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, 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，ASR（Automatic Speech Recognition）による文字起こし，さらに LLM による整形処理，そして自然言語処理ライブラリを用いたトークン化と構文解析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、他の調理コーパスとの比較を通じて，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 Zhou et al.(2024) は，さまざまな書籍や Web 上での質問，論文などのデータを収集・整備し，そのデータをもとに大規模コーパス「FoodEarth」を構築している．またそのデータをもとに，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」を開発している．

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」(Liu ほか, 2022) 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いる点である．しかしいずれも，動画をもとにした実際の話し言葉・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観察や，生成支援を目指していく．

以下に先行研究との比較を示す．

表 1: 先行研究との比較

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	レシピ共有サイト「下厨房」	動画（字幕・音声・OCR）
情報形式	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
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
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 等）

3. 構築方法

3. 1. 対象動画の選定

情報の発信手段として SNS (Social Network Service) の活用が一般化している．特に動画は，視覚・聴覚が同時に伝達され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ている．

本研究では，こうした背景を踏まえ，複数の調理系動画チャンネルを選定し，研究の対象とした．動画は主に YouTube および BiliBili の 2 つの動画投稿サイトより出典した．

まず YouTube であるが，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．今回の動画の多くはまずこのプラットフォームから取得した．

一方，BiliBili（哔哩哔哩）は中国国内の動画プラットフォームであり、若年層を中心に，日常的な投稿から専門的なものまで大変多くの動画が投稿・視聴される．

以下の表に今回取得した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。

表 2: 今回取得した調理系動画チャンネル一覧
(2025 年 8 月 1 日現在)

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および音声認識処理の準備

本研究では調理動画から音声・字幕・構造的な付帯情報などを取得するため，yt-dlp（YouTube Download Plus）というオープンソースツールを採用した. 今回は以下のような設定でダウンロードを行った．

- 画像および音声を bestvideo[height<=480]+bestaudio/best[height<=480] で選択
- 形式は MP4
- JSON形式でメタデータの抽出（動画タイトル、URL、投稿日、チャンネル名、再生数など）

また後述する音声認識処理（ASR: Automatic Speech Recognition）においては、一般に WAV（Waveform Audio File Format）と呼ばれる非圧縮の音声ファイル形式が広く用いられる．WAV は可逆的であり，音声劣化がないため，機械学習モデルや ASR の際に高精度な処理が可能である．

今回 YouTube 等から取得した動画は MP4 形式（映像と音声を含む圧縮コンテナ）で保存されるため，音声部分のみを WAV 形式に抽出・変換する必要がある．従ってこの変換に際し ffmpeg（(Tomar, 2006)）というライブラリを用い，16kHz モノラルの WAV ファイルを生成した．ここまでが ASR 処理の準備である．

3. 3. 音声認識処理 ASR 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 ASR モデルで

あり、多くの言語に対して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.以下が実際に音声から文字起こしされた結果の一部である．

表 3: 料理動画字幕：青椒牛肉**丝**（ピーマンと牛肉の細炒め）

時間	発話
0.00s - 1.20s	哈**喽**大家好我是王**刚**
1.20s - 3.00s	本期**视频**跟大家分享一道家常菜
3.40s - 4.60s	青椒牛肉**丝**
4.60s - 6.80s	首先我**们**准**备**牛里脊 300 克
6.80s - 8.60s	放入盆中清洗干**净**备用
9.00s - 11.00s	把牛肉**顺**着**纹**路片成薄片
13.60s - 16.00s	然后再**顺**着**纹**路切成粗**丝**备用

表 4: 日本語翻訳

時間	発話（日本語訳）
0.00s - 1.20s	こんにちは皆さん，王**刚**です
1.20s - 3.00s	今回の動画では家庭料理をひとつ共有します
3.40s - 4.60s	チンジャオロース（ピーマンと牛肉の細炒め）です
4.60s - 6.80s	まず牛ヒレ肉 300g を用意します
6.80s - 8.60s	ボウルに入れてきれいに洗い，準備しておきます
9.00s - 11.00s	牛肉を繊維に沿って薄切りにします
13.60s - 16.00s	そして繊維に沿って太めの千切りにして準備しておきます

