Large Language Models in Linguistic Research: the Pilot and the Copilot

Svetla Koeva
Institute for Bulgarian Language, Bulgarian Academy of Sciences
svetla@dcl.bas.bg

Abstract

In this paper, we present two experiments focussing on linguistic classification and annotation of examples, using zero-shot prompting. The aim is to show how large language models can confirm or reject the linguistic judgements of experts in order to increase the productivity of their work. In the first experiment, new lexical units evoking a particular FrameNet semantic frame are selected simultaneously with the annotation of examples with the core frame elements. The second experiment attempts to categorise verbs into the aspectual classes, assuming that only certain combinations of verbs belonging to different aspectual classes evoke a semantic frame. The linguistic theories underlying the two experiments, the development of the prompts and the results of the experiments are presented.

Keywords: LLMs, linguistic research, prompt engineering

1 Introduction

Until recently, natural language processing (NLP) relied on specialised language resources such as monolingual, bilingual and multilingual corpora as well as lexical and conceptual resources to develop functional applications. However, breakthroughs in artificial intelligence and the emergence of large language models (LLM) have changed the field, enabling the successful completion of a variety of NLP tasks in a completely different way.

Widely studied applications of LLMs include document intelligence tasks, such as sentiment analysis (Krugmann and Hartmann, 2024), text classification (Sun et al., 2023), risk prediction (Cao et al., 2024), information extraction (Peng et al., 2024), and many others. In addition, LLMs are used for machine translation (Zhu et al., 2024), content creation tasks such as creative writing, automatic sentence completion, paraphrasing, personalised decision making and code generation. LLMs play a crucial role in virtual assistants that facilitate various applications such as language understanding, speech generation and speech recognition (Wang et al., 2022; Arora et al., 2024).

The increasing use of large language models can be expected not only for traditional NLP tasks, but also to tackle typical linguistic challenges. The recent application of LLMs in linguistic research can be outlined in several directions. Most attempts aim at using LLMs for linguistic annotation to facilitate corpus-based linguistic studies by automatically annotating texts with targeted linguistic information (Kuzman et al., 2023; Gilardi et al., 2023), among others, but there are also attempts to use LLMs for theoretical investigations within specific linguistic frameworks (Beguš et al., 2023; Torrent et al., 2024).

In this paper we offer two experiments focusing on the linguistic classification and annotation of examples. The aim is to show how some newly released LLMs can confirm or reject the linguistic hypotheses of experts in order to increase the productivity of their work. In the first experiment, new proposals for lexical units evoking a particular FrameNet semantic frame are classified simultaneously with the annotation of examples with the core frame elements. In the second experiment, verb lexical units are categorised into different semantic classes, whereby it is assumed that a certain semantic frame can only be evoked by verbs that belong to certain combinations of semantic classes. The linguistic theories on which the two experiments are based, the development of the prompts and the results of the experiments are briefly presented.

The paper deals with the following topics: Section 2 gives an overview of the use of large language models in linguistic research. Section 3 briefly introduces the large language models used in the experiments. Section 4 describes experiments with two types of prompts: a) for the si-

multaneous augmentation of FrameNet semantic frames with new lexical units and annotated examples and b) for the classification of lexical units into semantic classes. The presentation of both experiments includes a description of the linguistic theories, the prompt structures and an overview of the results. Section 5 contains a conclusion on the benefits and challenges of using LLMs in linguistic research based on the experiments conducted.

2 LLMs and linguistic research

Prompt engineering has already been used to solve linguistically relevant tasks. In few-shot learning, only a few examples are given to the model during inference and task definition (Brown et al., 2020: 5-7). One-shot learning and zero-shot learning are similar to few-shot learning (Wei et al., 2023), but as the names suggest, only one or zero demonstrations are allowed to formulate a task description in natural language.

Recently, the ability of LLMs to recognise and classify specific language constructions, to analyse data within a theoretical framework, to annotate texts with relevant linguistic information to support corpus-based linguistic studies, etc. has been investigated. The following is a brief overview of the applications of LLMs in the field of linguistics, formally divided into: recognition of language constructions, analysis of data within a theoretical framework and linguistic annotation, as some of the studies combine more than one approach.

2.1 Recognition of language constructions

GPT-3 (Brown et al., 2020) knowledge of rare constructions with semantic and syntactic constraints, such as the construction *indefinite article + adjective + numeral + noun* in English (e.g. "a lovely five days"), was assessed (Mahowald, 2023). The acceptability judgements of the GPT-3 for this construction were compared with human judgements on a range of sentences to conclude that the GPT-3 judgements are broadly similar to human judgements.

