



# La langue Wu

Simeng SONG & Xiaobo WANG

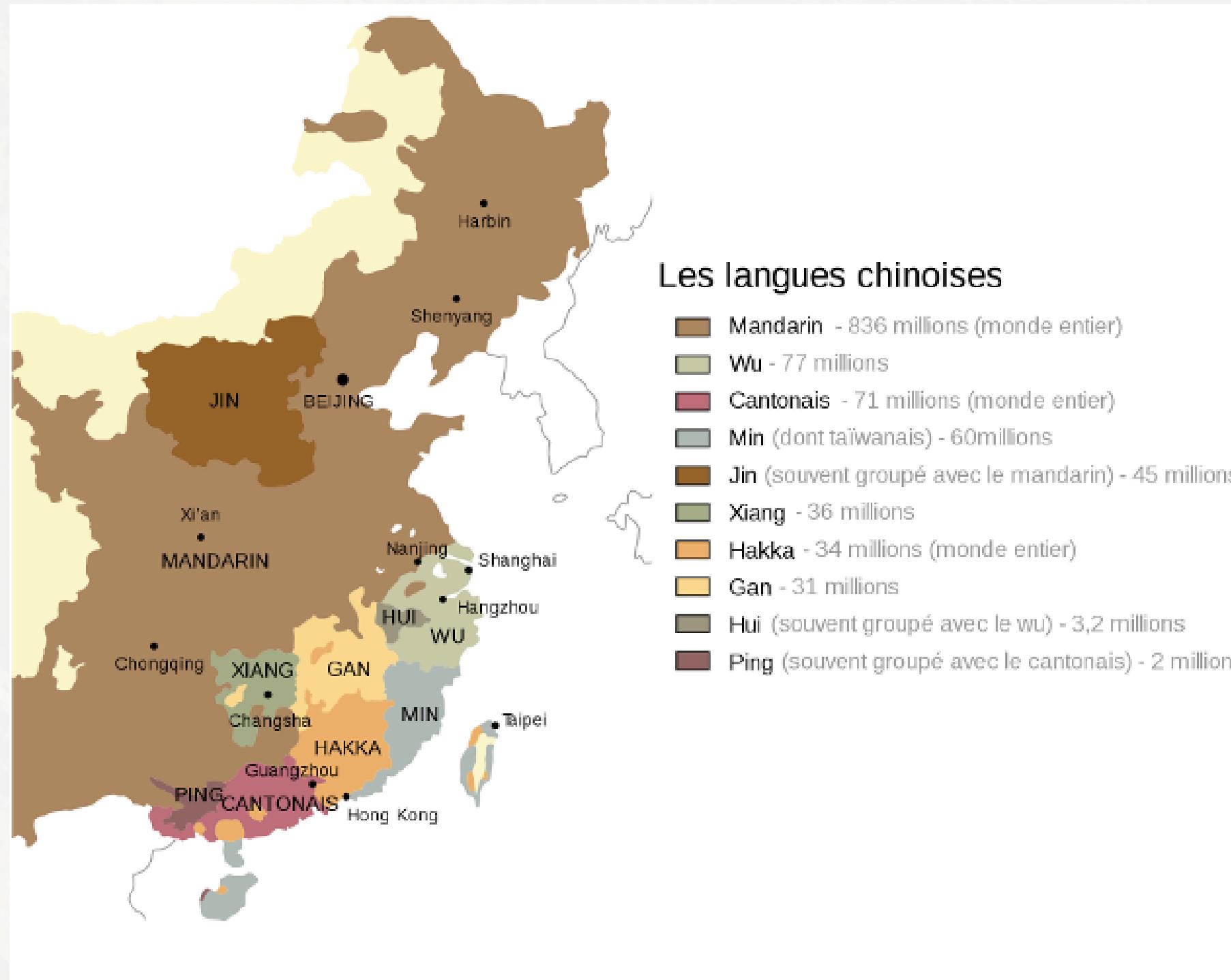
07/10/2025

# Le Wu (Wú yǔ)

- Originaire du delta du Yangtsé :  
**Shanghai**, le sud du **Jiangsu**, le nord  
du **Zhejiang**, ainsi que dans des  
petites parties des provinces du  
**Anhui**, du **Jiangxi** et du **Fujian**.
- Aussi parlé dans les communautés  
d'émigrés de Shanghai et Zhejiang (à  
Hong Kong, aux États-Unis, en  
Europe).
- Différent du mandarin, conserve des  
traits anciens



