Assessing the Pragmatic Competence of an LLM Regarding Novel Discourse Markers in Digital Communication

Ágnes Abuczki[1], **Giedre Valunaite Oleskevicienė**[2]

[1]Károli Gáspár University of the Reformed Church in Hungary; [2]Faculty of Human and Social Studies of Mykolas Romeris University
[1]Reviczky u. 4. 1088 Budapest, Hungary; [2]Ateities 20, Vilnius, Lithuania
[1]abuczki.agnes@kre.hu, [2]gvalunaite@mruni.eu

Abstract

The English language is changing faster than before, partly due to the influence of the Internet. Digital language includes a large number of discourse markers (DMs), many of which can be considered innovative. Acronymization, pragmatic specialisation, and compensatory lexical innovation are the most common lexical processes that can be witnessed in the DMs used in computer-mediated communication (CMC). The following novel DMs were identified in recent Twitter chats*: lol, tbh, omg, meh*, and *idk*. These DMs perform several functions, such as showing emotions, signaling uncertainty, hesitation, or mitigation. Interpreting these functions may not be an easy or obvious task for AI. The primary aim of the study is to evaluate the pragmatic competence of an LLM, Gemini 3 Pro, regarding the interpretation of these novel DMs. A mixed-method research process was employed: LLM-generated outputs were compared with the findings of the relevant literature, quantitative corpus analysis, and our qualitative human interpretation to assess the model's analytical usefulness. Gemini 3 Pro was found to show a high level of pragmatic competence in terms of interpreting the functions of DMs, but sometimes tended to overgeneralise, or failed to understand the tone of the text and the intention of the speaker to use a DM.

Keywords: discourse markers, LLMs, digital communication

1. Introduction, Research Objectives

Research on lexical innovation has become even more relevant in 21st-century society, particularly in the context of digital development, which is inevitably linked to multiple linguistic changes as well, including lexical innovations. Linguistic innovation is especially observable on social media platforms, whose discourse reaches broad audiences in a digital format that has become a norm in contemporary communication.

As Large Language Models (henceforth: LLMs) become more powerful and capable, it is now necessary to assess them beyond basic knowledge and data recall and focus on their ability to grasp nuance and context. In this pilot study, we focus on recent colloquial lexical innovations spreading on social media with discourse marking functions. In our pilot, we identified the following common discourse markers used in recent Twitter chats*: lol, tbh, omg, meh*, *idk*. These novel discourse markers (henceforth: DMs) are innovative tools that perform complex social tasks, such as showing emotions (such as *omg* showing surprise), signaling uncertainty and hesitation (by using *idk*), or ending a conversation in a text-based, digital environment (e.g. *so yeah*).

The primary aim of the study is to evaluate the pragmatic competence of an LLM (in our case study, Gemini 3 Pro) regarding novel discourse markers in computer-mediated communication (henceforth: CMC). (Gemini 3 Pro was released in November, 2025, described by Google as its most intelligent AI model yet, featuring enhanced reasoning and coding capabilities.) The secondary goal of the research is to compare and contrast the output of the LLM with human analysis, with the relevant findings of previous research on these items, as well as our own qualitative analysis of these items in our corpus.

2. Theoretical Background

2.1. Lexical Innovation

Social media remains the driving tool for versatile communication, reaching a large number of people and covering a wide variety of topics, including political, social, economic, and other information. It is also important to pay attention to the pragmatic side of such communication because of the way these topics are presented in the media, as the society's perception of social media discourse also depends on ideology, public acceptance, established stereotypes, accepted morality, gender perception, etc. (Pohorila, 2022). The dynamic phenomenon of modern media discourse also carries a powerful pragmatic and evaluative potential, influencing the views of society and forming subjective worldviews. In turn, research related to linguistic phenomena related to media discourse sheds light on linguistic form and corresponding genres guiding the pragmatic functions of social media discourses (Horbatko, 2021).