Another study investigated the extent to which the word frequency data of LLMs match the data of a large general corpus; the collocation data of LLMs match a large general corpus; and LLMs can recognise lexico-grammatical patterns and perform genre categorisation (Uchida, 2024). ChatGPT 3.5 showed a high agreement with the COCA (Corpus of Contemporary American English) ranking in terms of word frequency, but varied when analysing certain word types due to more repetitions. The overall agreement for the collocation patterns tested was 42.8%. When examining open slots in grammatical patterns, more than half of the words in ChatGPT 3.5 matched the top 20 of COCA, and more than 65% were within the top 40, showing effective verification of lexico-grammatical patterns. However, the hit rate for genre identification was low at both word and text level.

The book *Copilot for Linguists* introduces the concept of using LLM chatbots as a tool for trained linguists and shifts the focus from what these chatbots can achieve with language to how they can support linguists in their work (Torrent et al., 2024). Experiments are presented in which LLM chatbots were prompted to analyse grammatical constructions and to enrich FrameNet semantic frames in both English and Brazilian Portuguese. Prompt engineering techniques derived from these experiments were shared, and the potential of LLMs to act as copilots for construction grammarians in linguistic research was explored, highlighting their ability to recognise instances of fully developed constructions, analyse their syntax and understand their meaning. The study aims to demonstrate the valuable role that LLM chatbots can play in supporting the analytical endeavours of linguists. However, the limitations are also pointed out, e.g. when analysing constructions in languages other than English and when understanding the semantics of language constructions.

2.2 Formal analysis of linguistic data

Efforts have been made to show that LLMs can produce coherent and valid formal analysis of linguistic data (Beguš et al., 2023). The metalinguistic capabilities of GPT-3.5 and GPT-4 (OpenAI, 2023) were tested, focussing on three subdomains of formal linguistics: syntax, phonology and semantics. It was shown that GPT-4 is able to analyse both relatively simple and more complex syntactic structures largely correctly, while GPT-3.5 performs worse on the same tasks. GPT-4 was tested on two phonological problems by prompting the model with small datasets: a palatalisation process in Korean and a spirantisation process in an artificial language, and GPT-4 copes well with both. The ability of GPT-4 to produce lambda calculus analyses of English sentences, including cases of scopal ambiguity (e.g. *Every student likes a classmate*), was

tested. It was found that the model works well with simpler sentences and understands scopal ambiguity, but makes some significant errors when using the lambda calculus formalism. Overall, GPT-4 is largely (but not perfectly) able to produce coherent analyses of simple problems in each of the three domains tested, detect ambiguity, correct its own analytical errors, and comment on the feasibility of multiple solutions.

2.3 Linguistic annotation

The performance of ChatGPT and the multilingual XLM-RoBERTa language model (Conneau et al., 2019) was evaluated on the genre identification task to determine which model is best suited to enrich large web corpora for English and Slovenian with genre information (Kuzman et al., 2023). The two models are compared in three scenarios, switching between the languages of the prompts and the test datasets. The results show that Chat-GPT outperforms the fine-tuned model when applied to a completely new test dataset. However, when the model is fully prompted in Slovenian, performance drops significantly, demonstrating the limitations of using ChatGPT in smaller languages at the time of the experiments.

It was evaluated how ChatGPT performs subjective tasks related to social norms and cultural context, such as identifying implicit hate speech online and providing explanations for it (Huang et al., 2023). The results show that ChatGPT correctly identifies 80% of implicit hate tweets in the experimental setup, demonstrating its potential as a data annotation tool with a simple prompt design. However, it was noted that there is a risk of misleading non-experts if the model's decisions are incorrect.

Another study showed that ChatGPT classifications with zero-shots are better than MTurk (Amazon Mechanical Turk, a crowdsourcing marketplace) annotations (Gilardi et al., 2023). The analysis was performed on a sample of 6,183 documents, including tweets and news articles. Several annotation tasks were implemented, e.g. *relevance* to determine whether a tweet relates to content moderation or politics; *topic detection* to determine whether a tweet falls into one of six predefined topics; *stance detection* to determine whether a tweet supports, opposes or remains neutral on a US law, etc. The performance of ChatGPT was evaluated based on accuracy and intercoder agreement. The results showed that ChatGPT's zero-shot accuracy outperformed the crowd workers by an average of 25 percentage points, while ChatGPT's intercoder agreement outperformed both the crowd workers and the trained annotators on all tasks.

The performance of LLMs was evaluated on the task of annotating local grammars, focussing on the speech act of apology in English (Yu et al., 2024). The analysed corpus contained 5,539 instances of the word *sorry*, extracted from the Spoken British National Corpus 2014. The experimental setup involved few-shot prompting techniques and included the following steps: comparing the performance of GPT-4 based Bing Chatbot and ChatGPT 3.5 when annotating 50 instances and comparing the performance of Bing Chatbot and a human annotator when annotating 1000 instances. The results show that the Bing Chatbot performs better than ChatGPT 3.5. Although the human annotator achieved slightly higher accuracy than the Bing Chatbot, the latter showed robust abilities to understand both the semantic and pragmatic aspects of the language.