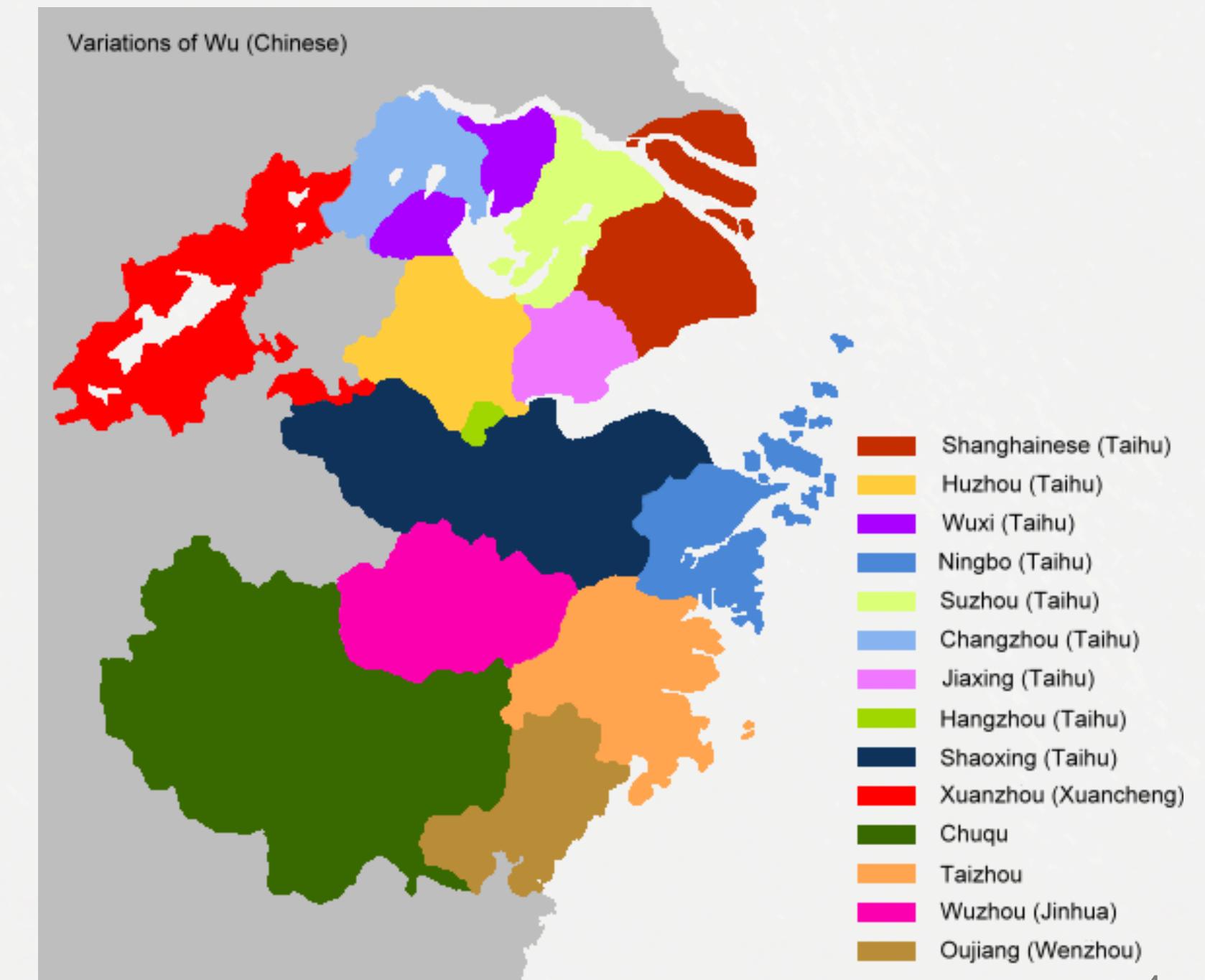
# Statut



- 77 millions de locuteurs natifs
- L'un des groupes linguistiques les plus importants de Chine, mais fortement menacée par le mandarin.
- Très peu présente dans les médias et dans les systèmes éducatifs officiels.
- En danger de déclin chez les jeunes générations