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 53–59

Lexical innovation has been studied from various perspectives by linguists, categorizing the processes of word formation (Miller 2014) and tracking the change of the meaning of words over time (Geeraerts 2010). In recent studies, researchers have focused on word lexicalization and their gradual acquisition of particular forms and meanings as well as the institutionalization of words as they enter into the standard vocabulary of a language (Brinton & Traugott, 2005). Lexicographical research focuses on identifying and defining neologisms by extensively using the corpus approach and internet search engine results (Kerremans et al. 2011). Traditional linguistic interest in the formation and development of new words is enhanced by corpus linguistics, which offers new solutions, as it allows for the open-ended analysis of language variation and change by searching large amounts of natural language data (Szmrecsanyi 2011; Grieve 2015). It should be admitted that small, regionalised corpora are not extensive enough to observe variations both in the use of common content words and rare new words, but the growth of social media has been changing the situation with the possibilities of computational linguists and LLMs to analyze incredibly large amounts of linguistic data harvested online, especially from Twitter, to understand certain patterns of lexical variation and change (Huang et al. 2016).

2.2. The Category of DMs

There is no widespread agreement in the literature about the name and the definition of this group of lexical items (including *tbh, omg, lol, idk, meh, so yeah*). Concerning the terminology of the present research, we refer to these items as discourse markers (DMs). DMs are traditionally defined as “sequentially dependent elements that bracket units of talk” (Schiffrin 1987: 31) or metalinguistic items that provide information about the segmentation and operation of a discourse (Fraser 1999). Schiffrin (1987) describes the role of DMs as “providing contextual coordinates for ongoing talk” that indicate for the hearer how an utterance is to be interpreted. This is the reason why they prove to be frequent and useful elements in CMC as well, since they help the readers disambiguate the intended meaning and tell us about the mental state/stance of the speaker or writer. Furkó (2014) overviewed earlier DM research and presented the criteria for DM identification by describing the key features of this morphologically, syntactically, and pragmatically heterogeneous group: high oral frequency, optionality (in a syntactic sense), low propositional contribution, procedural meaning, extreme multifunctionality (fulfilling pragmatic/interactional functions, such as stance, alignment, and mitigation), and context dependence. In summary, DMs are multifunctional pragmatic elements expressing various metacommunicative and cognitive functions. These functions of digital DMs will be explored in this research based on a small corpus and its mixed-method analysis, including analysis by an LLM.

2.3. Types of Lexical Innovation in DMs

We can classify the scrutinised DMs in terms of the type of lexical innovation present in them. The first feature is an abbreviation, as due to lexical economy, DMs shift from full phrases to initialisms. Such acronyms also turn into discourse markers with certain pragmatic functions. Another feature is pragmaticalisation, which means that acronyms undergo a lexicalisation process through which they become lexical units; for example, *lol* is used as a verb, e.g., in "*He literally lol'd*". Depending on the theoretical framework, scholars describe this process as pragmaticalisation (Ariel 1998), grammaticalisation (Traugott 1995), or constructionalisation (Traugott & Trousdale 2013).

2.4. Previous Research on the Selected Items in CMC

McCulloch (2019) describes language use on the Internet and argues that the English language is changing faster than before because of the influence of the Internet. Concerning the DMs under scrutiny, he finds that *lol* has become a softener. Using it at the end of a sentence (e.g., "*I'm so tired lol*") signals that the speaker is not actually complaining aggressively but rather seeking sympathetic feedback (e.g., a nod) from the reader.

Scott (2015) specifically analysed the one-to-many, asynchronous communication mode of Twitter and found that tweeters make their intended contextual assumptions accessible to a wide range of readers by using hashtags, which facilitate the use of an informal, casual style that fits the discourse context of Twitter. Scott suggests that expressions such as *tbh (to be honest)* and *ngl (not gonna lie)* serve as mitigators that soften the impact of a statement. These markers are typically used before giving potentially offensive or controversial opinions. Through the explicit reference to being honest, the writer signals a transition from polite conversation to a more authentic personal insight, which builds a sense of closeness with the listener (Scott 2015).

Tagliamonte & Derek (2008) analysed pragmatic particles in instant messaging among teens and found that *lol* and *omg* are used for discourse-pragmatic purposes. For instance, in their understanding, *lol* is used as a marker of empathy or a way to signal that the conversation is friendly. It serves a phatic function; it keeps the social connection open, rather than indicating actual humor. They claimed that teenagers use nonstandard language, but it should not be considered a degradation of language, but a new, innovative, and functional form of communication.

