調理動詞に注目した中国語料理動画コーパスの構築

──ローカル LLM による生成表現の改善──

博士前期課程 2 年次　篠﨑 秀紀

指導教員：Hodošček Bor 准教授, 今尾 康裕教授

研究発表構成

1. はじめに

本研究の目的は以下の 2 点である

1. 中国語の料理動画を対象とし，字幕や音声よるマルチモーダルな調理コーパスの構築を試みること．
2. コーパス内で用いられる調理動詞に注目し、共起関係や語彙クラスタリングを通じて、調理プロセスに内在する語彙的・構文的パターンの分析を行うこと．

これらを達成するために，本研究は調理動画に内在する言語的特徴を多角的に捉える枠組みを構築し，その手法と意義について以下で詳述する．

食は人間の生存と文化において根源的な役割を果たしており、社会的な価値観や個人の習慣とも密接に結びついている。近年では、料理や食に関する膨大な情報が動画共有サイトを中心に日常的に発信されており、料理動画はその中でも視覚的・音声的な情報が豊富なメディアとして注目を集めている。

その一方で、食に関する自然言語処理（NLP）の研究は、ニュースや商品レビュー、医療分野と比べると、まだ十分に整備された大規模なコーパスやリソースが限られている。特に、調理行為に関連する動詞表現の用法や手続き的知識の構造化といった観点から，動画音声・字幕を統合的に解析する研究は極めて少ないのが現状である。

したがってこのようなコーパスを整備することは，今後の言語研究の基盤となり，将来的にはローカル LLM との連携や，調理行為に特化した生成モデルの開発に資することも期待される．具体的には，今後はローカルで稼働する大規模言語モデル（LLM）との統合を目指しており，たとえば RAG（Retrieval-Augmented Generation）機構を介した生成制御や，意味役割付与による言語生成の改善に応用可能である。

本稿ではまず、対象動画の選定と収集手順について述べた後、Whisper や FunASR を用いた文字起こし、そして LLM による整形処理、HanLP によるトークナイズといった一連のコーパス構築工程について詳述する。さらに、他の料理コーパス（例：XiaChuFang

Recipe Corpus）との比較を通じて、本研究の位置付けを明確にし、現状の課題と将来的な言語応用の可能性について考察する。

2. 先行研究

近年、食分野における大規模言語モデル（LLM）の応用が進んでいる。たとえば FoodSky（Zhou et al., 2024）は、調理と栄養に関する知識を深く理解し、自然言語で助言・生成を行うことを目指した中国語特化の LLM である。FoodSky では、命令文ベースのデータを収集・整備した大規模食コーパス「FoodEarth」を構築し、多段階のデータ整備パイプラインによって、Web サイトや電子書籍などの多様な情報源から得られるノイズの多いデータを高品質に変換している。また、TS3M および HTRAG といった独自のモデル設計によって、細粒度な食セマンティクスと文化的多様性に対応可能な言語生成を実現している。

他方、料理レシピの自然言語データとしては、「下厨房（XiaChuFang）」という中国の人気料理共有サイトから収集された大規模レシピデータセット「XiaChuFang Recipe Corpus」も存在する。このコーパスには 150 万件以上のレシピが含まれ、それぞれが料理名、食材、調理工程、検索キーワード、記述的説明などの構造化情報を備えている。家庭的かつ実用的な口語表現が多く含まれており、実際の調理場面で使われる語彙や表現の分析に適している。

これらの先行研究に共通するのは、大規模な食関連テキストを活用して言語モデルの性能向上を図っている点である。しかしいずれも、動画や音声といった**実際の話し言葉・口頭指示**に基づくコーパスの構築や分析には踏み込んでいない。

本研究は、上記のような先行成果の上に立ちつつ、字幕および音声ベースで得られる自然発話に注目し、**調理動詞の語彙的構造の精緻な記述と生成支援**を目指す点に新規性がある。また、XiaChuFang に見られるような実用的なレシピ記述を参照しつつ、それらとは異なるモダリティ（映像・音声）の処理と統合に挑戦する点において、言語学的・工学的双方の意義を持つといえる。