# Variations

Le Wu n'est pas une langue uniforme, mais un ensemble de variétés très diversifiées. On distingue plusieurs sous-groupes principaux :



---

# Histoire

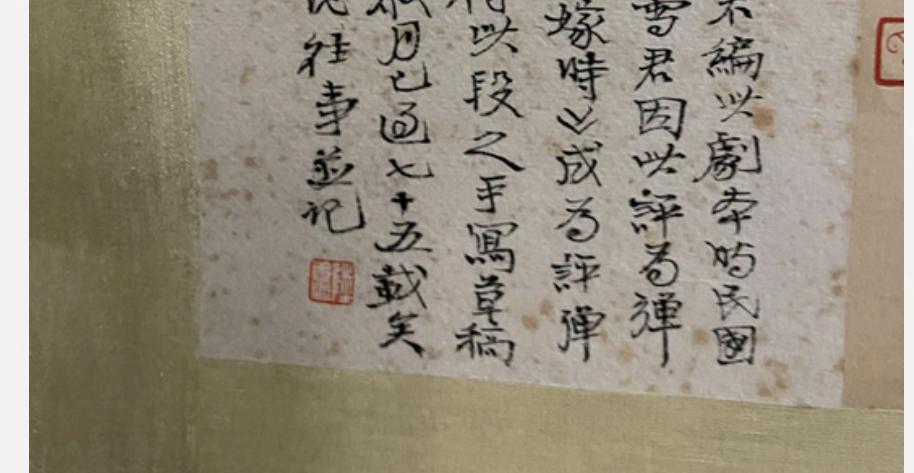
- né de l'évolution du chinois ancien parlé dans le bas Yangtsé
- III<sup>e</sup>-V<sup>e</sup> siècle : formé après les migrations du Nord vers le Sud
- VII<sup>e</sup>-XIII<sup>e</sup> siècle : développement important, la région devient un centre économique et culturel majeur
- XIV<sup>e</sup>-XIX<sup>e</sup> siècle : apparition des six variations et à partir du XIX<sup>e</sup> siècle, le Shanghaïen (appartenant au groupe Taihu) devient la variété plus influente, avec la montée de Shanghai comme centre urbain
- Aujourd'hui : concurrencé par le mandarin, mais reste porteur d'identité régionale

# Ecriture

- Contrairement au mandarin, le Wuyu n'a jamais eu de système d'écriture officiel. Historiquement, les locuteurs utilisaient des caractères chinois adaptés pour transcrire la langue orale, souvent de façon phonétique.
- Les manuscrits de chansons, de contes ou de “tanci” (弹词: un art narratif et musical typique du Jiangnan) montrent des caractères régionaux et des créations locales.



Page intérieure du roman Fan Hua (繁花) – un exemple contemporain d'usage du wuyu dans la littérature écrite.



Manuscrit du “tanci” Hèn bù xiāng féng wèi jià shí (恨不相逢未嫁时) écrit par l'artiste Lu Dan'an.

