Impact of ASR Transcriptions on French Spoken Coreference Resolution

Kirill Milintsevich
Institut National de l'Audiovisuel (INA)
France
kmilintsevich@ina.fr

Abstract

This study introduces a new ASR-transcribed coreference corpus for French and explores the transferability of coreference resolution models from human-transcribed to ASR-transcribed data. Given the challenges posed by differences in text characteristics and errors introduced by ASR systems, we evaluate model performance using newly constructed parallel human-ASR silver training and gold validation datasets. Our findings show a decline in performance on ASR data for models trained on manual transcriptions. However, combining silver ASR data with gold manual data enhances model robustness. Through detailed error analysis, we observe that models emphasizing recall are more resilient to ASR-induced errors compared to those focusing on precision. The resulting ASR corpus, along with all related materials, is freely available under the CC BY-NC-SA 4.0 license at: https://github.com/ina-foss/french-asr-coreference.

1 Introduction

Coreference resolution differs between written and spoken texts and is generally more challenging for spoken data, primarily because most existing corpora are based on written texts (Amoia et al., 2012). For the French language, the large-scale coreference corpus ANCOR (Muzerelle et al., 2014) is based on interview transcripts that were produced manually (Antoine et al., 2002; Eshkol-Taravella et al., 2011). With the help of recent state-of-the-art automatic speech recognition (ASR) systems, such as Whisper (Radford et al., 2023), we can automatically transcribe large amounts of audio data. For example, the *Institut national de l'audiovisuel* stores millions of hours of recorded French TV and radio broadcasts, which are continuously and automatically transcribed and used for research in the social sciences and digital humanities.

However, unlike the manual transcripts in ANCOR, Whisper produces text that includes punctuation, capitalization, occasional rephrasing, as well as ASR errors. This might lead to poor transferability of coreference resolution models trained on the ANCOR corpus when applied to ASR data.

To date, most studies on transferability in coreference resolution have focused on cross-corpus (Xia and Van Durme, 2021; Yuan et al., 2022) and cross-lingual (Lai and Ji, 2023; Pražák et al., 2024) transferability. However, pre-trained language models are sensitive even to small text perturbations, such as punctuation (Wang et al., 2023) and casing (Moradi and Samwald, 2021). Moreover, ASR errors have negative impact on downstream tasks, such as named entity recognition (Szymański et al., 2023) or spoken language understanding (Chang and Chen, 2022). Since these models are at the heart of most recent automatic coreference resolution models, such sensitivities might hinder their performance when resolving coreference on ASR texts.

In this study, we evaluate the transferability of coreference resolution models from human-transcribed to ASR-transcribed data. We create parallel silver training and gold validation datasets and conduct a comparative study using two distinct architectures. Finally, we perform a detailed error analysis to identify the types of ASR-induced errors that most affect model performance.

2 Automatic Coreference Resolution

Most widely used end-to-end coreference resolution systems are *mention-to-link*, meaning they first predict candidate mentions—phrases referring to some entity—and then establish coreference or anaphoric links between each pair of candidates. Lee et al. (2017) developed a model that lists all overlapping spans of a certain length as possible mention candidates. However, this approach incurs high computational overhead. Subsequently, Kirstain et al. (2021) reduced the computational

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 85–94
Suzhou, China, November 9, 2025.

complexity by using only the start and end tokens to construct the mention representation. CorPipe uses a similar approach by first predicting all mentions using a sequence tagging approach and then establishing coreference links with a self-attention layer (Straka, 2024). This system has repeatedly shown top performance at the CRAC Shared Task on coreference resolution (Novák et al., 2024).