高い認識精度を持つといえど，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の担保を試みた．

ASR の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．今後以下のような観点が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

ASR 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る：

- 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((Team, 2025))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る. 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. 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指示文(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た.

```
def make_prompt(whisper_text: str) -> str:
    return f"""# 指示
以下のテキストは，中国語の料理解説動画から取得された音声文字起こしの内容です.
タイムスタンプは削除済みです．文が途中で途切れていたり,句読点が不足しています.
これを自然な中国語の書き言葉に整形し，適切な位置に句読点（「,」「。」）を補ってください.
「。」のあとには改行を入れてください.
# 書き起こしされたテキスト
{whisper_text}
# 出力形式
整形された本文のみを出力してください．文意の削除・意訳・要約は行わないでください.
"""
```

図 1: 文整形タスクをおこわなせるプロンプト(一部抜粋)

その結果以下のような結果が得られた.

哈喽大家好，我是王刚。
本期视频跟大家分享一道家常菜，青椒牛肉丝。
首先，我们准备牛里脊 300 克，放入盆中清洗干净，备用。
然后，把牛肉顺着纹路片成薄片，再顺着纹路切成粗丝，备用。

図 2: プロンプト処理後の出力例

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．

4. 形態素解析と依存関係解析による構文情報の付与

本研究では，中国語向けの自然言語処理ライブラリである HanLP((He & Choi, 2021))を用いて，整形済みテキストに対する分かち書き，品詞タグ付け処理，依存関係解析などを行う準備を進めている。以下が先ほどのテキストに対して分かきおよびタグ付けを行った結果である．

哈喽/o 大家/r 好/a，/w 我/r 是/v 王刚/nr 。/w
本/r 期/q 视频/n 跟/p 大家/r 分享/v 一/m 道/q 家常菜/n，/w 青椒/n 牛肉丝/n 。/w
首先/c，/w 我们/r 准备/v 牛/n 里/n 脊/Ng 300/m 克/q，/w 放入/v 盆中/s 清洗/v 干净/a，/w 备用/v 。/w
然后/c，/w 把/p 牛肉/n 顺着/p 纹路/n 片/v 成/v 薄片/n，/w 再/d 顺着/p 纹路/n 切成/v 粗/a 丝/n，/w 备用/v 。/w

図 3: HanLP を用いた処理

本研究で扱った例文中に現れた品詞タグのうち、動詞に関するものを以下に示す

表 5: 本研究で確認された動詞関連 POS タグの一覧

POS タグ	意味（日本語）	説明
v	動詞	一般的な動作・状態を表す語（最も基本的な動詞タグ）
p	前置詞	動詞の補語や方向性を示す語（把、跟、顺着など）
c	接続詞	動作の順序や関係を表す語（首先、然后など）
d	副詞	動詞を修飾する語（再など）
a	形容詞	動詞の結果や状態を表す語（好、干净、粗など）