The conclusions that can be drawn from the brief overview of the use of LLMs for linguistic (or related) tasks are that they have the greatest application in the recognition of language units and their classification, i.e. in linguistic annotation. Automatic annotation of texts was not absolutely accurate even before LLMs (with varying degrees of success for different annotation tasks) (Liao and Zhao, 2019), but has been successfully used for applications requiring large annotated training data. Currently, the refinement of the prompting process to obtain correct information from the LLMs and the progress of the LLMs themselves can lead to positive results in several directions: creation of large semantic language resources and development of benchmarking datasets to evaluate LLMs for various linguistic tasks.

In this study, we will analyse the potential of current LLMs for the creation of **semantic language resources**, in particular for the enrichment of FrameNet's semantic frames with new lexical units and annotated examples, as well as for the semantic classification of lexical units.

3 Large language models used in the experiments

There are already several surveys on large language models describing the history of pre-training and

breakthroughs (Han et al., 2021; Zhou et al., 2023), the scaling and impact of pre-trained models (Wang et al., 2022), the prompting methods in natural language processing (Liu et al., 2023), the multimodal pre-trained models (Wang et al., 2023), the recent advances in LLMs with introduction to the background, key results and mainstream techniques (Zhao et al., 2023), the evaluation methods for LLMs (Chang et al., 2023), the comparison between some of the most popular LLMs, including the three families: GPT, LLaMA and PaLM, and the discussion of their features, contributions and limitations (Minaee et al., 2024), among others.

Here we will briefly introduce the LLMs that were used in the experiments to illustrate the capabilities of LLMs to assist the creation of large semantic language resources: (in alphabetical order) **Claude 3.5 Sonnet**, **Gemini 1.5 Pro**, **GPT-4o**, **GPT-4o mini**. The models were last accessed on 21 and 22 July 2024.

It is not our aim to compare these LLMs for several reasons: a) models are evolving very fast and new models and updates of existing models are constantly appearing; b) some of the models have been improved in a clearly defined direction; c) for some of the selected LLMs, detailed information about architecture, training data, model size, etc. is not available. Therefore, we do not focus on comparing models, but on exploring a possible way to work with LLMs on a specific linguistic task.

The **Claude 3** is a family of multimodal models released by Anthropic[1]: Claude 3 Opus, Claude 3 Sonnet and Claude 3 Haiku. Claude 3 has been trained with a mixture of public and proprietary data and is subject to rigorous cleaning and filtering methods. All models have image processing and show good performance in logical reasoning, maths and coding (Anthropic, 2023). Claude 3, especially the Opus model, reportedly outperforms other state-of-the-art models in various evaluation benchmarks such as the Google-Proof Question-Answering Benchmark (GPQA), Measuring Massive Multitask Language Understanding (MMLU) and others.

Sonnet has been updated to **Claude 3.5**, the latest version of the Claude LLMs[2]. In Anthropic's tests, the Claude 3.5 Sonnet outperforms some of the latest LLMs and other Claude models. The Claude 3.5 Sonnet has powered the Claude chatbot. It is also available via the Anthropic API.

The **Gemini 1.5** family of Google DeepMind aims to retrieve and analyse millions of context tokens, including multiple long documents, video and audio materials (Gemini Team, Google, 2024). To achieve this, the models are trained on multiple 4096-chip pods of TPUv4 accelerators, using a wide range of multimodal and multilingual data that includes image, audio and video content.

Gemini 1.5 Pro[3] is a sparse Mixture-of-Expert (MoE) Transformer-based model that builds on the advances of Gemini 1.0 (Gemini Team, Google, 2023) in a variety of multimodal tasks such as visual understanding, classification, summarisation and content creation from image, audio and video. The extensive evaluations with diagnostic multimodal long-context benchmarks show that Gemini 1.5 Pro is able to retrieve and understand large amounts of data. The model is available via the Gemini Chatbot and the Gemini API.

GPT-4 is a large multimodal model that accepts both image and text input and generates text output (OpenAI, 2023). It is reported that GPT-4 understands and generates text in more languages compared to its predecessor and outperforms GPT-3.5 and other large language models in a number of traditional NLP benchmarks and the MMLU benchmark.

The next generation, **GPT-4o** (omni)[4], has been integrated into the text, vision and audio modalities through techniques such as filtering training data and refining model behaviour through post-training. According to OpenAI, GPT-4o matches the performance of GPT-4 Turbo in the areas of text, reasoning and coding intelligence evaluated with traditional benchmarks, and offers multilingual, audio and vision processing. GPT-4o is available via the ChatGPT Plus and the OpenAI API.