Another approach to automatic coreference resolution is the *link-to-mention* approach, where anaphoric links are first predicted between the syntactic heads, and then the mention spans are reconstructed from the coreferent heads. This approach reduces computational overhead compared to the mention-to-link approach, as constructing span representations is unnecessary. Dobrovolskii (2021) presented WL-Coref, the first model that followed the link-to-mention approach. However, D'Oosterlinck et al. (2023) found that in the case of conjunctions of multiple mentions, the same syntactic head could correspond to multiple mention spans, leading to errors in the model of Dobrovolskii (2021). Subsequently, D'Oosterlinck et al. (2023) proposed moving the syntactic head to the coordinating conjunction instead (e.g., in a mention [Tom and Mary], the head is moved from "Tom" to "and"). Finally, Liu et al. (2024) proposed another iteration of the WL-Coref model, which added a special "antecedent link" to support singletons.

3 Data

The ANCOR corpus is the largest collection of spoken French text annotated for coreference. It consists of manual transcripts from four corpora: two representing socio-linguistic interviews (Eshkol-Taravella et al., 2011) and two representing highly interactive dialogues (Antoine et al., 2002). Originally in TEI format, the corpus is now available in the CorefUD format within the CorefUD collection (Novák et al., 2025; Nedoluzhko et al., 2022). The manual transcriptions in ANCOR do not include any punctuation or casing, except for question marks and proper names, and accurately retain speech discontinuities, including repetitions and stuttering.

The coreference annotation in ANCOR has several particularities that distinguish it from other corpora. First, the deictic pronouns (e.g., *I, you, we*) are always annotated as singletons, i.e. they are never linked to any other mentions. Second, the discontinuous mentions are present in the corpus.

Statistics	ANCOR		ASR	
	Train	Val	Train	Val
#documents	365	45	54	9
#sentences	25K	2,385	16K	2,628
#words	371K	38K	193K	31K
#entities	55K	5,827	25K	4,212
#mentions	91K	9,491	40K	6,751
%singletons	80.8%	79.9%	79.9%	79.1%
%disc. mentions	0.5%	0.6%	0.2%	0.3%

Table 1: Statistics of the datasets. Here, *disc.* stands for discontinuous.

Finally, in the original corpus, each utterance is attributed to a speaker, but this information was omitted in the CorefUD format.

3.1 Re-transcribing the Corpus

To build an ASR coreference corpus, we utilized the Whisper Large multilingual model[1] (Radford et al., 2023) to transcribe the ESLO corpus (Eshkol-Taravella et al., 2011) which constitutes the largest part of ANCOR. We then performed word-level alignment of the manual transcriptions with the ASR transcriptions using the `spacy-alignments`[2] library. Since the coreference annotation in CorefUD format is also word-level, we transferred it to the ASR data (see Annex B for an example). Next, we split the ASR data into training and validation sets using the same documents as in ANCOR. Finally, we added morpho-syntactic information (lemmas, part-of-speech tags, detailed morphological features, dependency trees) using Stanza's default model for French (Qi et al., 2020) and repositioned the syntactic head of each mention with heuristics from the `udapi-python` package.[3]

Due to imperfect automatic alignment, the resulting ASR corpus often contained invalid coreference annotations. For the validation set, we manually verified and corrected these annotations. For the training set, we removed sentences containing invalid CorefUD annotations, such as unclosed mention tags or closing tags without corresponding opening tags. Table 1 shows that the resulting ASR training set is almost half the size of its ANCOR counterpart, while the ASR validation set retains nearly 80% of its original size. Furthermore, the proportion of discontinuous mentions in the ASR dataset is smaller than in the original dataset. This

[1] Using the WhisperX implementation (Bain et al., 2023).