Vandekerckhove (2025) gave a detailed analysis of the use of *omg* in the digital language of Flemish adolescents. In general, he also shares the view that discourse markers function as

pragmatic signals that tell the reader exactly how to interpret the text, since digital writing lacks the nonverbal cues of face-to-face interpersonal communication. The paper highlights that, besides the primary function of *omg* to express shock, it fulfils discourse organizational functions as well. *omg* is often used to mark boundaries and bracket a message. It may signal a shift, e.g., from casual chat to a high-intensity narrative (as in *"omg you won't believe what happened..."*).

Softener markers like *lol* or *idk* are often used to mitigate face-threatening pragmatic acts. They are used when a speaker makes a request or expresses a slight criticism, and adding a marker at the end reduces the social risk of the interaction.

3. Methodology

We followed a mixed-method research process: a combination of quantitative corpus analysis (using a concordance), qualitative human interpretation, and AI-assisted analysis. LLM-generated outputs were compared with the findings of the relevant literature and our qualitative human interpretation to assess the model's analytical usefulness, pragmatic competence, and limitations, in line with recent computational studies on machine-learning methods for detecting hedges (Wise & El Barj 2023, p. 3). First, data was collected; a Twitter chat corpus was selected for this purpose, since it is a genre peculiar to recent digital communication. This text corpus was scraped from Twitter (242,170 words, 51 MB), where the odd lines are tweets and even lines are corresponding responded tweets. The corpus is formatted as a list of independent messages or short exchanges organised into one message per line. The text displays lexical variety, which is typical of social media, and includes a mix of standard English and slang (e.g., "*deadass*" and "*lowkey*"). The text is also highly informal, characteristic of non-standard capitalization, excessive punctuation (e.g., "*!!!!*"), and frequent use of emojis. In the subsequent stage of the experiment, AntConc 4.3.1, a freeware corpus analysis toolkit, was used for concordancing and quantitative text analysis (Anthony, 2024). We first carried out the quantitative and qualitative analysis of the pilot corpus and then prompted Gemini to carry out an LLM-based analysis. Finally, the different results were contrasted with one another.

First of all, we used AntConc 4.3.1, a freeware corpus analysis toolkit for concordancing and quantitative text analysis (Anthony, 2024). Most entries are single tweets or short chat messages, typically ranging from 10 to 25 words per line. Given the informal nature of the text, the TTR is relatively low for the entire corpus due to the repetition of common conversational phrases, though it contains a high number of unique informal variations and misspellings. The case-insensitive concordance searches of the pilot corpus gave the following counts for the target expressions: there are 201 occurrences of *lol* (including variations such as "*LOL", "lolol", "lolz"*), 35 occurrences of *tbh*, 21 *idk*, 15 *omg* and 4 *meh* items in our research corpus.

In the next stage, we discussed and agreed on the qualitative interpretation of the relevant lines (the left and right contexts of the word searches described above) as well as the functions of the digital DMs, driven by earlier works and the actual examples in the Twitter corpus. These findings were manually saved in a shared spreadsheet file for subsequent comparison.

As a next step, Gemini 3 Pro, an LLM-based generative AI chatbot, was used. Gemini 3 Pro was released in November 2025, described by Google as its most intelligent AI model yet, featuring enhanced reasoning and coding capabilities. In our experiment, Gemini was fed the full corpus file and was given the following prompt in thinking mode: "Collect and analyse the expressions *meh*, *lol, tbh, omg, idk* in the text file (Twitter chat corpus), analyse their uses, and classify the pragmatic and discourse functions of these expressions in digital communication." The prompt did not contain examples of classification or explicit category definitions. Subsequently, Gemini provided its answer about the common discourse-pragmatic functions and uses of the selected DMs (*mitigation, intensifying, hedging, expressing emotions, marking stance,* and *replacing facial expressions or gestures*), and it also reflected on usage in terms of the typical positions of the DMs, although it was not explicitly asked to do so. In the end of its reply, it also suggested giving an example for each function, if we need it. Therefore, we prompted in reply to: "give an example sentence for each function from the same text corpus attached." As a result, it gave a list of functions and an example for each, supposedly the most common usages (in the most frequent positions).

