a fixed factor, an interaction between proximity-to-stress and initiality when proximity-to-stress was included as a fixed factor (Models 2-5), and a random intercept by word. The first two models violate the assumption that the datapoints are independent because they lack the random intercept by speaker. For all models, the normality and homoscedasticity assumptions for the residuals were visually confirmed to hold.

Model	Data	Stress	Spk IDs	Seg.
1	noisy	no	no	auto.
2	noisy	yes	no	auto.
3	noisy	yes	yes	auto.
4	clean	yes	yes	auto.
5	clean	yes	yes	man.

Table 1: List of the five LMER models on the Portuguese data.

For French, four nested models were run. Model A included the right phone context and speech style (formal vs. conversational) as fixed factors – both derived directly from the corpus. Models B, C, and D each added one additional predictor: part of speech (POS), speaker profession, and speaker age category (< 40 vs. > 40), respectively. All models included speaker ID and word as random factors. As no significant effect of gender was found, the factor was excluded from the present analysis.

To assess the models' goodness-of-fit, both studies computed the conditional, marginal, and semi-partial R^2s using the `partR2` package (version 0.9.2; Stoffel et al. (2021)). The marginal R^2 is the proportion of variance in the dependent variable explained by the fixed factors, while the conditional R^2 captures the explained variance of the fixed and random effects together. The semi-partial

Predictors	A	B	C	D
right phone	yes	yes	yes	yes
speech style	yes	yes	yes	yes
POS	no	yes	yes	yes
profession	no	no	yes	yes
age	no	no	no	yes

Table 2: List of the four nested French mixed-effects logistic regression models.

R^2s are the proportion of variance explained separately for the fixed factors which also takes into account each factor's contribution to the interaction between them, if there is an interaction.

4. Results

Results of the R^2 analysis are shown in Figure 3 (Portuguese) and Figure 4 (French).

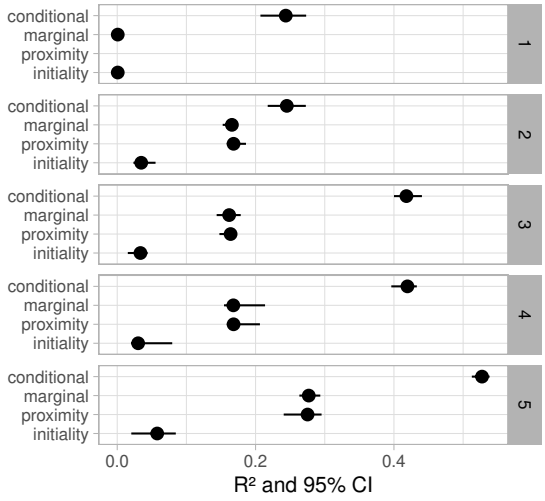

Figure 3: Conditional and marginal R^2s as well as semi-partial R^2s for proximity-to-stress and initiality, including their 95% confidence intervals, separately for the five LMER models on the Portuguese data.

For the Portuguese data, Model 1 largely explained the variance in the duration measurements as resulting from differences between word types: initiality barely captured any variation, yet the model as a whole still explained 24.4% of all variance. In Model 2, proximity-to-stress and an interaction between that fixed factor and initiality were added to the structure of Model 1.This change basically maintained the proportion of variance explained by the fixed and random effects together (24.5%), but increased the marginal R^2 to 16.6%, i.e. proximity-to-stress is a good predictor for the duration of /ia/ while initiality remains quite irrelevant. Recall, however, that the estimations resulting from these two models cannot be trusted because the missing by-speaker intercept violates the independence assumption. Adding this random intercept by speaker led to a large increase in the proportion of variance explained by the random effects structure, as can be seen from the higher conditional R^2 (41.8%) and virtually unchanged marginal R^2 (16.2%) for Model 3.