[2] `https://github.com/explosion/spacy-alignments`

[3] `https://github.com/udapi/udapi-python`

Model	Train	MUC	B^3	$CEAF_e$	BLANC	LEA	MOR	CoNLL
				Human transcription validation set				
WL-Coref	*Hum.*	**76**/77/76	59/68/63	67/58/62	55/68/59	55/65/60	**86**/85/**86**	67.26
	ASR	67/68/68	37/59/45	62/34/44	43/57/43	33/55/41	85/79/82	52.30
	H+A	**76/79/77**	**63/71/67**	**70/63/67**	**61/70/65**	**59/68/63**	**86/87/86**	**70.23**
CorPipe-24	*Hum.*	78/73/76	73/60/66	66/71/68	72/57/63	69/56/62	**86/84/85**	70.06
	ASR	64/66/65	58/54/56	55/61/58	55/53/54	52/48/50	72/82/76	59.50
	H+A	**80/75/77**	**74/63/68**	**67/72/69**	**73/60/66**	**70/58/64**	85/**84/85**	**71.37**
				ASR transcription validation set				
WL-Coref	*Hum.*	66/**74**/70	47/**67**/55	62/51/56	43/**65**/49	43/**63**/51	80/**84**/82	60.42
	ASR	69/64/66	46/54/50	62/42/50	46/50/44	42/50/46	**85**/76/80	55.43
	H+A	**74**/72/**73**	**63**/63/**63**	**66/61/64**	**60**/61/**61**	**59**/60/**59**	84/82/**83**	**66.55**
CorPipe-24	*Hum.*	74/**69**/71	68/**59/63**	**64/68/66**	**66/56/60**	63/**54**/58	85/82/**83**	66.79
	ASR	73/67/70	68/56/61	62/65/64	65/52/57	63/51/56	**86**/79/82	64.79
	H+A	**75/69/72**	**69/59/63**	**64**/67/65	**66**/55/**60**	**64/54/59**	**86**/81/**83**	**66.80**

Table 2: Results on the Human (upper part) and ASR (lower part) transcription validation sets. For each validation set, the best results for each model are shown in **bold**, and the best results across the models are underlined. All metrics are reported as Recall/Precision/F1, except for the CoNLL F1 score.

reduction is due to speech discontinuities (e.g., stutterings, repetitions, talking over) being preserved in the human transcription but absent from the ASR transcription. Finally, the ASR validation set has more sentences[4] despite being smaller. This results from Whisper producing text closer to written form, while human transcriptions split the text by pauses in speech or speaker changes.

4 Experimental Setup

We trained both WL-Coref (Dobrovolskii, 2021; D'Oosterlinck et al., 2023; Liu et al., 2024) and CorPipe (Straka, 2024) models with `camembertav2-base`[5] pre-trained encoder, which currently achieves state-of-the-art results on French NLP tasks (Antoun et al., 2024) (see Appendix A for more details). For each architecture, three model variants were trained according to the training data: 1) ***Hum.*** using the original ANCOR data; 2) ***ASR*** using the automatically transcribed subset of the original data; 3) ***H+A*** using the combination of the ANCOR and ASR training datasets.

To measure the performance of the models, in addition to the ASR validation set, we created a subset of the ANCOR human transcription validation set, which includes the same documents as the ASR validation set. For evaluation metrics, we used MUC (Vilain et al., 1995), B-cubed (Bagga and Baldwin, 1998), CEAFe (Luo, 2005), BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube, 2016), MOR (Mention Overlap Ratio) (Žabokrtský et al., 2022), and the CoNLL score which is the average of the first three metrics. All metrics were calculated using the CorefUD scorer[6] with exact mention matching and excluding all singletons.

5 Results and Discussion

The upper part of Table 2 presents the results on the manually transcribed validation set. For both the WL-Coref and CorPipe models, training on human transcripts (*Hum.*) yielded better performance compared to using only automatically constructed ASR training data (*ASR*), with CoNLL score drops of -14.96 and -10.56 for the WL-Coref and CorPipe models, respectively. Utilizing a mix of manual and ASR training data (*H+A*) slightly enhanced the performance of both models on the manually transcribed data, resulting in CoNLL score increases of $+2.97$ and $+1.31$ for the WL-Coref and CorPipe models, respectively. The lower part of Table 2 illustrates similar trends for the ASR validation set. However, the WL-Coref model appears to be more sensitive to changes in the data, whereas the CorPipe model shows almost no difference between the *Hum.* and *H+A* variants.

