Using LLMs for feature extraction: the case of language functions with deixis

Thomas Gaillat
Université Rennes 2
Place du Recteur Henri Le Moal
35043 Rennes Cedex, France
thomas.gaillat
@univ-rennes2.fr

Xiaobin Chen
University of Tübingen
Walter-Simon-Str. 12
72072 Tübingen, Germany
xiaobin.chen
@uni-tuebingen.de

Nelly Sagirov
University of Tübingen
Walter-Simon-Str. 12
72072 Tübingen, Germany
nelly.sagirov
@uni-tuebingen.de

Abstract

Language functions and notions have been central to language education since Ek and Trim's seminal work on proficiency thresholds. Their approach maps explicit language functions to potential forms, enabling teachers to design curricula aligned with communicative and task-based teaching objectives. While the English Grammar Profile attempted to automate form-to-function mapping, its deterministic heuristic approach risked overlooking alternative forms and misassigning functions. This study explores the potential of Large Language Models (LLMs) for annotating language functions in L2 English productions, focusing on deictic functions as a preliminary case study.

Our methodology involved creating a manually annotated gold standard using 15 randomly selected texts from the CELVA.sp learner corpus, containing 1,346 deictic form instances. Following Ek and Trim's categorization, we annotated three types: indefinite, definite non-anaphoric (exophoric), and definite anaphoric forms. Inter-annotator agreement showed high value (F1 scores between 0.983 and 0.993). We evaluated GPT-5.5 using few-shot prompting with Ek and Trim's specifications. Initial results showed 57.7% accuracy with macro F1 of 0.620, revealing systematic confusion between exophoric and anaphoric uses, particularly with first-person pronouns. After refining the prompt to explicitly address first-mention pronouns and bridging references, accuracy improved to 86.9% with macro F1 of 0.786. These findings suggest that LLMs can effectively identify language functions when provided with precise linguistic guidelines, offering a scalable approach to automatic function identification for CALL systems.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

1 Introduction

The identification and classification of language functions have been fundamental to language teaching and learning since the foundational work of Ek and Trim (1998b). Their *Threshold* and *Waystage* frameworks, developed under the auspices of the Council of Europe, established language functions as the backbone of communicative language teaching. These functions—such as imparting and seeking factual information, expressing attitudes, suasion, socializing, structuring discourse, and communication repair—provide a systematic way to map communicative intentions to linguistic forms, thereby enabling curriculum designers to create task-based learning objectives aligned with proficiency levels.

Traditionally, the mapping from functions to forms has been performed manually by language teachers and pedagogical material developers. More recently, the English Grammar Profile (EGP) (O'Keeffe and Mark, 2017) attempted to automate this process by using deterministic rules to classify forms into syntactic functions and uses. However, this approach has limitations since, by prioritizing form-to-function mapping, it risks overlooking alternative forms that realize the same function and may incorrectly assign forms to unintended functions, particularly in highly semantic or pragmatic contexts (Bannò et al., 2026; Sagirov and Chen, 2026).

The advent of Large Language Models (LLMs) presents a new opportunity to address these challenges. LLMs have demonstrated strong performance in semantic understanding and discourse annotation tasks (Eguchi and Kyle, 2024). Crucially, they provide access not only to the most likely form for a given function but also to the paradigmatic choices—the alternative forms that a learner might use to realize that function, along with their probabilities. This capability is par-

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 119–129, November 19–20, 2026, University of Tartu Library

ticularly valuable for investigating L2 interlanguage, where learners often exhibit creative and non-native-like alternation patterns (Mallart et al., 2025; Dubois et al., 2023).

In this study, we investigate whether LLMs can effectively identify language functions in L2 English productions, focusing on deictic functions as a preliminary use case. Deixis, defined as "any form referring or identifying by means of linguistic items belonging to closed sets the reference of which is dependent upon the context of the utterance" (Ek and Trim, 1998b, p. 57), presents a significant challenge due to its heavy reliance on semantic and pragmatic context. Our research question is: *Can LLMs perform accurate form–function mappings for deictic reference in learner corpora?*

The contributions of this work are threefold:

- We create a manually annotated gold standard for deictic functions in L2 English using the CELVA.sp learner corpus (Mallart et al., 2023).
- We evaluate the performance of GPT-5.5 in identifying deictic functions using few-shot prompting with explicit linguistic guidelines.
- We demonstrate that prompt refinement, particularly the explicit distinction between anaphoric and exophoric uses, significantly improves LLM performance.