For Model 4, all tokens labelled as noisy were excluded from the original dataset, so the difference between Models 3 and 4 is the presence/absence of 3116 tokens (31.4% of all tokens) with background noise, music, or unacceptable audio quality. The resulting R^2s for Model 4 are very similar to those of Model 3, except the 95% confidence intervals are larger for Model 4 because of the smaller data sample. Finally, Model 5 was run on the 6785 non-noisy tokens whose duration was derived from hand-corrected segment boundaries. Out of the five models, Model 5 resulted in the highest marginal and conditional R^2s. It is still the case that proximity-to-stress explained the overwhelming amount of variance in the duration measurements (27.5%) and that initiality was not a good predictor. In addition, the by-word and by-speaker intercept also explained 25.0% of the variance (i.e. the difference between the conditional and marginal R^2s).

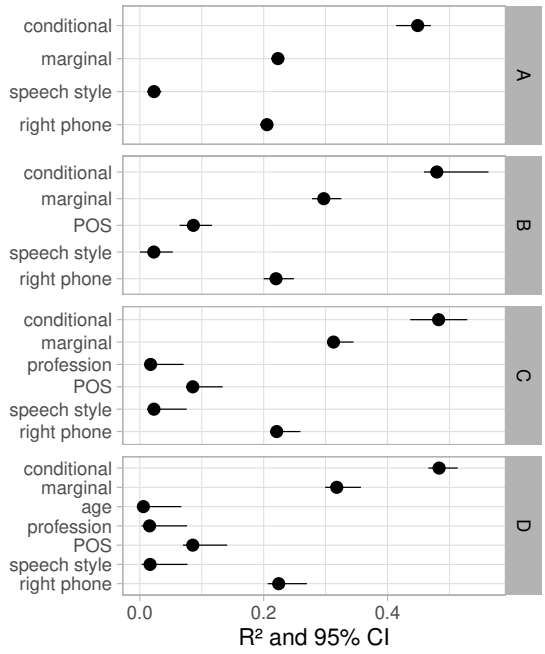

Figure 4: Conditional, marginal, and semi-partial R^2 estimates (95% confidence intervals) for multiple predictor sets across the four nested mixed-effects logistic regression models fitted to the French data.

For the French data, Model A shows that the fixed predictors explain 22% of the variance, with right phone context contributing the most (approximately 20%). Including random effects increases the ex-

plained variance by an additional 23%, highlighting the importance of allowing for speaker- and word-specific variability in this statistical analysis. Adding POS in Model B improves model fit, increasing the variance explained by the fixed factors to 29% and the total (conditional) R^2 to 47%. Incorporating speaker profession in Model C yields only a modest increase in marginal R^2 (1.8%), and speaker age in Model D contributes less than 1%. Despite the small effects of profession and age, Model D achieves the highest conditional R^2(48.3%) and provides the best overall fit according to the Akaike Information Criterion (AIC).

These findings show that filtering the data and sourcing additional information impacts the statistical analysis and renders large non-linguistic corpora viable for phonetic analysis.

5. Discussion

The aim of this paper was to illustrate some of the challenges associated with the linguistic analysis of large corpora. Using broadcast corpora of French and Portuguese, we demonstrated possible solutions to deal with sub-optimal audio quality, missing speaker information, and a lack of additional annotations. The incremental R^2 analyses showed that each processing step and added predictor changed the goodness-of-fit of the statistical models, but not to equal extents. For example, making sure that the results on Portuguese were not influenced by varying noise levels or background music was a labor-intensive process, but it was ultimately shown that excluding noisy audio files did not significantly alter the goodness-of-fit of the LMER (cf. R^2s of Models 3 to 4 in Figure 3). That is, the automatic segmentation of the data was not influenced by the variable audio quality. This finding is very encouraging for analyses that depend solely on the accuracy of the aligner: segmental duration, speech rate measurements, the identification of segments of interest in the stream of fluent speech, and the presence of pronunciation variants. The same cannot be said for studies involving automatic measurements of formants or fundamental frequency which might be affected more strongly by noisy recordings, even though audio enhancement and sound source separation algorithms are getting better at cleaning up audio data without influencing phonetic measurements (Stanley et al., 2025; Liu et al., 2021). Furthermore, it should be noted that the performance of ASR systems and forced-aligners can vary depending on the training data, e.g. the transcription and automatic segmentation might be less accurate if the aligner was not trained on data with similar noise levels as the corpus data. As in the case of French (see section 2.1), the Portuguese acoustic models were trained on similar corpora, thus a high accuracy rate in the segmentations, independent of noise level, was to be expected.