# — État de l'art en TAL

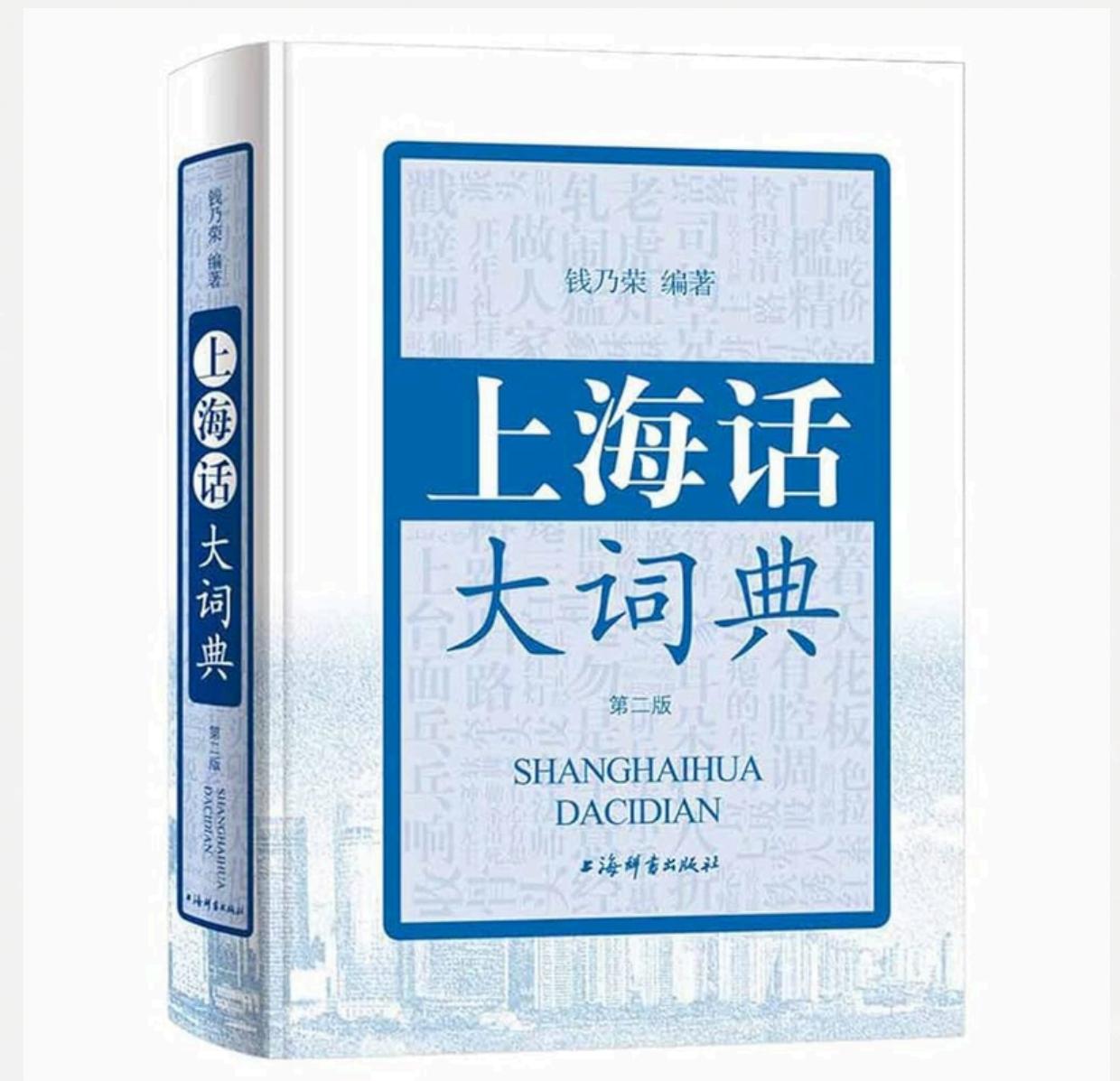
Le Wu est une langue peu dotée : très peu de ressources et d'outils.

Corpus existants :

- Wenzhou Spoken Corpus (orales)
- ASR-SCShhiDiaDuSC (Shanghaïen)
- The FLORES+ Wu dataset (English - Wu)
- Dictionnaires locaux

Outils :

- Analyse phonétique avec Praat
- Reconnaissance vocale (projets Alibaba Qwen-ASR)
- Traduction automatique (intégré dans FLORES+)



# Alibaba's New Speech Recognition Model Pushes Accuracy But Keeps Weights Closed



On September 8, 2025, Alibaba's Qwen team [introduced](#) Qwen3-ASR Flash, an automatic speech recognition (ASR) system covering 11 languages — as well as multiple dialects and accents — and a range of acoustic conditions, [positioned](#) as an all-in-one transcription service.