Across both validation sets, WL-Coref achieved higher precision in all metrics except $CEAF_e$, while

[4]Defining a sentence in spoken text can be challenging. In the context of this work, a sentence is defined as a continuous sequence of words where all mentions are fully contained within it, meaning that a mention cannot span across sentence boundaries.

[5]`https://huggingface.co/almanach/camembertav2-base`

[6]`https://github.com/ufal/corefud-scorer`

Model	Train	Head Error	Span Error	Conflated Entities	Extra Mention	Extra Entity	Divided Entity	Missing Mention	Missing Entity
Human transcription validation set									
WL-Coref	*Hum.*	188	43	85	286	70	62	100	125
	ASR	163	53	58	322	48	54	133	171
	H+A	188	49	76	303	71	46	79	111
CorPipe	*Hum.*	3	64	112	277	81	97	121	73
	ASR	3	103	143	446	150	142	127	79
	H+A	3	63	105	257	79	96	93	70
ASR transcription validation set									
WL-Coref	*Hum.*	167	48	54	377	85	50	119	125
	ASR	144	51	58	257	42	50	171	197
	H+A	154	53	64	287	68	46	118	137
CorPipe	*Hum.*	0	85	91	307	87	76	130	86
	ASR	0	82	101	301	78	79	133	103
	H+A	0	86	89	285	77	77	129	102

Table 3: Error analysis on the Human (upper part) and ASR (lower part) transcription validation sets. For each validation set, the cells are color-coded in a gradient column-wise, with red representing the highest value and green representing the lowest value.

CorPipe showed higher recall. When applied to the ASR validation set, WL-Coref trained on human transcribed data exhibited a significant drop in recall and only a slight drop in precision. In contrast, CorPipe showed only a moderate decrease in recall.

We hypothesize that this discrepancy may occur because the WL-Coref model predicts links between mention heads, making it more susceptible to errors from ASR and automatic syntactic parsing, which in turn affect its performance. In contrast, the CorPipe model employs a sequence tagging approach to detect mentions, which does not rely on additional syntactic information.

5.1 Error Analysis

To better understand the impact of ASR transcriptions on the performance of coreference resolution models, we conduct an error analysis based on the work of Kummerfeld and Klein (2013). To adapt this analysis to the exact mention matching scenario, we introduce a Move Head operation. This operation corrects a predicted mention head if the spans of the predicted and ground truth mentions match exactly, corresponding to what is termed a Head Error. The remainder of the analysis largely adheres to the methodology outlined by Kummerfeld and Klein (2013).

Table 3 presents the error analysis for the WL-Coref and CorPipe models (see Annex C for examples of errors). The WL-Coref model exhibits a high number of Head Errors but fewer Span Errors. This can be explained by the design of the CorPipe model, which is specifically tailored for the CorefUD shared task where head matching is used for evaluation. Interestingly, the WL-Coref model produces more accurate spans even when starting from incorrect heads. Lastly, WL-Coref consistently has more Missing Entity errors which is explained by the lower recall.

When evaluated on the ASR validation set, models trained on *Hum.* data demonstrate more Conflated Entities, where a predicted entity includes mentions from different ground-truth entities, and fewer Divided Entities, where different predicted entities include mentions from the same ground-truth entity. This behavior suggests that when applied to ASR transcriptions, the models group mentions into tighter clusters. A possible reason is that the lack of filler words and repetitions in ASR transcriptions reduces the distance between mentions.

The large number of Extra Mention errors mostly stems from assigning mentions, which should otherwise be singletons, to a coreference chain and linking the pronouns *ce* and *ça* (it) when they are non-referential. The increase in such errors on the ASR validation data could be explained by Whisper producing more "grammatically valid" transcriptions, adding these pronouns, e.g., by inserting *c'est* (it is) when they are absent from speech and consequently from human transcriptions.

Finally, we found that both human and ASR validation sets contain annotation errors. However, their exact impact on the evaluation is beyond the scope of this study and requires further study.