The remainder of this paper is structured as follows. Section 2 presents the theoretical background on form–function mapping and the motivation for using LLMs. Section 3 describes our methodology, including corpus selection, annotation scheme, and LLM prompting strategy. Section 4 presents the results, while Section 5 discusses the findings and their implications. Section 6 concludes with perspectives for future work.

2 Theoretical Background

2.1 Form–Function mapping in language acquisition

The relationship between linguistic forms and their communicative functions has been a central concern in both first and second language acquisition research. As Ellis (1994, p. 135) observes, "Form-function variation is also systematic and is contextually induced. All three types of context—linguistic, situational, and psycholinguistic—may influence which forms learners employ to perform specific language functions." This systematic variability gives rise to L2-specific microsystems of alternations, where learners may use non-native-like distributions of forms to realize particular functions (Mallart et al., 2025; Dubois et al., 2023).

Ek and Trim's work 1998b) represents a seminal effort to catalog language functions and map them to potential forms. As the authors explain, language functions are intimately linked to the needs humans have when expressing. for the authors, "the starting-point of the specification of our objective, then, is a list of the kind of things people may do by means of language. These are things such as describing, enquiring, denying, thanking, apologising, expressing feelings, etc.[...] We refer to these things as 'language functions' (Ek and Trim, 1998b, p 23).

Their final classification includes six main categories: imparting and seeking factual information; expressing and finding out attitudes; suasion; socializing; structuring discourse; and communication repair. Each of these categories is further subcategorized, including specific functions and notions with examples of forms. For instance one specification focuses on anaphora and indicates that it can be realized by "personal pronouns (subject forms and object forms) possessive adjectives possessive pronouns as complement (as in: 'You .take it, it's 'yours.)" (Ek and Trim, 1998b, p.57).

This resulted in a comprehensive inventory that has informed national language teaching programs across Europe, including the French secondary school English curriculum in the 1990s and 2000s.

2.2 Limitations of deterministic approaches

The English Grammar Profile (O'Keeffe and Mark, 2017) attempted to automate the mapping from forms to functions using deterministic rules. While this approach worked well for identifying specific constructions and uses in large corpora, it struggled with highly semantic or pragmatic contexts. As (Bannò et al., 2026) and (Sagirov and Chen, 2026) demonstrate, deterministic rules can underperform compared to more sophisticated systems.

A key limitation of the EGP approach is its directionality. It starts from forms and attempts to map them to functions, which predetermines the functional assignments and may overlook alter-

native forms that realize the same function. For the purposes of language teaching and assessment, however, we often need to work in the opposite direction: starting from a communicative function and identifying the range of forms that learners use to realize it.

2.3 The potential of Large Language Models

Large Language Models offer several advantages for language function identification. First, they excel at tasks requiring deep semantic comprehension, which is crucial for identifying functions that depend heavily on context, such as deictic reference. Second, unlike deterministic systems that output a single classification, LLMs can provide the probability distribution over possible forms for a given function, giving insight into learners' paradigmatic choices. Third, once an effective prompting strategy is developed, it can be easily scaled to identify a wide range of language functions across large corpora. Fourth, LLMs can be fine-tuned on domain-specific data to improve performance for particular tasks or languages.

Previous work has demonstrated the effectiveness of LLMs for discourse-functional feature annotation (Eguchi and Kyle, 2024) and for identifying EGP grammatical constructions (Bannò et al., 2026). However, the application of LLMs to the identification of language functions *per se*—particularly in L2 productions—remains underexplored.

2.4 Deixis as a use case

We selected deictic functions as our preliminary use case for several reasons. Deixis has been extensively studied and hinges around the notions of external reference (exophora) vs. contextual reference (endophora) (Halliday and Hasan, 1976; Stirling and Huddleston, 2002). In addition, some linguists also see an opposition between new vs. already mentioned referents (Cornish, 1999). Syntactically, deixis is realized via various forms such as *this*, and *that* proforms, but it is also realized pragmatically as the forms enable the speakers to take difference stances regarding the focus they want to place in particular situations.

Ek and Trim's definition frames deixis as a notion which "involves referring or identifying by means of linguistic items belonging to closed sets the reference of which is dependent upon the context of the utterance, e.g., time, place, persons involved." For them, the notion "may be definite or indefinite, non-anaphoric or anaphoric, i.e. referring to an item already mentioned". Their definition implies several points. First, it involves reference that is dependent on the context of utterance, including time, place, and the persons involved, which makes it a challenging test of semantic understanding. Second, they provide explicit definitions and examples for deictic forms, which can be operationalized for annotation. Third, deictic reference is fundamental to communicative competence, and errors in deictic usage can significantly impact comprehension. Fourth, the distinction between anaphoric (referring to a previously mentioned entity) and exophoric (referring to entities in the situational context) uses provides a clear test of an LLM's ability to understand contextual reference.