The **GPT-4o mini**[5] is a smaller version of the GPT-4o model. GPT-4o mini has reportedly outperformed GPT-3.5 Turbo in several LLM benchmarks and is trained with an instruction hierarchy method (Wallace et al., 2024) that improves the model's resistance to jailbreaks and system prompt extraction. In ChatGPT, GPT-3.5 was replaced by GPT-4o mini. It is also available via the OpenAI API.

[1] https://www.anthropic.com/news/claude-3-family
[2] https://www.anthropic.com/news/claude-3-5-sonnet
[3] https://deepmind.google/technologies/gemini/pro/
[4] https://www.infoq.com/news/2024/05/openai-gpt4o/
[5] https://platform.openai.com/docs/models/

4 Prompting-based experiments to facilitate linguistic research

We present two experiments with zero-shot prompting aimed at linguistic classification and annotation of examples. The aim is to demonstrate the effectiveness of LLMs in making linguistic decisions, thus increasing the expert's confidence in creating semantic resources with reliable information without the need for a second or third expert to perform the same activities.

The first experiment (Augmenting FrameNet semantic frames with lexical units and annotations) simultaneously aims to select new lexical units that evoke a particular FrameNet semantic frame, suggest relevant examples and annotate them with the core elements of the frame. The second experiment (Classification of FrameNet lexical units into semantic classes) aims to categorise verbs that evoke a particular FrameNet semantic frame into relevant semantic classes.

The general framework can be summarised as follows: a) formulation of a linguistic task; b) selection of suitable LLMs (mainly according to two criteria: novelty and accessibility); c) formulation of a prompt template aimed at fulfilling the linguistic task; d) execution of the experiment and collection of data from the selected LLMs; e) human evaluation of the results obtained.

4.1 Augmenting FrameNet semantic frames with lexical units and annotations

The annotation of new examples for existing frames with the syntactic realisation of frame elements and lexical units and the discovery of new lexical units for existing frames is the most common extension to improve lexical coverage and representation (Torrent et al., 2024).

An approach to expanding lexical units that evoke a semantic frame utilises the links between lexical units that evoke a frame and lexical units in other resources to discover potential new lexical units. For this task, vector representations of lexical units and clustering techniques are used (Yong and Torrent, 2020).

Our experiment aims at a simultaneous extension of the semantic frames of FrameNet with new lexical units and annotated examples. For the expansion of lexical units, we use a list of potential candidates that are accepted or rejected by LLMs, in contrast to other approaches that use direct instructions to LLMs to propose new lexical units (Torrent et al., 2024). Furthermore, we aim at a full-text annotation with the core frame elements of the synthetic examples provided by the LLMs.

4.1.1 Frame semantics in brief

FrameNet is based on the theory of frame semantics (Fillmore, 1976, 1982).

The central idea of frame semantics is that word meanings are described in terms of semantic frames, which are schematic representations of the conceptual structures and patterns of beliefs, practices, institutions, images, etc. that provide a foundation for meaningful interaction in a given speech community (Fillmore et al., 2003: 235).

FrameNet[6] is a collection of semantic frames that contain a common abstract semantic representation for a set of lexical units and valency patterns that represent semantic and syntactic descriptions based on the annotation of examples. The semantic frame in FrameNet includes the following components: frame name; informal definition of the situation that the frame represents; semantic type of the frame (optional); set of frame elements associated with the frame (core and non-core: peripheral, extrathematic and core-unexpressed); relations between frame elements, if any; frame-to-frame relations, if any; and the lexical units that evoke the frame.

The frame element information includes the name of the frame element, its informal definition, the semantic type (optional), and examples illustrating the use of the frame element (optional).

The information on the lexical units includes a definition, a semantic type (optional), annotated examples and derived valency patterns that show the correspondence between the frame elements and their syntactic realisation.

An excerpt from the semantic frame **Arriving** is shown to illustrate this. **Arriving** frame in FrameNet has the definition: "An object **Theme** moves in the direction of a **Goal**. The **Goal** may be expressed or it may be understood from context, but its is always implied by the verb itself." The core frame elements are: **Theme** defined as "**Theme** is the object that moves. It may be an entity that moves under its own power, but it need not be" and **Goal** defined as "**Goal** is any expression that tells where the **Theme** ends up, or would end up, as a result of the motion".

The verbs that evoke this frame are: appear.v, ap-

[6] http://framenet.icsi.berkeley.edu

proach.v, arrive.v, come.v, crest.v, descend (on).v, enter.v, find.v, get.v, hit.v, make it.v, make.v, reach.v, return.v, visit.v.