また上記の分析をもとに，以下のような依存関係を図式化した．

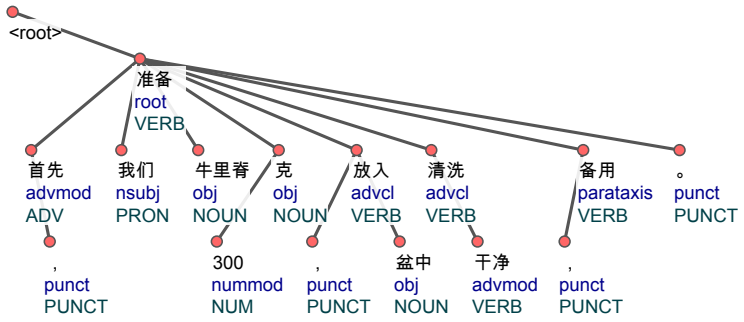

図 4: 構文解析ツリーの例（HanLP + UD Viewer）

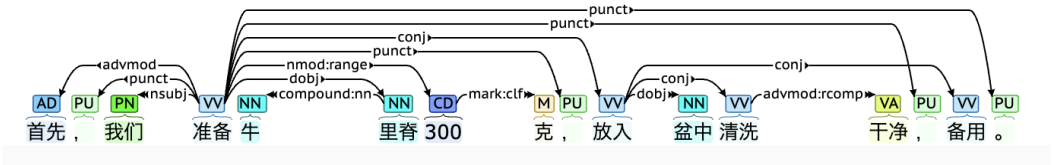

図 5: 依存構造解析図（HanLP による出力）

今後の分析においては，構文解析結果を用いて各動詞に対する依存関係を抽出し，その動詞と関係の深い品詞を洗い出す．一例で言えば，直接目的語（dobj）との関係に注目することで，調理動詞がどのような対象を操作しているかを明らかにすることができると考える．具体的に言えば，「炒」や「煮」などの主要な調理動詞に対して，その直後に現れる目的語句を抽出・集計し，頻出組み合わせを統計的に分析する．さらに動詞に付随する副詞的要素（advmod）や形容詞的修飾（amod）なども合わせて抽出することで，動作の様態（例：「素早く炒める」「充分に煮る」）に関する共起パターンを明らかにすることを目指す．

これらの処理は，HanLP によって得られた依存構造の出力を用い，対象となる動詞をキーとし，その係り先（依存子）や係り元（支配語）の構造を逐語的に解析することで実装する予定である．

加えて，本研究では動画というマルチモーダルな情報源を出発点としているため，文字起こしされた言語情報とともに，各調理工程にかかる実際の時間や調理対象の変化といった視覚的特徴を併せて確認できるという利点がある． これにより各動詞が示す意味が，時間的持続性や調理結果との対応関係を伴った実体的な知識として捉えられる．将来的には，このような視覚情報との統合を通じ，生成モデルによる料理説明の自然さや正確さをさらに高める応用も期待される．

5. コーパスの設計

本研究におけるコーパスの設計に際し，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．今後は動画音声から抽出されたテキストやメタ情報をもとに，タイトル・投稿者・調理法・使用食材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．

表 6: 動画字幕コーパスへの応用提案

項目	説明	例
video_title	動画のタイトル	青椒牛肉**丝**的正宗做法
author	投稿者	王**刚**
date	投稿された日付	20240228
dish_name	調理名	青椒牛肉**丝**
ingredients	主な材料	牛里脊, 青椒
cooking_methods	調理法(動詞)	切**丝**, 炒制, **调**味
cooking_time	調理(動画)時間	10 分**钟**

最終的には，上記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指す.

6. まとめ

本研究では，中国語調理動画を対象とした包括的なマルチモーダルコーパスの構築手法を提案し，その実用性と拡張性の検証を目指している．音声文字起こし，テキスト整形，品詞解析，構文解析という一連の処理パイプラインを通じて，従来には見られなかった，動画コンテンツからの言語学的に価値の高いデータを体系的に抽出することを計画している． 特に，音声認識と LLM を用いたテキスト整形の組み合わせにより，中国語の口語的表現の文字起こし際の高い精度の実現を目指し，後続の言語処理の品質向上への貢献を期待している．構築予定のコーパスでは，調理動詞の使用傾向分析，語彙共起パターンの抽出，文体・ジャンル特性の定量化など，多角的な言語分析を可能にしたいと考えている．これにより，従来の小規模なテキストコーパスでは実現困難であった，実用的言語使用の大規模解析の実現を目指している．今後は，意味役割ラベル付与の自動化，RAG システムとの統合，多言語展開などを通じて，本コーパスの学術的・実用的価値の向上を図っていく予定である．特に，ローカル LLM との連携による調理支援システムの構築は，言語学研究の社会実装という観点からも含め，意義のあるものとしたい.

参考文献

He, H., & Choi, J. D. (2021 年). The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5555–5577. https://aclanthology.org/2021.emnlp-main.451

Liu, X., Feng, Y., Tang, J., Hu, C., & Zhao, D. (2022 年,). *Counterfactual Recipe Generation: Exploring Compositional Generalization in a Realistic Scenario.* https://arxiv.org/abs/2210.11431

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

Team, Q. (2025 年,). *Qwen3 Technical Report.* https://arxiv.org/abs/2505.09388

Tomar, S. (2006 年). Converting video formats with FFmpeg. *Linux Journal*, *2006*(146), 10.