In contrast, identifying unique speakers is crucial not only to avoid violations of the independence assumption in regression models, but also because speaker information might lead to model improvement. As reported above, adding the speaker IDs as a random factor increased the overall goodness-of-fit of the models by 13% in the French case and 17% in the Portuguese case. More detailed information about speakers does not always significantly increase the marginal R^2, as e.g. in the case of the speakers' age (Model D) and gender (no effect) in the French data, but they might still be relevant for the research question or provide further insights. Especially in working with large corpora, there may be imbalances or biases in the data that need to be uncovered. In both the Portuguese and the French studies, two thirds of the speakers were male and it is likely that the speakers who were recorded on news or interview shows for national radio and TV had relatively homogeneous levels of education. For instance, for the French study, which focused on socio-economic status, the data did not permit a binary classification of professions (e.g., working vs. middle class), as the large majority of the speakers held white-collar occupations – mainly journalists and politicians – which are commonly overrepresented in broadcast media. So while corpus data comprises thousands of speakers which can potentially cover different genders, age ranges, education levels, and socioeconomic statuses, diversity is not an automatic consequence of large-scale data.

Data cleaning and augmentation for both studies has been done (semi-)manually. For the Portuguese study, noisy recordings were discarded based on auditory judgements, and stress information was annotated (semi-)manually, using a combination of automated rule-based stress assignment and manual labelling based on pronunciation dictionaries. For the French study, speaker metadata (profession, age, gender) were manually sourced based on each speaker's identity. Similarly, part of speech was manually annotated for over 300 distinct word types. Recent advances in Deep Learning algorithms and Large Language Models, however, can significantly reduce this manual workload. For instance Wu et al. (2023) proposed a method capable of predicting a speaker's profession from recordings with up to 76% accuracy by combining acoustic (speech rate, pitch and formant measurements) and transcription (textual feature vectors) features. Regarding part-of-speech tagging, a recent review (Chiche and Yitagesu, 2022) highlights that modern deep learning approaches can achieve up to 99% accuracy, but this comes with high computational costs. A recent study on data

from the West Point Brazilian Portuguese (Harmath-de Lemos, 2022) has shown that MFCC-based HMM-GMM models can reliably detect word stress in continuous speech. Despite significant advances in automatic methods, manual intervention remains indispensable, both for configuring the algorithms and for output verification. This process can prove to be as time-consuming as manual annotations, keeping the overall workflow labor-intensive, regardless of the chosen approach.

The question, then, is whether data cleaning and optimization is worth the investment. We argue here that there is a strong argument that it is. High-quality linguistic metadata and annotation enable more robust analyses which facilitate replication and allow the corpora to be used for further studies. Without adequate preparation, large-scale datasets, that were originally designed for non-linguistic purposes, risk providing the sought-after quantity without the necessary interpretability. Moreover, the process of annotating and error checking can be revealing in itself, offering insights on imbalances in the data, subtle trends or mismatches between automatic and human assessments.

Ultimately, the decision on whether preprocessing is worth the cost depends on the research goal. Efforts devoted to corpus preparation should be proportional to the theoretical linguistic demands and to the long-time exploitation potential of the dataset. The long-term payoff in data reliability, reusability, and methodological transparency often justifies the investment, especially in naturalistic speech, where variability is the goal.