4.1.2 Prompt design

For our experiment, we use semantic and lexical information from FrameNet, in particular from the semantic frame **Arriving**, focussing on lexical units of verbs. There are several preparatory steps: the selection of potential new lexical units for the **Arriving** frame and the experimental formulation of an appropriate prompt, aiming at a simultaneous evaluation of the selected lexical units and suggestions for annotated examples to illustrate the core elements of the frame.

We consider WordNet (Miller et al., 1990) as a natural source of new lexical units for existing frames in FrameNet, since WordNet contains about 117 000 synsets[7].

The first step is to match the lexical units of a particular semantic frame with the corresponding synsets from WordNet. Although previous mappings between FrameNet and WordNet are known (Shi and Mihalcea, 2005), the method used is relatively simple and combines the exact matching of lexical units from FrameNet with literals from WordNet and the evaluation of the similarity of the definitions. This approach takes into account the fact that WordNet contains many literals with the same form but different meanings. After mapping, a human judgement is made to decide whether a particular mapping between a lexical unit in FrameNet and a WordNet literal is correct.

Among the lexical units that evoke a particular semantic frame, there are synonyms, hypernyms and hyponyms (troponyms for verbs), although the existing semantic relations between lexical units are not explicitly labelled. All troponym synsets up to a hypernym synset whose literal(s) are mapped with a lexical unit from FrameNet are considered potential candidates. If the mapping is rejected by an expert, the corresponding troponyms are also ignored.

The results of the mapping show that one lexical unit is not present in WordNet, three mapped synsets have no troponyms and in two mappings the synset appears as a troponym of another mapped synset. The number of potential new lexical units evoking the semantic frame **Arriving** is **137**.

Each prompt contains the name of the targeted semantic frame, its definition, the core elements of the frame with their definitions and a list of the new lexical units provided with definitions from WordNet. The desired output format is also included. The prompt template has the following structure:

The semantic frame "frame name" *in FrameNet has the definition:* "frame definition". *The core frame elements are:* "name of core frame element" *defined as* "frame element definition".

Indicate which of the following verbs: "verb": "definition", *evoke the semantic frame* "frame name" *and those that do not; give three examples of each verb and annotate the core frame elements in the examples. Use this pattern:*

"verb": "definition"

Example 1: "example" "core frame element name": "annotation"

The Appendix A illustrates a prompt with four verbs.

Preliminary tests were carried out to determine the optimal format for the prompts according to human judgement, with the aim of achieving a satisfactory level of completeness of responses.

4.1.3 Results and discussion

The prompt requires an assessment of the relevance of the lexical units to a semantic frame, the provision of relevant examples and the annotation of the examples with the core elements.

The classification of lexical units according to whether or not they evoke the semantic frame **Arriving** is linked to the correct interpretation of the verb meaning by the LLMs. In some cases, the use of the lexical units in the synthetic examples that do not illustrate the intended meaning is classified, e.g. *The company **scaled** back production to match declining demand*. However, some unintended verb meanings that are illustrated with examples are correctly categorised as not evoking the semantic frame **Arriving**, e.g. *The new policy aims to **get at** the root causes of inequality*. In some cases, models provide clues that can be helpful for an expert's final decision, e.g. ***To get at** implies reaching an abstract or physical destination making it relevant to the **Arriving** frame as it involves a **Theme** moving towards a defined **Goal**.*

The examples given are most likely synthetic, but it is also possible that they are contained in the training data. Regardless, all examples are grammatically correct and sound natural. They are structured to consist of parts interpreted as **Theme** and **Goal**, but usually do not contain more than one

[7] https://wordnet.princeton.edu

LLM	Lexical units			Annotation		
	P	R	F1	P	R	F1
Claude 3.5 Sonnet	0.87	0,58	0.73	0.77	0.53	0.63
Gemini 1.5 Pro	0,80	0,66	0,72	0.80	0.40	0.53
GPT-4o	0.87	0.53	0.66	0.86	0.48	0.61
GPT-4o mini	0.58	0.87	0.69	0.50	0.75	0.60

Table 1: Results from the experiment "Augmenting FrameNet semantic frames").

clause and do not illustrate complex syntactic structures.

The annotation of core frame elements is correct in most cases, as long as the examples represent the given lexical unit and not others with a different meaning, e.g. *The hikers surmounted the peak and gazed out at the breathtaking view* (**Theme**: hikers, **Goal**: peak). Again, the use of simple syntactic structures, which are always explicit, is a shortcoming when it comes to illustrating the varied use of language.