Regarding L2 acquisition, the deictic construct is actually not so easy to acquire, and it is interesting to study it to understand language development. For instance, in the case of proforms *this, that* and *it* proforms, it has been shown that learners get confused between each of the forms in their proform use (Gaillat, 2016). Pointing out entities in discourse appears to raise confusions in the cues that trigger a proform instead of the two others. Deixis represents a challenge for learners that needs to be addressed in instruction, be it in class or in CALL systems.

3 Methodology

3.1 Corpus and data selection

Our study uses the CELVA.sp learner corpus (Mallart et al., 2023), a collection of over 1,700 texts written by university students across various academic domains. For our gold standard, we randomly selected 15 texts from this corpus, resulting in a dataset of 293 unique sentences containing 1,346 deictic form instances. The texts were distributed across four CEFR levels (A1-B2), as shown in Table 1. The data is available on the IRIS database[1].

The CELVA.Sp corpus is particularly suitable for this study because it contains authentic L2 English productions from learners at various proficiency levels, it covers a range of academic domains providing diversity in topics and vocabulary, and it has been previously used in studies of syntactic alternations and proficiency assessment (Mallart et al., 2025).

[1] see https://www.iris-database.org/details/2qKsR-xul7i/

CEFR level	Number of texts
A1	4
A2	3
B1	5
B2	3

Table 1: Distribution of texts across CEFR levels.

3.2 Annotation scheme

Following (Ek and Trim, 1998a), we adopt a three-way classification of deictic forms, operationalizing their definitions into three categories:

- **Indefinite:** Forms like *someone*, *something*, *anywhere* that do not refer to specific entities. Example: *I need a pen from **somewhere** in the room.*
- **Definite non-anaphoric (exophoric):** Forms like *this*, *that*, *here*, *now* that refer to entities in the situational context or to new entities introduced without prior mention. Example: ***This** is heavy.* (speaker pointing to an object)
- **Definite anaphoric:** Forms like *he*, *she*, *it*, *the* that refer to entities already mentioned in the discourse. Example: *The keys are on the table. **Those** are mine.*

We used the corresponding tagset, including some examples.

1. **DEIXIS-INDEF:** Indefinite deictic forms
2. **DEIXIS-DEF-NON-ANAPHORA:** Definite non-anaphoric (exophoric) forms
3. **DEIXIS-DEF-ANAPHORA:** Definite anaphoric forms

3.3 Annotation process

Annotation was performed using Argilla (Daniel and Francisco, 2023), an open-source framework for data-centric NLP that supports collaborative span annotation (See Figure 1). Argilla was chosen for its intuitive interface, which allows non-computer specialists to participate in annotation projects, and its support for Docker-based deployment, enabling easy setup on local servers. Each targeted span can be selected with different colors depending on the label. The data can be extracted via a Python Argilla client that allows dataset creation for subsequent machine learning tasks.

Two of the annotators, both experts linguists, participated in the annotation process. First, guidelines were developed on the basis of the chapter dedicated to language functions (Ek and Trim, 1998a, p 27-47). these include definitions of anaphora and exophora regarding which our modeling approach encounters challenges (see Section 4.2) Second, both annotators discussed the guidelines and annotated a small sample of texts together to ensure mutual understanding. Third, each annotator separately annotated the same 15 texts. Finally, disagreements were discussed and resolved through consensus. Most disagreements were attributed to fatigue from processing large numbers of annotations rather than fundamental differences in interpretation.

3.4 Inter-rater agreement

Inter-rater agreement was calculated before the final discussion using standard classification metrics. As shown in Table 2, agreement was exceptionally high across all categories, with F1 scores ranging from 0.983 to 0.993. This high level of agreement indicates that the annotation scheme was clear and that the guidelines were effectively communicated.

3.5 LLM prompting strategy

We evaluated `gpt-5.5-pro-2026-04-23` using a few-shot prompting approach. Two modes were tested: a reasoning mode and a NULL mode (without reasoning). The results reported in this paper are from the reasoning models, which had better performance than the non-reasoning model. The prompt structure was designed to provide the LLM with comprehensive linguistic context:

- **Mission statement:** "You are an agent that does linguistic annotation of deixis."
- **Linguistic guidelines:** i) A detailed explanation of deixis based on (Ek and Trim, 1998a), including the definition "Deixis involves referring or identifying by means of linguistic items belonging to closed sets the reference of which is dependent upon the context of the utterance (e.g., time, place, persons involved)", ii) the categorization (definite vs. indefinite, anaphoric vs. non-anaphoric), iii) a candidate word list of deictic forms for each category, and contextual examples illustrating each category.