6. Conclusion

The studies presented here show that large-scale phonetic research depends not on the availability of data, but on metadata richness which can be achieved through the strategic use of shared resources, annotation tools, database management systems, computational processing, modern statistical techniques, and manual work. We argue that large-scale, naturalistic data should be used more frequently in the phonetic sciences, both to test whether results from controlled experimental studies hold up in natural settings, and because data found "in the wild" can reveal insights that are beyond the reach of experimental studies. At the same time, phenomena attested in naturalistic speech can be investigated more systematically in laboratory settings. Thus, we do not propose that corpus studies should take over traditional laboratory approaches, but rather, that the two complement each other: large-scale, naturalistic speech represents the virtually infinite variability of spoken language including imbalances and messiness, whereas laboratory speech is often more formal but offers more accurate control and allows for hypothesis testing.

7. Acknowledgements

This research was partially supported by the ANR project DIPVAR (ANR-21-CE38-0019), PI Ioana Vasilescu. We thank four anonymous reviewers and one meta-reviewer for their insightful comments and feedback.

8. Bibliographical References

Martine Adda-Decker and Lori Lamel. 1999. Pronunciation Variants across System Configuration, Language and Speaking Style. *Speech Communication*, 29:83–98.

Emily Ahn and Eleanor Chodroff. 2022. VoxCommunis: A Corpus for Cross-linguistic Phonetic Analysis. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 5286–5294.

Alan W. Black. 2019. CMU Wilderness Multilingual Speech Dataset. In *Proceedings of the 2019 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(10).

Eleanor Chodroff and Miao Zhang. preprint. A Crosslinguistic Corpus Phonetic Analysis of Intrinsic Vowel Duration. *arXiv*.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Abigail C. Cohn and Margaret E. L. Renwick. 2019. Doing Phonology in the Age of Big Data. In *Cornell Working Papers in Phonetics and Phonology 2019*, pages 1–36.

John Coleman, Mark Liberman, Greg Kochanski, Lou Burnard, and Jiahong Yuan. 2011. Mining a year of speech. In *Proceedings of New Tools and Methods for Very-Large-Scale Phonetics Research*, pages 16–19.

BNC Consortium. 2007. The British National Corpus, XML Edition.

Johanna Cronenberg, Ioana Chitoran, Lori Lamel, and Ioana Vasilescu. 2024. Crosslinguistic Comparison of Acoustic Variation in the Vowel Sequences /ia/ and /io/ in Four Romance Languages. In *Proceedings of the 25th Interspeech Conference*, pages 3689–3693.

Johanna Cronenberg, Lori Lamel, and Ioana Chitoran. In prep. Reduction of the Hiatus /ia/ in Naturalistic European Portuguese.

Thiago Fraga-Silva, Jean-Luc Gauvain, and Lori Lamel. 2011. Lattice-based Unsupervised Acoustic Model Training. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Thiago Fraga-Silva, Jean-Luc Gauvain, and Lori Lamel. 2013. Interpolation of acoustic models for speech recognition. In *Proceedings of the Interspeech*, pages 3347–3351.

Sylvain Galliano, Edouard Geoffrois, Guillaume Gravier, Jean-Francois Bonastre, Djamel Mostefa, and Khalid Choukri. 2006. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proc. LREC 2006*, volume 6, pages 315–320.

Guilherme D. Garcia. 2017. Weight Gradience and Stress in Portuguese. *Phonology*, 34(1):41–79.

John J. Godfrey and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62. Web Download.

Guillaume Gravier, Gilles Adda, Niklas Paulson, Mathieu Carré, Aude Giraudel, and Olivier Galibert. 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *Proc. LREC 2012*.

Simone Harmath-de Lemos. 2022. Detecting word-level stress in continuous speech: A case study of brazilian portuguese. *Journal of Portuguese Linguistics*, 20(101).