Table 1 shows the calculated results compared to the manual annotation of 20% of the outputs. As mentioned above, these calculations cannot be considered relevant for the evaluation of the LLMs and some of the results are not readily comparable. The manual annotation for lexical units simultaneously takes into account the correct classification of the lexical unit as part of the semantic frame **Arriving** and the correct suggestion of examples; and for the annotation – the correct suggestion of examples and the correct annotation of the core frame elements. With this approach of simultaneous evaluation of two components, the precision and recall values are reduced. For example, the combination of a correct classification of a verb and an example that illustrates a different meaning is scored as a true negative, as is the correct annotation of a core frame element in an inappropriate example.

The partial manual annotation of the LLM output shows that the proposed approach cannot completely replace the manual development of semantic resources or manual annotation: both the resulting verb list, the examples and the annotations have to be manually evaluated and in some cases rejected or re-annotated. However, the experiment shows that the LLMs can be used as a second annotator when enriching the FrameNet with new lexical units (different models can be selected as more suitable for different linguistic tasks). Furthermore, since the annotation with the core frame elements was correct in a large number of examples reflecting the meaning of the verbs, it is possible to successfully use LLMs for the automatic enrichment of FrameNet annotations for examples selected by experts. There is also great scope for improving the prompt(s) with instructions on exactly how to annotate the frame elements (e.g. inclusion of modifiers and articles).

4.2 Classification of FrameNet lexical units into semantic classes

The second experiment aims to categorise the verbs that evoke a certain semantic frame into semantic classes, in this case aspectual classes (related to situation types, also called eventuality types): *states*, *activities*, *accomplishments* and *achievements*. This classification is relevant for the differentiation of activities, accomplishments and achievements in frames that represent events, as the states in FrameNet are grouped in separate semantic frames.

4.2.1 Verb aspectual classes in brief

The categorisation of verbs into acpectual classes is based on the following characteristics (Vendler, 1957): **change** with the values dynamicity and stativity; **temporal extent** with the values durativity and punctuality; **defined endpoint** (homogeneity) with the values telicity and atelicity, and comprises four situation types:

– **states** – continuous stative situations that are atelic;

– **activities** – continuous dynamic situations that are atelic;

– **accomplishments** – continuous dynamic situations that are telic;

– **achievements** – punctual dynamic situations that are telic.

This classification was followed by numerous (more detailed) classifications of situation types, including (Kenny, 2003; Dowty, 1997; Piñón, 1997), which aim to describe in a conventional way the situations that are important for the semantic representation of verbs and their argument structure. The grouping of verbs within a semantic frame ac-

cording to aspectual classes is also reflected in the grouping of the frame elements and the valency patterns associated with the verbs.

The proposed experiment aims to classify verbs that evoke a particular semantic frame into three semantic classes: activities, accomplishments and achievements.

4.2.2 Workflow and prompt design

The identification of semantic classes of verbs is intended to show whether activities, accomplishments and achievements can coexist in semantic frames.

The workflow includes the selection of verbs to be tested from FrameNet, the selection of suitable linguistic tests and the creation of the prompt(s).

The selected verbs are verbs that evoke the semantic frame **Arriving** from FrameNet, as well as verbs that were added by mapping with WordNet and evaluated by an expert (20 in total). Due to their similarity in meaning, they can be conditionally divided into two groups: the group of ***arrive*** verbs and the group of ***approach*** verbs.

There are several linguistic tests that have been formulated to distinguish between different aspectual classes of verbs (Dowty, 1997). We have selected tests that are only relevant for distinguishing between activities, accomplishments and achievements:

– The verb occurs with expressions such as *for an hour*, which means that it occurs at any point in the hour. When the result is positive, activities are clearly distinguished from achievements and accomplishments.

– The verb can occur with the verb *finish*. If the result is positive, accomplishments are clearly distinguished from activities and achievements.

– The verb can occur with the verb *stop*. If the result is negative, achievements are clearly distinguished from accomplishments and activities.

After some experiments, the prompt is structured as follows:

There is a list of verbs:

"verb": "definition";

"verb": "definition".

Give examples with each of the verbs from the list in the following constructions if the examples are grammatically correct.

"Theme" is verb-*ing* "Goal" ***for an hour***.

"Theme" ***finish*** verb-*ing* "Goal".

"Theme" ***stop*** verb-*ing* "Goal".

The Appendix B illustrates a prompt. The experiment was conducted with the same LLMs.

4.2.3 Results and discussion

The list of verbs for the experiment contains only verbs that can be categorised as activities, accomplishments and achievements. Some models, such as Claude 3.5 Sonnet and Gemini 1.5 Pro, correctly distinguish between two classes: activities/accomplishments (the group of **approach** verbs) and achievements (the group of **arrive** verbs), assuming that activity verbs can co-occur with the *for*-expression and the verb *stop*, accomplishment verbs can co-occur with the verbs *finish* and *stop*, while achievement verbs cannot co-occur with the *for*-expression and the verbs *finish* and *stop*. The output also contains some explanations:

Claude 3.5 Sonnet: ***Note***: *Verbs that describe punctual actions (arrive, drive in, enter, get, come) or don't imply a gradual process (come on, go up) don't fit well with these constructions in their given meanings*.