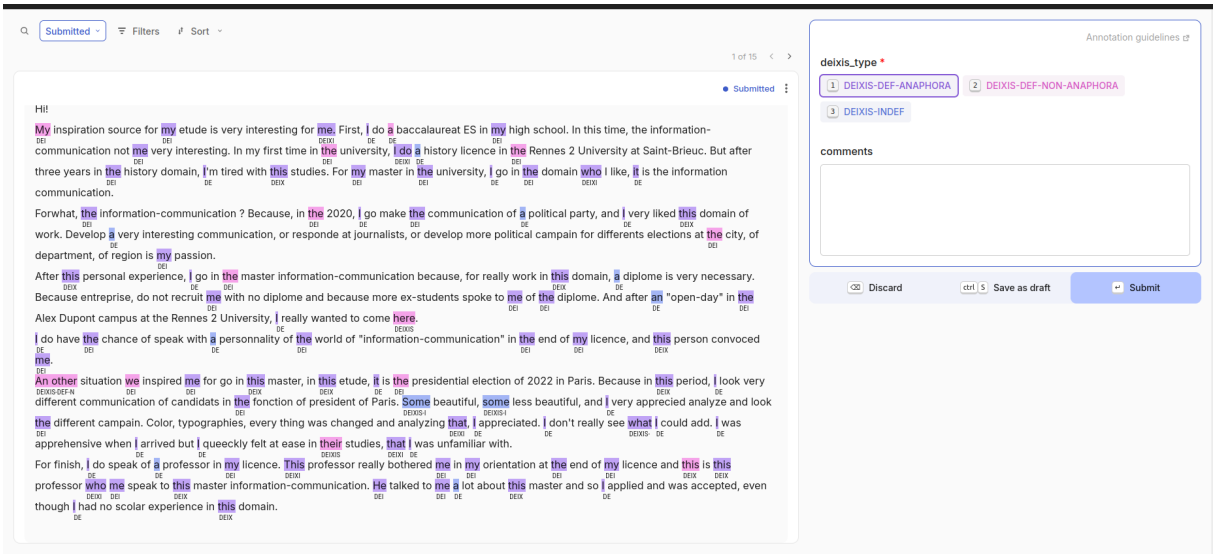


Figure 1: Argilla annotation tool

Label	Precision	Recall	F1	# of forms
DEF-ANAPHORA	0.9919	0.9950	0.9935	613
DEF-NON-ANAPHORA	0.9783	0.9890	0.9836	91
INDEF	0.9742	0.9930	0.9835	152
TOTAL				856

Table 2: Inter-rater agreement for deictic function annotation.

- **Task description:** "I will give you the context of the sentence as well as the text in which the sentence comes from. Use them as the context to decide on the label of the deixis. Only output the label, nothing else."
- **Target presentation:** The sentence with the deictic form marked in double brackets (e.g., "Dear college. Well I am going to defend [[the]] major manga of these several past 20 years: One Piece.").
- **Full context:** The entire learner text in which the sentence appears, introduced by "The text in which the sentence is taken from:"

This comprehensive prompt structure was designed to mimic the information available to a human annotator, providing both the immediate sentence context and the broader discourse context. The evaluation was done on OpenAI Platform's Evaluation environment, with the following settings:

- Reasoning mode: pro
- Reasoing effort: medium
- Verbosity: medium
- Summary: auto

3.6 Evaluation metrics

We evaluated the LLM's performance using standard classification metrics: accuracy, precision, recall, F1 score, and macro averages across all classes, treating each class equally regardless of size.

It must be noted that random classification would be expected to reach 33% accuracy, establishing a baseline against which to interpret results.

4 Results

4.1 Initial prompt performance

The initial evaluation using the first version of the prompt yielded the results shown in Table 3. Overall accuracy was 0.577, with a macro F1 score of 0.620. Performance varied significantly across categories: the model performed extremely well on indefinite deictic forms (F1 = 0.962) but struggled with the distinction between definite anaphoric and definite non-anaphoric forms. The

lowest performance was on definite non-anaphoric forms (F1 = 0.307), with high recall (0.837) but very low precision (0.188), indicating that the model was over-predicting this category.

4.2 Error analysis

To understand the sources of error, we examined the confusion matrix (Table 4) and conducted a qualitative analysis of the misclassifications. Claude AI was used to compared all occurrences of spans between the GS and the predictions.