比較項目	**FoodEarth (FoodSky)**	**XiaChuFang Corpus**	**本研究（構築予定）**
データ起点	Web、書籍、論文、ChatGPT 生成	レシピ共有サイト「下厨房」	動画（字幕・音声・OCR）
モダリティ	テキスト（書き言葉中心）	テキスト（投稿者による自然文）	音声＋映像＋口語表現
形式	Q&A 形式（instruction-based）	料理名、食材、手順、説明の構造化	メタ情報、テキスト、動詞・工程データベース
フィルタリング	ChatGPT + 専門家の 2 段階精査	不明（公開済み投稿を収集）	手動 + 時系列整合 + OCR 補正
注目点	栄養学・食習慣・健康提案	家庭的調理表現の多様性	調理動詞の用法・語彙的意味

比較項目	FoodEarth (FoodSky)	XiaChuFang Corpus	本研究（構築予定）
LLM 連携	クラウド LLM で fine-tuning	用途明示なし（NLP 研究に活用）	ローカル LLM（Gemma, Qwen など）

3. 構築方法

3. 1. 対象動画の選定

近年、料理や生活に関する情報の発信手段として、動画共有サイトの活用が一般化している。特に料理動画は、視覚・聴覚・言語が同時に伝達される**マルチモーダルな情報資源**として、自然言語処理や意味解析の研究対象としても注目されつつある。

本研究では、こうした背景を踏まえ、実際に視聴・研究可能な**調理系動画チャンネル**を複数選定した。動画の出典は主に YouTube および BiliBili の 2 つの動画投稿プラットフォームである。

YouTube はアメリカ発のグローバルな動画共有サービスであり、世界中の個人・団体がコンテンツを配信している。特に中国語圏のユーザーによるチャンネルも数多く存在し、海外に居住する中国語話者による料理動画も豊富に見られる。

一方、BiliBili（哔哩哔哩）は中国国内で人気の動画プラットフォームであり、若年層を中心にニッチかつ専門的な動画が投稿・視聴される傾向がある。コメントの弾幕文化や字幕支援機能の充実など、言語資料としての利点も多く、本研究では BiliBili も対象とした。

以上を踏まえ、YouTube では主に「阿朝哥美食」「美食作家王刚」「尚食厨房」などの登録者数・投稿数ともに安定しているチャンネルを選出し、BiliBili では「老东北美食」を例外的に収集対象とした。

以下の表に、各チャンネルの出典・登録者数・動画本数などの概要を示す。

対象動画チャンネル一覧

チャンネル	出典	登録者数	動画数
阿朝哥美食	YouTube	90.5 万人	1,906 本
阿慶師	YouTube	83.9 万人	738 本
大师的菜	YouTube	20.5 万人	756 本
尚食厨房	YouTube	27 万人	411 本
美食作家王刚	YouTube	214 万人	1,074 本
老东北美食	BiliBili	106.2 万人	3,753 本

3. 2. 動画のダウンロードとメタ情報取得

本研究では、調理動画から音声・字幕・構造的な付帯情報を高精度に取得するため、yt-dlp（YouTube Download Plus）というオープンソースツールを採用した。

yt-dlp は YouTube をはじめ、Vimeo や BiliBili など 1,000 以上の動画共有サイトに対応しており、以下のような多彩な機能を備えている：

- 動画・音声ファイルのダウンロード（フォーマット・解像度指定も可能）
- 字幕ファイル（`.srt`, `.vtt`）の取得と埋め込み
- JSON 形式のメタデータの抽出（動画タイトル、URL、投稿日、チャンネル名、再生数など）

加えて、`.json` メタ情報からは、再生数や動画長を用いたフィルタリング、また投稿時期ごとの分布分析、動画ごとの字幕品質の比較など、多角的な探索的データ解析（EDA）の基盤が構築されつつある。

3. 3. Whisper による文字起こし

調理動画に含まれる音声情報をコーパスとして活用するには、高精度な文字起こしが不可欠である。本研究では、OpenAI によって開発された多言語対応の音声認識モデル Whisper (Radford ほか, 2022) を用い、各動画から抽出した WAV 音声に対して文字起こしを行った。