Gemini 1.5 Pro: ***Approach***: *Indicates movement towards a goal, but doesn't necessarily mean reaching it*. ***Arrive*****:** *Indicates the completion of a journey, reaching the intended destination*.

The other models give examples for all constructions, even if they are not always grammatically correct. All models provide some examples that are correct but illustrate either a different meaning or a different construction.

A clear distinction between activities and accomplishments cannot be made in this experiment.

The group of **approach** verbs is provided with some examples with *for*-expressions that are typically combined with activities, such as:

The ship approached the island for an hour. (This implies a slow, gradual approach.)

The car neared the city for an hour. (Similar to approach, implies a gradual movement.)

The *for*-expressions, like the *in*-expressions, are treated semantically as a way of measuring the scope of eventualities. An *in*-adverbial measures the time span in which eventualities expressed by telic predicates culminate, while a *for*-adverbial measures the temporal duration of eventualities denoted by atelic predicates (Filip, 2011). In our experiment, the combination with a *for*-expression should signal that the verbs are activities and not accomplishment or achievements. However, there can be shifts between telic and atelic interpretations of a verb and in some cases the *for*-expression

can be combined with accomplishments. In the sentence *The guests **arrived** at the wedding **for an hour***, the *for*-expression means that the guests intended to stay at the wedding for an hour, and the interpretation is telic.

Many linguistic works emphasise that aspectual distinctions are distinctions between linguistic expressions and not properties of events (Rothstein, 2004). For example, accomplishment verbs can differ in their telicity according to the properties of their direct objects (Verkuyl, 1972, 1993). An accomplishment verb is usually the head of a telic verb phrase, but the verb phrase is atelic if the direct object is a bare plural or a mass noun. When a verb is an activity, the properties of the direct object have no influence on the telicity of the verb phrase.

One of the main differences between activities and accomplishments are the **goals** incorporated in the meaning of accomplishments, which is also a feature of the **approach** verbs.

The linguistic tests with the verbs *finish* and *stop* aim to distinguish achievement verbs that do not normally come after the verbs *finish* and *stop* because they do not describe any kind of process. The achievement verbs can occur with *stop* if they only express a habitual, repetitive event, e.g. *The visitors stopped coming to the museum* (every day). In the experiment, examples are regularly given with both *finish* and *stop* with the **approach** verbs:

The hikers finished drawing close to the summit.

The boat stopped coming near the shore.

To summarise, the results of the experiment show that the verbs in the **approach** group can be defined either as activities or accomplishments, depending on the context in which they are used, while the verbs in the **arrive** group can be defined as achievements.

Different examples, even if they express the same meaning of a verb, may refer to different situation types, so the linguistic tests may or may not work, mainly because they were not constructed for LLMs. While an expert may be able to make a relatively quick decision about a particular example, the task of classifying verbs based on multiple constructions in which they may appear with different meanings should be further refined.

For the reasons mentioned above and because of the nature of the prompt, this experiment cannot be used to evaluate the models. However, the results may lead the expert to make the correct decision based on the examples and interpretations provided by the LLMs. The fact that one and the same semantic frame can be evoked by both activity/accomplishment verbs such as ***approach*** and achievement verbs such as ***arrive*** could raise the question of the reorganisation of some frames (the general meaning of the frame **Arriving** is that of achievement). Such an indication could arise from the fact that other verbs that are among the selected potential candidates for the frame **Arriving** are actually part of other frames, e.g. the verb *land* evokes the frame **Vehicle_landing** with the definition: "A flying **Vehicle** comes to the ground at a **Goal** in a controlled fashion, typically (but not necessarily) operated by an operator".

5 Conclusions

In general, LLMs cope quite successfully with the linguistic annotation of words or phrases expressing a given FrameNet core frame element, so it can be expected that the automatic annotation of the (core) frame elements for previously selected examples can be successfully performed by LLMs. Furthermore, the use of LLMs as a bank of examples illustrating different linguistic phenomena can be perfected with more specific instructions in the prompts.

Some difficulties in using LLMs for linguistic classification into aspectual classes may arise from the nature of linguistic tests that only work for a particular linguistic context, while there may even be a shift between aspectual classes within a verb meaning for different constructions and contexts. This means not only that linguistic tests need to be further elaborated in order to work with LLMs, but also that a theoretical justification needs to be provided for the contextual conditions under which aspectual class shifts occur, as has already been done for a number of cases in English.

The use of more than one LLM with the same prompts allows the expert to confirm or reject an initial hypothesis, i.e. a second and third annotator may be superfluous. On the other hand, carefully analysing the errors of the LLMs can also help the expert to make one or the other decision.