The confusion matrix reveals that the primary source of error was the misclassification of definite anaphoric forms as definite non-anaphoric. Out of 591 actual DEF-ANAPH instances, 333 were incorrectly classified as DEF-NON-ANAPH.

Qualitative analysis revealed two main issues. First, the model frequently misclassified repeated first-person pronouns (*I*, *me*, *my*, *we*, *us*, *our*) as exophoric rather than anaphoric. In reality, only the very first reference to the writer in a text should be considered exophoric; all subsequent references are anaphoric, as they refer back to the previously introduced speaker. Second, the model struggled with bridging references, where a definite noun phrase refers to a part or aspect of a previously mentioned entity (e.g., *the background* after introducing a painting). These should be classified as anaphoric, but the initial prompt did not provide clear guidance on this point.

4.3 Prompt refinement

Based on the error analysis, we refined the prompt to address these specific issues. The modified prompt included four explicit rules.

i) A general rule for anaphoricity states that "a marked form is *def_anaph* whenever its referent has been mentioned anywhere earlier in the text, in any grammatical form—and the model should scan the whole preceding text for an antecedent before choosing between *def_anaph* and *def_non_anaph*."

ii) For first- and second-person forms (*I*, *me*, *my*, *we*, *you*, ...), "only the very first reference to the writer (or reader) in the text is *def_non_anaph*; every later one is *def_anaph*."

iii) Regarding bridging, or associative anaphora, with article *the*, "*the* is anaphoric not only on repeated mention but also when its noun is a part or aspect of a previously mentioned entity (e.g., *the background*, *the dominant color* after a painting has been introduced)."

iv) Finally, *def_non_anaph* is specified as residual, "reserved for referents new to the text, identified exophorically from the situation or from material inside the phrase itself."

The final prompt for the experiment can be found in the Appendix 6.

4.4 Improved performance

The refined prompt led to substantial improvements in performance, as shown in Table 5 and the confusion matrix in Table 6.

The amended prompt improved overall accuracy by 28.4 percentage points (from 0.585 to 0.869) and macro F1 by 16.6 points (from 0.620 to 0.786). The most dramatic improvement was for DEF-ANAPH, where F1 increased from 0.595 to 0.910. DEF-NON-ANAPH also saw substantial improvement, with F1 increasing from 0.349 to 0.508.

While the amended prompt significantly improved performance, some confusion between DEF-NON-ANAPH and DEF-ANAPH remains. This is partly due to a residual issue in the prompt, i.e., the phrase "identified exophorically from the situation or from material inside the phrase itself" can lead to misclassification of postmodified definites (e.g., *the way the painting was created* should be DEF-ANAPH if the painting was previously mentioned, but the wording of the prompt might suggest DEF-NON-ANAPH).

5 Discussion

5.1 Key findings

The results of this study demonstrate that Large Language Models can effectively identify language functions in L2 English productions when provided with clear and explicit linguistic guidelines. Our gold standard achieved exceptional inter-annotator agreement (F1 scores of 0.983–0.993), indicating that the annotation scheme based on (Ek and Trim, 1998b) is clear and reliable. GPT-5.5 achieved 57.7% accuracy with the initial prompt, demonstrating a baseline capability for deictic function classification. Explicit linguistic rules significantly improved performance. The amended prompt, which included clearer distinctions between anaphoric and exophoric uses, achieved 86.9% accuracy and a macro F1 of 0.786. This improvement of 28.4 percentage points underscores the importance of precise task instructions for LLMs. Finally, the model performed best

Class	Precision	Recall	F1
INDEF	0.927	1.000	0.962
DEF-NON-ANAPH	0.188	0.837	0.307
DEF-ANAPH	0.952	0.429	0.592
Accuracy 0.577; Macro P 0.689; Macro R 0.755; Macro F1 0.620			

Table 3: Performance of the initial prompt for deictic form classification.

Gold\Pred	INDEF	NON-ANAPH	ANAPH
INDEF	153	0	0
DEF-NON-ANAPH	2	77	13
DEF-ANAPH	10	333	258

Table 4: Confusion matrix for the initial prompt.

Metric	Initial	Amended
Overall accuracy	0.585	0.869
Macro F1	0.620	0.786
INDEF F1	0.915	0.941
DEF-ANAPH F1	0.595	0.910
DEF-NON-ANAPH F1	0.349	0.508

Table 5: Performance comparison between the initial and amended prompts.

on indefinite deictic forms (F1 between 0.941 and 0.962), while the primary source of error was the confusion between exophoric and anaphoric uses, particularly for first-person pronouns and bridging references.