Two other LLMs were consulted and were given the same corpus file and prompt, but GPT-5.2 Auto by OpenAI and Claude 4.5 Sonnet did not upload the corpus file and did not perform the task, as the size of the attached file must have been too large. Claude 4.5 specifically highlighted that files larger than 31 mb cannot be uploaded, so these LLMs were not involved in the analysis, which has left the experiment as a single-model pilot study rather than a comprehensive comparative research study.

4. Research Findings

4.1 Human Interpretation: The Discourse-Pragmatic Functions of the Novel DMs in CMC

The utterances (in Table 1) demonstrate the typical functions and different positions of the scrutinised novel DMs: *lol, tbh, omg*, *meh* (expressing various emotions), and *idk* (expressing a hedge and functioning as a mitigator). These examples have been selected

by us from our corpus to illustrate the most common uses of the DMs in digital communication.

Marker	Example 1	Example 2
meh	*"A: Are you excited? B:* ***Meh****, not really."* (indifference)	*"The food was okay, but the service was just...* ***meh****."* (dissatisfaction)
lol	*"I can't believe I just sent that to the wrong person* ***lol****."* (mitigation)	*"****Lol****, that is literally the funniest thing I've seen today."* (irony)
tbh	*"I think the first season was better* ***tbh****."* (stance marking)	*"****Tbh****, I never really understood why that show was popular."* (hedging)
idk	*"****Idk*** *how females fuck with this."* (preface)	*"****Idk,*** *I'm just trying to help here"* (mitigation).
omg	*"****omg*** *same, I'm dying, it's all I've been thinking about"* (marks emotional state)	*"****OMG****, my phone has been jumping from like 39 to 0 if I open a new app; I'm fed up."* (intensifier)

Table 1. Examples from our corpus, complemented with human coding of the functions in brackets

4.2 Contrasting Human Interpretation and LLM Analysis

Analysis of "*meh*"

According to Urban Dictionary, *meh* represents doubt and functions as a shoulder shrug, with its meaning described in the entry:https://www.urbandictionary.com/define.php ?term=meh. In the Urban Dictionary it is not explicitly categorised as a DM, but in many contexts, we consider it a DM due to its low propositional contribution, procedural meaning and multifunctionality. In the tweets, it often expresses indifference and an evaluative stance as well, carrying an expression of dissatisfaction (in contrast with previous higher expectations), as in *"A: Are you excited? B: Meh, not really" in our chat corpus.* This signals that the writer or speaker finds the topic uninteresting and shows discouragement of further deep engagement on that specific point.

The LLM analysis does not always comply with reality, as *meh* is not always used initially, as described by the LLM. In fact, *meh* is common in mid-position as well in the research dataset. Moreover, it does not always serve as a DM (as was suggested by Gemini); sometimes it serves as a predicative adjective, as in "The *new update is a bit meh, I expected more features,"* where *meh* is remodified by *"a bit,"* and it means something like disappointing or unimpressive. The online version of the Cambridge Dictionary also mentions its use as an adjective.

Analysis of "*lol*" (laughing out loud)

Theoretically, this DM most commonly expresses the pragmatic function of mitigation. In our corpus, *lol* rarely indicates its literal meaning, laughter; instead, it more frequently functions as a mitigation in order to soften the blow of a critique. It is often used to signal irony or to suggest a friendly or non-confrontational tone.

Gemini analysis emphasizes the final position of this DM; however, we found several examples in our corpus where *lol* is used initially to indicate laughter, such as in "*lol that is literally the funniest thing I've seen today.*" In fact, it can be placed both at the beginning or at the end of a sentence (e.g., *"I'm dead, not looking forward to this lol"*).

Analysis of "*tbh*" (to be honest)

The main functions of this DM include stance marking and hedging. Concerning the hedging function, it is used for signaling an opinion that might be unpopular or controversial. It helps manage the user's face by showing that the statement is a subjective observation rather than an objective fact. It often appears at the beginning of a turn to frame the entire message as a moment of sincerity.

Gemini highlighted the sentence-initiality of *tbh*, but in fact, in our corpus of tweets, *tbh* was often placed on the left periphery (at the end) of sentences, as in the following examples: "*I'm just really tired of the constant drama, tbh*." "*I think the first season was better tbh*."