6 Conclusions

In this study, we evaluated the performance of coreference resolution models trained on human-transcribed data when applied to ASR-transcribed data, observing a general trend of decreased performance on ASR data for models trained on manual transcriptions. We proposed an approach to automatically transfer coreference annotations from human to ASR transcriptions and discovered that training only on silver ASR data harms model performance, whereas combining silver ASR data with gold manual data enhances it. Further error analysis revealed that ASR systems, which tend to overcorrect transcriptions, introduce potential errors to coreference resolution systems. We found that models prioritizing higher recall are more robust to these errors than those focusing on precision.

Limitations

Given the scarcity of spoken coreference datasets in French, this study is confined to a single corpus, primarily comprising socio-linguistic interviews. These interviews have low interactivity and cover a limited range of topics. Furthermore, participants are sampled from a restricted geographic area, specifically Orléans and Tours, which narrows the vocabulary used in the interviews. A more topically diverse corpus would be essential for a broader evaluation.

Regarding coreference resolution models, this study evaluates only two architectures: WL-Coref and CorPipe. While a more diverse set of models would enhance the robustness of the comparison, hardware limitations and variations in coreference data formats present significant challenges. Additionally, the prevalence of English-specific or OntoNotes-specific architectures complicates the adaptation of existing models to other languages and the CorefUD format, which is beyond the scope of this study.

Finally, this study only uses Whisper as the ASR system for automatically transcribing the dataset recordings. We acknowledge that other ASR systems may produce different transcriptions, potentially leading to different effects on automatic coreference resolution performance.

Acknowledgments

This work is partially funded by the ANR Pantagruel project ANR-23-IAS1-0001-02.

References

Marilisa Amoia, Kerstin Kunz, and Ekaterina Lapshinova-Koltunski. 2012. Coreference in spoken vs. written texts: a corpus-based analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 158–164, Istanbul, Turkey. European Language Resources Association (ELRA).

Jean-Yves Antoine, Sabine Letellier-Zarshenas, Pascale Nicolas, Igor Schadle, and Jean Caelen. 2002. Corpus OTG et ECOLE_MASSY : vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement. In *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles. Posters*, pages 319–324, Nancy, France. ATALA.

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. CamemBERT 2.0: A smarter french language model aged to perfection. *arXiv preprint arXiv:2411.08868*.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Ya-Hsin Chang and Yun-Nung Chen. 2022. Contrastive learning for improving ASR robustness in spoken language understanding. In *Interspeech 2022*, pages 3458–3462.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karel D'Oosterlinck, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and Chris Develder. 2023. CAW-coref: Conjunction-aware word-level coreference resolution. In *Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 8–14, Singapore. Association for Computational Linguistics.

Iris Eshkol-Taravella, Olivier Baude, Denis Maurel, Linda Hriba, Celine Dugua, and Isabelle Tellier. 2011. Un grand corpus oral disponible : le corpus d'orléans 1968-2012 [a large available oral corpus: Orleans corpus 1968-2012]. *Traitement Automatique des Langues*, 52(3):17–46.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 14–19, Online. Association for Computational Linguistics.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.

Tuan Lai and Heng Ji. 2023. Ensemble transfer learning for multilingual coreference resolution. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 24–36, Toronto, Canada. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Houjun Liu, John Bauer, Karel D'Oosterlinck, Christopher Potts, and Christopher D. Manning. 2024. MSCAW-coref: Multilingual, singleton and conjunction-aware word-level coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 33–40, Miami. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 843–847, Reykjavik, Iceland. European Language Resources Association (ELRA).