5.2 Comparison with previous work

Our findings are consistent with previous studies that have applied LLMs to discourse-functional annotation. Eguchi and Kyle (2024) demonstrated that LLMs can effectively annotate discourse-functional features in L2 writing, though their focus was on discourse-level rather than semantic-pragmatic functions. Bannò et al. (2026) showed that LLMs can identify EGP grammatical constructions, but also noted that deterministic systems like POLKE may outperform LLMs for purely syntactic tasks (Sagirov and Chen, 2026). Nevertheless deixis span extraction being highly semantic, an approach relying strictly on form matching is likely to fail to identify lesser-used forms operationalizing a language function.

Our study extends the discourse functional direction by focusing on semantic-pragmatic functions (deixis), but it is not a direct match to (Eguchi and Kyle, 2024). While their approach focuses on identifying rhetorical stance expressions in academic English, we conduct LF identification based on Ek and Trim's publication (Ek and Trim, 1998b). We show that LLMs can achieve high accuracy when provided with explicit linguistic guidelines. The improvement from prompt refinement suggests that LLMs are particularly sensitive to the clarity and specificity of task instructions for semantic tasks.

5.3 Implications for language learning and automatic assessment

First, it must be said that our approach to LF extraction is designed to be scaled up, as more prompts can be added in the architecture to accommodate other language functions (as skills.md file for instance).

The ability to automatically identify many language functions in learner corpora has several important implications. Language functions can serve as new metrics for proficiency prediction. Unlike traditional complexity metrics (Bulté and Housen, 2012), function-based metrics can provide insights into learners' communicative competence and their ability to use language appropriately in context. The automatic identification of language functions could enable adaptive learning systems that adjust tasks based on the detected functions and forms in learners' productions, which is currently missing in most CALL systems. Existing systems mostly focus on providing learners with interactive activities and feedback on grammar errors, instead of modeling what the learner can or can not do at a functional level. CALL systems embarking automated function identification could include proficiency pre-

Gold\Pred	INDEF	NON-ANAPH	ANAPH
INDEF	245	2	8
DEF-NON-ANAPH	6	79	88
DEF-ANAPH	15	57	846

Table 6: Confusion matrix for the amended prompt.

diction models to support teachers in diagnosing learners' strengths and weaknesses in specific functional areas, designing targeted instructional activities, and tracking learners' progress in acquiring specific language functions. A functional approach to L2 learning also aligns better with the Common European Framework of References for Languages (Council of Europe, 2001, CEFR). Finally, researchers can use automated function identification to analyze large learner corpora, investigating questions such as how learners' functional repertoires develop across proficiency levels, what the typical alternation patterns for specific functions in L2 interlanguage are, and how functional distributions differ between L1 and L2 productions.

5.4 Limitations

Despite the promising results, this study has several limitations. Our gold standard consisted of only 15 texts and 1,346 deictic instances; while this was sufficient for a preliminary study, larger datasets would be needed for fine-tuning and for generalizing to broader populations. Our pilot texts are from learners similar in proficiency and studying in the same university. It is still unclear if the LLMs would perform at the same level when texts by learners of other proficiency levels are analyzed. Furthermore, We focused on a single language function (deixis) with three subcategories; extending the approach to the full inventory of language functions (Ek and Trim, 1998b) would require additional annotation and prompt development.

Our results are specific to GPT-5.5, and performance may vary with other LLMs or versions. The significant improvement from prompt refinement highlights the sensitivity of LLMs to task instructions, and developing effective prompts for other language functions may require similar iterative refinement. Finally, even with the amended prompt, some confusion between DEF-NON-ANAPH and DEF-ANAPH remained, indicating that further refinement or fine-tuning may be necessary for optimal performance.

6 Conclusion and future work

This study demonstrates that Large Language Models can effectively identify language functions in L2 English productions when provided with clear and explicit linguistic guidelines. Our focus on deictic functions as a preliminary case study yielded promising results: with an initial prompt, GPT-5.5 achieved 57.7% accuracy, and with a refined prompt that explicitly addressed anaphoricity, accuracy improved to 86.9%. The key contribution of this work is the demonstration that LLMs can be used for automatic language function identification, a task that has traditionally required manual annotation. This capability opens up new possibilities for language teaching, assessment, and research, including the development of adaptive learning systems, teacher support tools, and large-scale analyses of learner corpora.