Analysis of "*idk*" (I don't know)

It usually carries a function of an epistemic hedge or mitigator used to signal uncertainty, a lack of commitment to a statement, or to soften the impact of a potentially controversial opinion. In this specific corpus, *idk* is pragmatically used to soften a potentially controversial opinion. It allows the speaker to simultaneously distance themselves from being expressively certain. For example**:** "idk how females fuck with this ". In this sentence, it functions as a preface to an observation, showing a personal confusion rather than an attack.

Concerning the functions of this DM, Gemini provided an analysis similar to human interpretation.

Analysis of "*omg*" (oh my god)

As indicated in the scientific literature, in our corpus *omg* signals high emotional arousal, expressing various emotions, such as surprise, shock, frustration, or excitement. It often serves to invite the interlocutor to share in their emotional state, as in *"omg, same. I'm dying, it's all I've been thinking about."* In this context, it expresses enthusiasm and acts as an intensifier for agreement.

Overall, the LLM, Gemini, provided quite an extensive pragmatic analysis in the context of digital communication (specifically Twitter), and its findings were most of the time in line with our own analysis. It managed to illustrate that the analysed expressions function less and less as literal semantic units and acquire the status of discourse markers indicating stance, which is a sign of the lexical innovation in the items. However, it included a few generalisations about the initial position of the DMs and the tone they express (e.g., *lol* expressing laughter, which was not always the case).

4.3 The Uses and Functions of Lexical Innovation in Digital Communication

Initialism and acronymization are the most visible forms of innovation representing the shift from full phrases to initialisms (*LOL, TBH, IDK, OMG*). The benefits and functions of the discussed innovative markers in digital media include speed, linguistic economy, and subcultural signalling. The innovation of lexical economy is driven by the principle of least effort. For example: *"Wait, you actually did that? lol stop."* Lexicalization demonstrates that the discussed acronymic DMs do not remain just short forms; they have become lexical units in their own right, as discussed above, *lol* being used as a verb (in *"He literally lol'd"*).

Obviously, digital communication lacks paralinguistic cues, such as tone of voice and facial expressions. Compensatory lexical innovation fills this gap with items such as *meh* or *omg* as innovative ways to represent a facial expression or a tone of voice. For instance, the DM *omg* represents innovative ways to display emotional intensity without audio or visual signals, and the DM *meh* functions as a translation of physical sounds into lexical entries. The transliteration of nonverbal cues clearly adds to lexical innovation, as *meh* represents an innovation where a nonlexical sound of a grunt is turned into a written word.

Lexical innovation may also involve pragmatic specialisation, which often involves a word becoming specialised for a specific social function. The best example is *tbh* used as a stance marker, which is almost always placed at the start of a sentence to affect the entire following message, to manage face, and to keep social politeness in a public or semi-public forum, such as Twitter.

5. Discussion about the Usability of LLMs in Discourse Analysis

The current experiment aligns with the study by Furkó (2025), stating that LLMs show the ability to identify common discourse markers and their functions. Our pilot experiment with Gemini was restricted, but it showed that the model proved surprisingly capable of explaining how new DMs function in digital conversations, often aligning well with established pragmatic theories. This suggests that AI could be an excellent help for researchers, sifting through large amounts of data to identify DMs and their environment for deeper human analysis. However, the study by Furkó (2025) shows that the AI analysis still has its limits. It tends to struggle with the subtler side of language, particularly when the meaning depends heavily on context. We saw a similar pattern in our own trial, as Gemini frequently missed the tone in our corpus and interpreted sarcastic comments as completely literal. This underscores a deficiency in pragmatic competence, because LLMs lack the situational awareness needed to safely make decisions about language appropriateness, e.g., if someone is being sincere, joking, or being sarcastic. Generally, AI sometimes still fails to grasp tone or irony because of the lack of real-world experience.

Additionally, in the current experiment, Gemini occasionally made broad assumptions about where DMs usually appear and let its own biases colour the findings in order to illustrate the initial usage of DMs.

6. Conclusion

Based on the analysis of our chat corpus, it was found that digital language includes a large number of DMs, many of which can be considered innovative. Acronymization, pragmatic specialisation, and compensatory lexical innovation are the most common lexical processes in CMC. The scrutinised digital DMs represent a class of true lexical neologisms, as they are built through compensatory innovation, which means that while tweets lack nonverbal cues such as tone of voice or facial expressions, a lexical innovation fills this gap. Words such as *meh* or the repetitive use of *omg* serve as innovative digital body language to replace a facial expression or a tone of voice that would otherwise be lost in text.