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, and 23 others. 2025. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Ondřej Pražák, Miloslav Konopík, and Pavel Král. 2024. Exploring multiple strategies to improve multilingual coreference resolution in CorefUD. *Preprint*, arXiv:2408.16893.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

M Recasens and E Hovy. 2011. BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. 2023. Why aren't we NER yet? artifacts of ASR errors in named entity recognition in spontaneous speech transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1761, Toronto, Canada. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Wenqiang Wang, Chongyang Du, Tao Wang, Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, and Xiaochun Cao. 2023. Punctuation-level attack: Single-shot and single punctuation can fool text models. In *Advances in Neural Information Processing Systems*, volume 36, pages 49312–49324. Curran Associates, Inc.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

A Implementation Details

For WL-Coref, we utilized the implementation by Stanza (Qi et al., 2020), while for CorPipe, we used the implementation from their official repository.[7] The original implementation of CorPipe-2024 is based on an older version of TensorFlow, making it challenging to run on modern systems. We updated the original code to be compatible with the most recent version of TensorFlow without altering the model's architecture. Since neither model supports discontinuous mentions, they were removed from the training data. All models were trained for 40 epochs on an NVIDIA A100 40GB GPU.

B Examples of Data

Table 4 shows an example of human and ASR transcribed data.

C Examples of Errors

In this section, we demonstrate the examples of different errors. Table 5 shows an example of a Head Error, Table 6 shows an example of a Span Error, Table 7 shows an example of an Extra Mention, Table 8 shows an example of an Extra Entity, Table 9 shows an example of a Missing Mention, Table 10 shows an example of a Missing Entity, Table 11 shows an example of a Divided Entity, and Table 12 shows an example of a Conflated Entity.

[7] https://github.com/ufal/crac2024-corpipe

Human	eh bien [**monsieur**]$_s$ [**je**]$_s$ vais commencer par [**vous**]$_s$ poser [des petites **questions** préliminaires toutes simples]$_s$ n' est -ce pas et depuis combien de temps habitez-[**vous**]$_s$ [**Orléans**]$_1$? euh [dix-neuf **ans**]$_2$ [dix-neuf **ans**]$_2$ oui et qu' est -ce qui [**vous**]$_s$ a amené à vivre à [**Orléans**]$_1$?
ASR	Eh bien, [**Monsieur**]$_s$, [**je**]$_s$ vais commencer par [**vous**]$_s$ poser [des petites **questions** préliminaires, toutes simples]$_s$, n'est-ce pas ? Et depuis combien [de **temps**]$_s$ habitez[**-vous**]$_s$ à [**Orléans**]$_1$? [19 **ans**]$_2$. [19 **ans**]$_2$, oui. Et qu'est-ce qui [**vous**]$_s$ a amené à vivre à [**Orléans**]$_1$?

Table 4: Examples of human and ASR transcribed data. Each new line represents a sentence break. Mentions are enclosed in square brackets with mention heads highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton.

Gold	Est-ce que [la **langue** française] est aussi bien enseignée, ou mieux enseignée, ou moins bien enseignée, que de le temps où vous étiez vous-même à l'école ?
Predicted	Est-ce que (la langue **française**) est aussi bien enseignée, ou mieux enseignée, ou moins bien enseignée, que de le temps où vous étiez vous-même à l'école ?

Table 5: Examples of Head Error. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold and underlined**.

Gold	Je trouve que c'est [des **différences** assez grandes].
Predicted	Je trouve que c'est (des **différences**) assez grandes.

Table 6: Examples of Span Error. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**.

Gold	Demain, je serai peut-être partie ou prête à partir et je peux rester encore [six **ans**]$_s$. Je ne sais pas. J'aimerais rester. J'aimerais rester, mais... On nous a dit, n'est-ce pas, ailleurs, que la ville d'Orléans est une ville assez froide, mais je ne sais pas si vous avez des visites là-dessus, puisque vous êtes... Il y a [quelques **années**]$_s$, oui.
Predicted	Demain, je serai peut-être partie ou prête à partir et je peux rester encore (six **ans**)$_1$. Je ne sais pas. J'aimerais rester. J'aimerais rester, mais... On nous a dit, n'est-ce pas, ailleurs, que la ville d'Orléans est une ville assez froide, mais je ne sais pas si vous avez des visites là-dessus, puisque vous êtes... Il y a (quelques **années**)$_1$, oui.