Future work will focus on several directions. We plan to increase the size of our annotated dataset to enable fine-tuning of LLMs for deictic function identification, which will also allow us to investigate the performance of fine-tuned models compared to prompt-based approaches. We will extend our approach to other language functions from the inventory (Ek and Trim, 1998b), such as those related to expressing attitudes, suasion, and socializing. We will explore methods for automatically converting function definitions and examples into effective prompts for LLMs, reducing the need for manual prompt engineering. Finally, we will develop prototypes of adaptive learning systems and teacher support tools that leverage automated function identification to provide targeted feedback and instruction.

Ultimately, this line of work aims to create a comprehensive map of language functions and their realized forms in learner productions. Such a map could serve as the foundation for a new generation of language learning technologies that are more closely aligned with communicative competence and the needs of real-world language use.

Appendix: Complete prompt for deixis function annotation

Deixis involves referring or identifying by means of linguistic items belonging to closed sets the reference of which is dependent upon the context of the utterance (e.g. time, place, persons involved). Deixis may be definite or indefinite (he vs. someone), non-anaphoric or anaphoric (i.e. referring to an item already mentioned) (Why don't *you* come? vs. buy those books because I need *them*.). Here is a non-exhaustive list of potential candidates classified per category:

A. Non-Anaphoric Deixis (`def_non_anaph`)

1. Personal Pronouns
    - Subject forms: I, you, he, she, it, we, they
    - Object forms: me, you, him, her, it, us, them
2. Possessive Forms
    - Possessive adjectives: my, your, his, her, its, our, their
    - Possessive pronouns as complement: mine, yours, his, hers, ours, theirs
        - Example: This is mine.
    - Possessive pronouns as subject:
        - Example: Mine is better.
3. Demonstrative Forms
    - Demonstrative adjectives: this, that, these, those, such
    - Demonstrative pronouns: this, that, these, those
4. Independent Relative Pronoun
    - what
        - Example: What you say is true.
5. Definite Article
    - the
6. Interrogative Forms
    - Interrogative pronouns: who, whom, whose, what, which
    - Interrogative adjectives: whose, what, which

B. Anaphoric Definite Deixis (`def_anaph`)

1. Personal Pronouns
    - Subject forms: I, you, he, she, it, we, they
    - Object forms: me, you, him, her, it, us, them
2. Possessive Forms
    - Possessive adjectives: my, your, his, her, its, our, their
    - Possessive pronouns as complement:
        - Example: You take it, it's yours.
3. Demonstrative Forms
    - Demonstrative adjectives: this, that, these, those
    - Demonstrative pronouns: this, that, these, those
4. Relative Pronouns
    - who, whom, whose, which, that
    - Omission of relative pronoun:
        - Example: The book (that) I read was interesting.
5. Reflexive/Emphatic Pronouns
    - myself, yourself, himself, herself, itself, ourselves, yourselves, themselves
        - Example: I hurt myself.
        - Example: I've done it myself.
6. Reciprocal Pronoun
    - each other
7. Definite Article
    - the
8. Adverbs
    - here, there, now, then, so
        - Example: He wanted to go out, but he didn't say so.
9. Prop Word
    - one, ones
        - Example: I like the red one.
10. Substitute Verb
    - do
        - Example: He asked me to help him, and I did.

C. Indefinite Deixis (`indef`)

1. Indefinite Article
   - a, an
2. Indefinite Pronouns
   - People/Entities: someone, somebody, no one, (not) ... anybody, (not) ... anyone, nobody, each, everybody, everyone
   - Things/Events: something, (not) ... anything, nothing, everything
   - Quantitative:
     - all (e.g., They all went home; I want all of it.)
     - both (e.g., They both went home; I want both of them.)
     - some (e.g., Some of them went home.)
   - Impersonal:
     - it (e.g., It's raining.)
     - you (e.g., It's a nice record if you like modern music.)
3. Indefinite Determiners (Quantity)
   - all, a lot of, any (also: hardly any, not any), both, each, enough, (a) few, (a) little, many, more, most, much, no, several, some
   - half (e.g., Give me half of it; Give me the other half; Give me half a bottle.)
   - at least (e.g., I need at least $5.)
4. Adverbs
   - somewhere, nowhere, everywhere, (not) ... anywhere, sometimes, never, always
5. Semi-Deictics
   - Person/Entity:
     - person (e.g., There are five persons present.)
     - man (e.g., There were animals here before man came.)
     - people (e.g., What do people think about the government?; There are five people present.)
   - Thing/Action:
     - thing (e.g., What do you call that thing?)
     - do (e.g., What are you going to do tonight?)

Annotation Task

You will be given a sentence that contains deixis forms. The deixis form is marked with bracket pairs such as [[word]]. Decide the type of deixis of the marked word using one of the following labels:

indef: referring to indefinite deixis, e.g.