Gemini 3 shows a high level of pragmatic competence in terms of interpreting the functions of DMs but sometimes tends to overgeneralise (e.g. about the sentence-initial position of certain DMs). Gemini provides general linguistic rules about the position of DMs and simply does not mention the cases where DMs are used in uncommon positions, but we do not consider this generalisation a hallucination since the initial position is typical of DMs.

In brief, our pilot study indicates that an effective analytic approach may involve using AI tools to process large amounts of data to establish patterns, but it must be followed by critical human interpretation, especially when speakers/writers express sarcasm or irony or when semantic ambiguity is present in the text. Text generation by LLMs is not really genuine reasoning, though, and what is even more challenging for AI is to replicate pragmatic performance. While the scalability of AI is incredibly promising for linguistics, it is clear that a human analyst is still required to verify if the AI's interpretations are valid in a specific context.

7. Limitations

The limitation of the present pilot study is due to the lack of model diversity and the limited range of the analyzed data. Although we intended to include multiple LLMs, technical constraints regarding file size prevented GPT-5.2 Auto and Claude 4.5 Sonnet from performing the task, leaving the experiment as a single-model case study rather than a comprehensive comparative evaluation. The findings about the functions of these DMs in Twitter chats are in line with the functional categories of the related literature pertaining to computer-mediated communication. However, because the research is conducted on a relatively small dataset, consisting only of Twitter chats, the findings may not be fully representative of broader digital communication patterns. Besides, a monolingual approach was used. Our future research plans include testing further novel discourse markers (including multiword DM patterns as well) in multilingual and more extensive datasets.

8. Acknowledgments

This publication is based upon work from COST Action CA23147 GOBLIN—Global Network on Large-Scale, Cross-domain, and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

9. Bibliographical References

Ariel, M. 1998. Discourse Markers and Form-Function Correlations. In: Jucker, A. H. & Ziv, Y.. (Eds.) *Discourse markers: descriptions and theory*. Pragmatics and Beyond Series, 57. Amsterdam and Philadelphia: John Benjamins.

Brinton, L. and Traugott, E. 2005. *Lexicalization and language change*. Cambridge, UK: Cambridge University Press.

Fraser, B. 1999. What are discourse markers? *Journal of Pragmatics* 31, 931–952.

Furkó, P. 2014. Cooptation over grammaticalization. *Argumentum* 10, 289-300.

Furkó, P. 2025. Pragmatic markers and ideological positioning in EUROPARL: A corpus-based study. *Russian Journal of Linguistics* 29 (4). 795–816.

Geeraerts, D. 2010. *Theories of lexical semantics*. Oxford, UK: Oxford University Press.

Grieve, J., Nini, A. and Guo, D. 2017. Analyzing lexical emergence in American English online. *English Language and Linguistics* 21(1). 99-127.

Horbatko, A. O. 2021. Approaches to the definition of media discourse in modern English-language mass media. *Current issues of philology and methodology* (36-42). Sumy: Publishing and Printing Enterprise Printing Factory LLC.

Huang, Y., Guo, D., Kasakoff, A. & Grieve, J. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59. 244-255.

Kerremans, D., Stegmayr, S. and Schmid, H. 2011. The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In Kathryn Allan & Justyna Robinson (eds.), *Current methods in historical semantics*, 59-96. Berlin: Mouton de Gruyter.

McCulloch, G. 2019. *Because Internet: Understanding the New Rules of Language.* Riverhead Books.

Miller, G. 2014. *Lexicogenesis*. Oxford, UK: Oxford University Press.

Pohorila, A. I. 2022. The functioning of euphemisms in the English media discourse. *Transcarpathian Philological Studies*, 21(2), 100-103.

Schiffrin, D. 1987. *Discourse Markers.* Cambridge: Cambridge University Press.

Scott, K. 2015. The pragmatics of hashtags: Inference and conversational style on Twitter. *Journal of Pragmatics.* 81. 10.1016/j.pragma.2015.03.015.