Whisper は、世界中の複数言語の音声データで学習された大規模な自動音声認識（ASR）モデルであり、特に標準中国語（普通话）に対して非常に高い認識精度を持つ。本研究では、現在利用可能な中でも最も大規模で精度の高い `large-v3` モデルを選定した。

実際の処理では、各音声ファイルを一定のビームサイズと中国語指定で推論させ、タイムスタンプと発話単位が付加された出力を得た。これにより、後続の整形処理や意味的区切りの解析にも対応できる柔軟なデータ形式を確保することができた。

ただし、強い方言や背景音が多い動画、話速の早い場面では一部精度が低下する傾向も確認された。これらの課題については、後述する**ローカル LLM を用いた整形処理**によって補完することで、実用的なコーパス構築に資する文字列品質を担保している。

以上のように、Whisper の導入により、本研究は調理音声の大規模な自動文字起こしを可能とし、マルチモーダルコーパスの基盤構築に大きく貢献した。

また、中国語に特化した音声認識モデルとしては、中国・アリババ社が開発した FunASR (Gao ほか, 2023) も存在する。FunASR は、標準語に加えて北京方言などの地域変種にも対応した軽量モデル（SenseVoiceSmall）

一方で、Whisper は多言語に対応しつつ、中国語においても非常に高い精度を持ち、特に句読点や構文の安定感、雑音耐性において実用上の信頼性が高いことから、本研究ではまず Whisper を主要な ASR エンジンとして採用した。

今後の課題としては、Whisper と FunASR の出力結果を比較分析し、具体的にどちらが調理動画に適しているかを評価することが挙げられる。特に、方言や口語的表現が頻出する場面において、両モデルの出力傾向や誤認識のパターンを可視化・評価していく必要がある。

音声認識において「どこまでを正解とみなすか」は常に課題であり、単一の正解テキストが存在しない以上、一定の基準や検証方針をもって妥当性を担保していく姿勢が求められる。例えば、以下のような観点が今後の評価指標として想定される：

- 意味の通る文になっているか（可読性）
- タイムスタンプの正確さ（文単位の分離）
- 動詞や重要語の正確性（語彙情報の抽出に影響）

- 特定話者の訛り・速さへの耐性

本稿では音声認識精度の検証を定量的には行っていないが、今後の展望として、複数モデルを用いた出力比較および人手による部分的アノテーションによって、現実的な精度評価基準を構築していく必要がある。

3. 4. LLM を用いた整形処理

Whisper によって得られた文字起こし結果には、以下のような課題が見られた：

- タイムスタンプによる不自然な文分割（途中で切れる文）
- 同一発話内容の繰り返しや冗長な記述
- 意味のない繰り返し語（例：「然后、然后……」）

これらを放置すると、コーパスとしての構造的な一貫性が損なわれ、後続処理（動詞抽出、クラスタリング）にも悪影響を及ぼす。そのため、本研究ではローカル大規模言語モデル（LLM）である Qwen3-14B を用いて、文整形（cleaning）を自動化する手法を導入した。

Qwen3-14B は、Alibaba Group によって開発された中規模のオープンウェイト言語モデルであり、14.8 億パラメータという比較的軽量な構成ながら、推論精度と実行効率のバランスに優れる点が特徴である。多言語対応の設計で、中国語と英語に特に高い性能を発揮し、日本国内の研究環境でもローカル実行可能な LLM の選択肢として注目されている。 具体的には、Ollama で Q4_K_M として量子化されたバージョンを使用した。

整形処理では、Qwen3-14B に「重複文の除去」「無意味な繰り返しの削除」などのタスクを指示するプロンプトを与え、句読点の整備を含む自然な文単位への整理を実施した。なお、語彙の言い換えや再構成は明示的に禁止し、コーパスの忠実性を保持する設計とした。