To summarise, it can be said that the use and importance of LLMs in linguistic work, as in many other areas, will increase. In any case, LLMs are useful to linguists as interlocutors who can surprise them with unexpected linguistic usages.

Acknowledgments

The present study is carried out within the project *Ontology of Activity Predicates – Linguistic Modelling with a Focus on Bulgarian* funded by the Bulgarian National Science Fund, Grant Agreement No KP–06–H80/9 from 8.12.2023.

References

Anthropic. 2023. The Claude 3 Model Family: Opus, Sonnet, Haikus.

Siddhant Arora, Ankita Pasad, Chung-Ming Chien, Jionghao Han, Roshan Sharma, Jee weon Jung, Hira Dhamyal, William Chen, Suwon Shon, Hung yi Lee, Karen Livescu, and Shinji Watanabe. 2024. On the Evaluation of Speech Foundation Models for Spoken Language Understanding. *ArXiv*.

Gašper Beguš, Maksymilian Dabkowski, and Ryan Rhodes. 2023. Large Linguistic Models: Analyzing theoretical linguistic abilities of LLMs. *ArXiv*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yupeng Cao, Zhi Chen, Qingyun Pei, Fabrizio Dimino, Lorenzo Ausiello, Prashant Kumar, K. P. Subbalakshmi, and Papa Momar Ndiaye. 2024. RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data. *ArXiv*.

Yu-Chu Chang, Xu Wang, Jindong Wang, Yuanyi Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Weirong Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qian Yang, and Xingxu Xie. 2023. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15:1 – 45.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

David R. Dowty. 1997. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Studies in Linguistics and Philosophy. Springer Dordrecht.

Hana Filip. 2011. Lexical aspect. In Robert Binnick, editor, *The Oxford Handbook of Tense and Aspect*, pages 721–752. Oxford University Press, Oxford.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280:1, pages 20–32. New York.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the morning calm*, page 111–137. Hanshin Publishing, Seoul.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Gemini Team, Google. 2023. A Family of Highly Capable Multimodal Models.

Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23. ACM.

Anthony Kenny. 2003. *Action, Emotion and Will*, 2nd edition. Routledge.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment Analysis in the Age of Generative AI. *Customer Needs and Solutions*, 11:1–19.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. *ArXiv*.

Xiaofeng Liao and Zhiming Zhao. 2019. Unsupervised Approaches for Textual Semantic Annotation, A Survey. *ACM Computing Surveys*, 52(4):66:1–66:45.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9).

Kyle Mahowald. 2023. A Discerning Several Thousand Judgments: GPT-3 Rates the Article + Adjective + Numeral + Noun Construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

George Miller, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large Language Models: A Survey. *ArXiv*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*.

Letian Peng, Zilong Wang, Feng Yao, Zihan Wang, and Jingbo Shang. 2024. MetaIE: Distilling a Meta Model from LLM for All Kinds of Information Extraction Tasks. *ArXiv*.

Christopher Piñón. 1997. Achievements in an event semantics. *Semantics and Linguistic Theory*, 7:276–293.

Susan Rothstein. 2004. *Structuring Events: A Study in the Semantics of Lexical Aspect*. Blackwell, Oxford.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: combining framenet, verbnet and wordnet for robust semantic parsing. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'05, page 100–111, Berlin, Heidelberg. Springer-Verlag.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text Classification via Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Tiago Timponi Torrent, Thomas Hoffmann, Arthur Lorenzi Almeida, and Mark Turner. 2024. *Copilots for Linguists: AI, Constructions, and Frames*. Elements in Construction Grammar. Cambridge University Press.

Satoru Uchida. 2024. Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics*, 4(1).

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66:143–160.

Henk J. Verkuyl. 1972. *On the Compositional Nature of the Aspects*. Foundations of Language Supplementary Series 15. D. Reidel Publishing Company, Dordrecht, Netherlands.

Henk J. Verkuyl. 1993. *A Theory of Aspectuality: the Interaction between Temporal and Atemporal Structure*. Cambridge Studies in Linguistics; 64. University Press, Cambridge.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. *ArXiv*.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-Trained Language Models and Their Applications. *Engineering*, 25:51 – 65.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023. Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey. *ArXiv*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-Shot Information Extraction via Chatting with ChatGPT. *ArXiv*.

Zheng Xin Yong and Tiago Timponi Torrent. 2020. Semi-supervised Deep Embedded Clustering with Anomaly Detection for Semantic Frame Induction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3509–3519, Marseille, France. European Language Resources Association.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of AI-assisted pragmatic annotation: The case of apologies. *ArXiv*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *ArXiv*.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 2023. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *ArXiv*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.