Szmrecsanyi, B. 2011. Corpus-based dialectometry: A methodological sketch. *Corpora* 6(1). 45-76.

Tagliamonte, S. and Derek, D. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* - AMER SPEECH. 83. 3-34.

Traugott, E. G. 1995. The Role of the Development of Discourse Markers in a Theory of Grammaticalization. *Paper given at the 12th International Conference on Historical Linguistics.* Manchester; 13–18, August, 1995.

Traugott, E. and Trousdale, G. 2013. *Constructionalization and constructional changes*. Oxford: Oxford University Press. pp. 304.

Vandekerckhove, R. 2025. "OMG! Why discourse markers thrive in interactive social media writing" In: Fábián, A. & Trost, I. (eds.) *Impulses and Approaches to Computer-Mediated Communication: Proceedings of the 12th International Conference on Computer-Mediated Communication and Social Media Corpora.* University of Bayreuth.

Wise, M. & El Barj, H. N. 2023. PragMaBERT: Analyzing pragmatic markers in political speech. *CS224N Project Report.* Stanford University.

10. Language Resource References

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc (accessed on 6 October, 2025)

Anthropic. (2025). *Claude 4.5 Sonnet* [Large language model]. https://claude.ai/ (accessed on 27 December, 2025)

Google. (2025). *Gemini 3 Pro* [Large language model]. https://gemini.google.com/ and https://aistudio.google.com/ (accessed on 28 December, 2025)

Marsan's Twitter chat corpus repository: https://github.com/marsan-ma/chat_corpus/blob/master/twitter_en.txt.gz, subpart of this dataset: https://github.com/marsan-ma/chat_corpus (accessed on 5 November, 2025)

"Meh" in *Cambridge Dictionary*: https://dictionary.cambridge.org/dictionary/english/meh (accessed on 15 November, 2025)

"Meh" in *Urban Dictionary*: https://www.urbandictionary.com/define.php?term=meh (accessed on 15 November, 2025)

OpenAI, Inc. (n.d.). Models. *GPT-5.2.* [Large language model]. Platform.openai.com; OpenAI, Inc. https://platform.openai.com/docs/models/ (accessed on 28 December, 2025)

LREC 2026

NEOLOGY AND LARGE LANGUAGE MODELS

Workshop Proceedings

Editors
Valunaite Oleskeviciene Giedre and Giouli Voula

16 MAY 2026

ISBN 978-2-493814-70-8

EAN 9782493814708

Organizing Committee

Florentina Armaselu (University of Luxembourg)

Voula Giouli (Aristotle University of Thessaloniki)

Barbara Lewandowska-Tomaszczyk (University of Applied Sciences in Konin)

Chaya Liebeskind (Jerusalem College of Technology)

Barbara McGillivray (King's College London)

Giedre Valunaite Oleskeviciene (Mykolas Romeris University)

Scientific Committee

Florentina Armaselu (University of Luxembourg, Luxembourg)

Giorgio Maria Di Nunzio (University of Padua, Italy)

Radovan Garabík (Ludovit Stur Institute of Linguistics, Slovak Republic)

Voula Giouli (Aristotle University of Thessaloniki)

Anas Fahad Khan (CNR-Istituto di Linguistica Computazionale "Antonio Zampolli," Italy)

Barbara Lewandowska-Tomaszczyk (University of Applied Sciences in Konin)

Chaya Liebeskind (Jerusalem College of Technology, Israel)

Elpida Loupaki (Aristotle University of Thessaloniki, Greece)

Barbara McGillivray (King's College London)

Liudmila Mockiene (Mykolas Romeris University, Lithuania)

Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences, Poland)

Atul Kr. Ojha (University of Galway, Ireland)

Ana Ostroški Anić (Institute of Croatian Language and Linguistics, Croatia)

Marko Robnik-Šikonja (University of Ljubljana, Slovenia)

Purificação Silvano (University of Porto, Portugal)

Enriketa Sogutlu (Beder University, Albania)

Ranka Stankovic (University of Belgrade, Serbia)

Giovanni Luca Tallarico (University of Verona, Italy)

Giedre Valunaite Oleskeviciene (Mykolas Romeris University, Lithuania)

Andrius Utka (Vytautas Magnus University)

Federica Vezzani (University of Padua, Italy)

Table of Contents