現時点では、この整形処理の精度や妥当性に関する定量的な評価（例：文区切りの正確性や冗長率の変化）は行われておらず、今後の検証課題とする。今後は、正規表現やルールベース手法と比較しながら、LLM による整形の意義と限界を明らかにしていく予定である。

このような整形工程は、単なる見た目の調整にとどまらず、後続の形態素解析や意味役割ラベリングの品質に直結する重要な処理であり、ローカル LLM の応用可能性を示す実践例でもある。

3. 5. HanLP によるトークナイズ

本研究では、中国語向けの高性能 NLP ライブラリである HanLP（Heterogeneous Annotated Natural Language Processing）を用いて、整形済みテキストに対する分かち書きおよび品詞タグ付け処理を行う準備を進めている。

HanLP は中国語処理を中心に設計されたオープンソースの総合 NLP ツールキットであり、以下のような特徴を持つ：

- 分詞（word segmentation）：文字列を意味単位で分割（例：「自然语言处理」⟶「自然语言」「处理」）
- 品詞付与（POS tagging）：単語に対して名詞・動詞・副詞などの品詞情報を付加
- 固有表現認識（NER）：人名・地名・組織名などの検出
- 構文解析・依存関係解析：文法構造の分析

これらはすべて、Transformer ベースの事前学習モデルにより実装されており、

現在のところ、本研究では HanLP を用いたトークナイズの基礎的な環境構築を完了している段階にある。今後は、分かち書きにより得られた単語リストから動詞・名詞など意味的に重要な語彙を抽出し、共起ネットワークの構築やクラスタリング分析へと展開する予定である。

これにより、調理動詞の用法やレシピ手続きの言語的構造を定量的に可視化し、調理行為の意味役割ラベル付与や LLM による言語生成の精度向上にもつなげていくことが期待される。

3. 6. XiaChuFang Corpus の参照

既存の中国語料理コーパス「XiaChuFang Recipe Corpus」を分析し、タグ構造や語彙階層の設計思想を参考にした。これにより本研究の設計方針を補強した。

特に、name（料理名）や recipeInstructions（調理手順）、recipeIngredient（材料）などのフィールドは、語彙・構文・意味の三層にまたがる言語データの抽出を可能にしており、自然言語処理タスクとの親和性が高い。本研究でもこの構造を参考に、動画音声から抽出されたテキストをもとに、料理タイトル・発話者・調理手順・使用食材・出典プラットフォーム（YouTube/BiliBili）・タグ情報といった多様な属性を含む構造化データとしてコーパスを整備することを目指している。

最終的には、以下のような構造を持つ JSON 形式のデータとして格納・再利用可能な設計を採用し、調理動詞の使用傾向、語彙の共起構造、ジャンル間の表現差などを定量的に分析できる基盤を構築することを目的としている。こうした形式は、将来的に RAG による検索・生成システムへの統合や、専門的調理文書の自動生成、学術的な料理文化研究への応用も視野に入れている。

4. 1. Chinese_Video_Corpus 構築における課題

現在、OCR 精度や字幕ファイルの解析処理に問題があり、完全な構築には至っていない。音声起こしは済んでいるが、字幕との対応づけやタイミング整合に課題が残る。

5. 今後の展望と応用

クラスタリングに基づく意味役割ラベル付与や、RAG による LLM 生成制御の実装を視野に入れる。さらに他ジャンル動画との比較や、中国語特有の構文への応用も計画中。

7. まとめ

本研究では、中国語の調理動画を対象とした実用的かつ拡張性のあるマルチモーダルコーパスの構築を進めており、その応用先としてローカル LLM との連携を見据えている。今後は、より精緻な整形処理、クラスタリングの自動化、および文法構造の抽出などを通じて、意味のある生成支援を実現していく。

参考文献

Gao, Z., Li, Z., Wang, J., Luo, H., Shi, X., Chen, M., Li, Y., Zuo, L., Du, Z., Xiao, Z., & Zhang, S. (2023 年,). FunASR: A Fundamental End-to-End Speech Recognition Toolkit. *INTERSPEECH*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022 年,). *Robust Speech Recognition via Large-Scale Weak Supervision*. https://arxiv.org/abs/2212.04356