Unlike conventional systems that require separate models for different languages or conditions, Qwen3-ASR Flash consolidates capabilities into a single API-based model. It supports Mandarin, Cantonese, Sichuanese, Hokkien, Wu, and English (with British and American accents), alongside French, German, Spanish, Italian, Portuguese, Russian, Japanese, Korean, and Arabic.

## WenZhou Spoken Corpus

Department of Linguistics, University of Alberta

Jingxia Lin and John Newman

Word count in all the documents				
Category	Word Count	Punctuation Count	Total	Character Count
News commentary	111861	14488	126349	174361
Song	894	33	927	1395
Story	1050	217	1267	2460
Phone Call	20887	4514	25401	36257
Face to face conversation	13012	2878	15890	23582
Internet chat	7006	1490	8496	13132
<b>Total</b>	<b>154710</b>	<b>23620</b>	<b>178330</b>	<b>251187</b>

## 4 Data Samples

This section lists the first 5 lines of translation along with their English counterparts.

1. 斯坦福医学院个科学家勒礼拜一公布一种可以按种类划分细胞个新个诊断家生个发明：一种可以用标准喷墨打印机大量生产，差弗多小到只有一美分一只个可印芯片。

On Monday, scientists from the Stanford University School of Medicine announced the invention of a new diagnostic tool that can sort cells by type: a tiny printable chip that can be manufactured using standard inkjet printers for possibly about one U.S. cent each.

---

# Perspective

- Améliorer la normalisation du Wu, pour faciliter la segmentation.
- Créer un petit corpus annoté pour entraîner ou ajuster les modèles.
- Adapter ou fine-tuner les outils de segmentation existants (jieba, pkuseg) sur des données du Wu.
- Combiner des règles linguistiques spécifiques au Wu avec des méthodes statistiques.
- Étendre l'étude à d'autres variétés régionales du Wu (Suzhou, Ningbo...).
- Contribuer à la préservation et la numérisation des langues régionales chinoises.

# Perspective

## Qu'est-ce que nous pouvons faire comme projet ?

- Corpus : ASR-SCShhiDiaDuSC
- Tâche : évaluation de la segmentation du Wu (Shanghaïen) avec les outils de mandarin
- Méthodes :
  - Exécuter la segmentation automatique avec *jieba* et *pkuseg*
  - Créer une référence manuelle (“gold standard”)
  - Comparer les résultats et analyser les erreurs
- Résultats attendu :
  - Visualisation des différences entre segmentation automatique et manuelle
  - Valuation entre deux modèle par F1-score

---

# Bibliographies

- **Yue, X.; Miao, L.; Ding, J.** Research on Wu Dialect Recognition and Regional Variations Based on Deep Learning. *Appl. Sci.* 2025, 15, 10227. <https://doi.org/10.3390/app151810227>
- **Newman, John & Lin, Jingxia & Butler, Terry & Zhang, Eric.** (2007). The Wenzhou Spoken Corpus. *Corpora*. 2. 97-109. [10.3366/cor.2007.2.1.97](https://doi.org/10.3366/cor.2007.2.1.97).
- **Hongjian Yu, Yiming Shi, Zherui Zhou, and Christopher Haberland.** 2024. Machine Translation Evaluation Benchmark for Wu Chinese: Workflow and Analysis. In *Proceedings of the Ninth Conference on Machine Translation*, pages 600–605, Miami, Florida, USA. Association for Computational Linguistics.
- <https://slator.com/alibaba-speech-recognition-model-pushes-accuracy-keeps-weights-closed/>
- <https://magichub.com/cn/datasets/shanghai-dialect-scripted-speech-corpus-daily-use-sentence/>

# MERCI

Simeng SONG & Xiaobo WANG

07/10/2025