Jonathan Harrington, Sallyanne Palethorpe, and Catherine I. Watson. 2000. Does the Queen speak the Queen's English? *Nature*, 408(6815):927–928.

Jennifer B. Hay, Stefanie Jannedy, and Norma Mendoza-Denton. 1999. Oprah and /ay/: Lexical Frequency, Referee Design and Style. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 1389–1392.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26.

William Labov. 1972. *Sociolinguistic Patterns*. Blackwell.

Lori Lamel and Jean-Luc Gauvain. 1992. Continuous Speech Recognition at LIMSI. In *Proceedings of the Final Review of the DARPA ANNT Speech Program*, pages 1–7.

Mark Y. Liberman. 2019. Corpus Phonetics. *Annual Review of Linguistics*, 5(1):91–107.

Shuo Liu, Gil Keren, Emilia Parada-Cabaleiro, and Björn Schuller. 2021. N-HANS: A Neural Network-Based Toolkit for in-the-wild Audio Enhancement. *Multimedia Tools and Applications*, 80(18):28365–28389.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, third edition edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Anisia Popescu, Mathilde Hutin, Lori Lamel, Ioana Vasilescu, and Martine Adda-Decker. 2023. Stop devoicing and place of articulation: a cross-linguistic study using large-scale corpora. In *Proceedings of the 20th ICPhS*, pages 3186 – 3190.

Anisia Popescu, Lori Lamel, Marc Evrard, and Ioana Vasilescu. 2025. Tracking /r/ deletion: Forced alignment of pronunciation variants and sociophonetic insights into post-obstruent final /r/ in french. In *Proceedings of the Interspeech*, pages 2945 – 2949.

Rasmus Puggaard-Rode, Camilla Søballe Horslund, and Henrik Jørgensen. 2022. The Rarity of Intervocalic Voicing of Stops in Danish Spontaneous Speech. *Laboratory Phonology*, 13(1):1–47.

Hugo Quené. 2013. Longitudinal Trends in Speech Tempo: The case of Queen Beatrix. *The Journal of the Acoustical Society of America*, 133(6):EL452–EL457.

Morgan Sonderegger, Jane Stuart-Smith, Jeff Mielke, and The SPADE Consortium. 2023. How variable are English sibilants? In *Proceedings of the 20th International Congress of Phonetic Sciences*, pages 3196–3200.

Joseph A. Stanley, Lisa Morgan Johnson, and Earl Kjar Brown. 2025. Testing the Effect of Speech Separation on Vowel Formant Estimates. *Linguistics Vanguard*.

Martin A. Stoffel, Shinichi Nakagawa, and Holger Schielzeth. 2021. partR2: Partitioning R2 in Generalized Linear Mixed Models. *PeerJ*.

Ioana Vasilescu, Yaru Wu, Adèle Jatteau, Martine Adda-Decker, and Lori Lamel. 2020. Alternances de Voisement et Processus de Lénition et de Fortition: Une Étude Automatisée de Grands Corpus en Cinq Langues Romanes. *Traitement Automatique des Langues*, 61(1):11–36.

Yaru Wu. 2018. *Étude de la réduction segmentale en français parlé à travers différents styles : apports des grands corpus et du traitement automatique de la parole à l'étude du schwa, du /K/ et des réductions à segments multiples*. Phd thesis, Université Sorbonne Nouvelle - Paris 3.

Yaru Wu, Lihu Chen, Benjamin Elie, Fabian Suchanek, Ioana Vasilescu, and Lori Lamel. 2023. Who's speaking? predicting speaker profession from speech. In *Proc. ICPhS 2023*, pages 3086–3090.

Yaru Wu, Fabian Suchanek, Ioana Vasilescu, Lori Lamel, and Martine Adda-Decker. 2022. Using a knowledge base to automatically annotate speech corpora and to identify sociolinguistic variation. In *Proc. LREC 2022*, pages 1054–1360.