- Then with Internet witch arrived before the Web , [[a]] whole network developped with also economic functions on this market.
- In fact, people remember [[more]] informations and historical event when they play than they don't.
- I will most take attention on video games but it is something we can also find in [[other]] aspects of entertainment.

def_anaph: referring to definite anaphoric deixis, anaphoric referring to a deixis that refers to something that has been mentioned previously. e.g.

- My parents found it a bit dangerous and warned me of [[the]] risks of social networks.

[We] can move freely in all Paris and in Paris, a Paris's region abandonned nowadays.

[My] opinion in the invention on the web is that is allowed at the time to start to communicate more easily, to exchange document.

def_non_anaph: referring to definite non-anaphoric deixis. e.g.

- Then with Internet witch arrived before the Web , a whole network developped with also economic functions on [[this]] market.
- I will most take attention on video games but it is something [[we]] can also find in other aspects of entertainment.

[I] would like to discuss about the amazing way people can remember huge informations throught playing.

Definite non-anaphoric forms may includes forms referring to new referents in the context or to exophoric referents.

There are basically two aspects of the deixis you need to consider, definiteness and whether it is anaphoric or not. Definiteness is about whether the deitic word is referring to something specific or general. Anaphoricness is about whether the word is referring to an entity that has been mentioned before. Therefore, for both aspects, you would need to context in which the word is used to decide its definiteness and anaphoricness. I will give you the context of the sentence as well as the text in which the sentence comes from. Use them as the context to decide on the label of the deixis.

Only output the label, nothing else.

Here is the sentence with the marked word:

`{{highlighted_sentence}}`

Here is the text in which the sentence is taken from:

`{{text}}`

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: Mistral Vibe was used to rephrase and improve language.

References

Stefano Bannò, Penny Karanasou, Kate Knill, and Mark Gales. 2026. Exploiting the English Grammar Profile for L2 grammatical analysis with LLMs. ArXiv:2603.17171 [cs.CL].

Bram Bulté and Alex Housen. 2012. *Defining and Operationalising L2 Complexity*. John Benjamins Publishing Company.

Francis Cornish. 1999. *Anaphora, Discourse, and Understanding. Evidence from English and French*. Oxford University Press, Oxford.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Vila-Suero Daniel and Aranda Francisco. 2023. Argilla - Open-source framework for data-centric NLP.

Tanguy Dubois, Magali Paquot, and Benedikt Szmrecsanyi. 2023. Alternation phenomena and language proficiency: the genitive alternation in the spoken language of EFL learners. *Corpus Linguistics and Linguistic Theory*, 19(3).

Masaki Eguchi and Kristopher Kyle. 2024. Building custom NLP tools to annotate discourse-functional features for second language writing research: A tutorial. *Research Methods in Applied Linguistics*, 3(3):100153.

J. A. van Ek and J. L. M. Trim. 1998a. Language functions. In *Threshold 1990*, Threshold, pages 27–47. Cambridge University Press, Cambridge.

Jan Ate van Ek and John Leslie Melville Trim. 1998b. *Threshold 1990*. Cambridge University Press, Cambridge (England).

Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press, Oxford, United Kingdom.

Thomas Gaillat. 2016. *Reference in Interlanguage: the case of this and that. From linguistic annotation to corpus interoperability*. Thesis, Université Paris-Diderot, Paris.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. English Language Series. Pearson Education Limited, Harlow.

Cyriel Mallart, Andrew Simpkin, Nicolas Ballier, Paula Lissón, Rémi Venant, Bernardo Stearns, Jen-Yu Li, and Thomas Gaillat. 2025. Assessing the validity of syntactic alternations as criterial features of proficiency in L2 writings in English. *Research Methods in Applied Linguistics*, 4(3):100238.

Cyriel Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen Yu Li, and Thomas Gaillat. 2023. A new learner language data set for the study of English for Specific Purposes at university level. In *Proceedings of the 4th Conference on Language, Data and Knowledge - LDK 2023*, volume 1, pages 281–287, Vienna, Austria.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

Nelly Sagirov and Xiaobin Chen. 2026. POLKE: A system for comprehensively annotating pedagogically-oriented grammatical structure use in language production. *Research Methods in Applied Linguistics*, 5(3):100336.

Lesley Stirling and Rodney Huddleston. 2002. Deixis and anaphora. In Rodney Huddleston and Geoffrey K. Pullum, editors, *The Cambridge Grammar of the English Language*, pages 1449–1564. Cambridge University Press, Beccles, Suffolk.