Table 7: Examples of Extra Mention. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown.

Gold	Parce que [les **Américains**]$_s$, [les **Allemands**]$_s$, les Suisses, les Japonais, [tout **ça**]$_s$, [**ça**]$_s$ ne parle pas latin.
Predicted	Parce que (les **Américains**)$_1$, (les **Allemands**)$_1$, les Suisses, les Japonais, (tout **ça**)$_1$, (**ça**)$_1$ ne parle pas latin.

Table 8: Examples of Extra Entity. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown.

Gold	C'est peut-être moi qui écris le plus. Et à [votre **famille**]$_1$? disons que ma femme écrit plutôt à sa famille et moi j'écris plutôt à [la **mienne**]$_1$, encore qu'il arrive très fréquemment que j'écrive à la sienne et qu'elle écrive à [la **mienne**]$_1$.
Predicted	C'est peut-être moi qui écris le plus. Et à (votre **famille**)$_1$? disons que ma femme écrit plutôt à sa famille et moi j'écris plutôt à la mienne, encore qu'il arrive très fréquemment que j'écrive à la sienne et qu'elle écrive à la mienne.

Table 9: Examples of Missing Mention. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown.

Gold	Mais alors, ce qui est embêtant, c'est que vous avez des gosses qui demandent à travailler. et qui ne veulent pas être les bras coincés, ou qui ne veulent pas faire de [la **pâte** à modeler]$_1$ parce qu'on [**l'**]$_1$a déjà fait à la maison.
Predicted	Mais alors, ce qui est embêtant, c'est que vous avez des gosses qui demandent à travailler. et qui ne veulent pas être les bras coincés, ou qui ne veulent pas faire de (la **pâte** à modeler)$_s$ parce qu'on (**l'**)$_s$a déjà fait à la maison.

Table 10: Examples of Missing Entity. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown.

Gold	Je ne parle évidemment pas des dictionnaires de [**langue** ancienne]$_1$ que nous avons à la maison. ... En tout cas, en ce qui concerne [les **langues** anciennes]$_1$, il y a 20 ans, plus de la moitié des élèves faisaient [des **langues** anciennes]$_1$, alors que maintenant, [**ça**]$_1$ représente 1 %.
Predicted	Je ne parle évidemment pas des dictionnaires de (**langue** ancienne)$_1$ que nous avons à la maison. ... En tout cas, en ce qui concerne (les **langues** anciennes)$_2$, il y a 20 ans, plus de la moitié des élèves faisaient (des **langues** anciennes)$_2$, alors que maintenant, (**ça**)$_2$ représente 1 %.

Table 11: Examples of Divided Entity. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown. Three dots (...) show that there are several sentences in between.

Gold	Alors, [au **bureau**]$_1$, À [mon **bureau**]$_1$, j'ai un petit Larousse, mais j'ai chez moi un dictionnaire en dix volumes. ... Je parle pour mon foyer, je ne parle pas [du **bureau**]$_1$. ... Est-ce que vous pourriez dire combien par [**mois**]$_2$? Oui, au point de vue personnel, pas plus de deux ou trois lettres personnelles par [**mois**]$_2$.
Predicted	Alors, (au **bureau**)$_1$, À (mon **bureau**)$_1$, j'ai un petit Larousse, mais j'ai chez moi un dictionnaire en dix volumes. ... Je parle pour mon foyer, je ne parle pas (du **bureau**)$_1$. ... Est-ce que vous pourriez dire combien par (**mois**)$_1$? Oui, au point de vue personnel, pas plus de deux ou trois lettres personnelles par (**mois**)$_1$.

Table 12: Examples of Conflated Entity. Gold mentions are enclosed in square brackets and predicted mentions are between the round brackets. Correct mentions are higlighted in green while incorrect mentions are in red. Mention heads are highlighted in **bold**. Subscripts are entity identifiers with $_s$ denoting a singleton. Only mentions relative to the error are shown. Three dots (...) show that there are several sentences in between.