



Leveraging Artificial Intelligence for Assigning ILR Ratings to Authentic Content

Jordan M. Eason

Doctoral Student, University of Coimbra, Portugal

When teaching Brazilian Portuguese and other languages for the Department of Defense (DoD), the decision on which authentic texts to choose can be difficult. One complicating factor is choosing an authentic text that is at the correct proficiency level for the students—not too challenging and not too easy. The DoD uses the Interagency Language Roundtable (ILR) levels to indicate the complexity of any text or listening passage. Learning how to “level” (i.e., assign a rating score) a passage according to the ILR is a skill that must be developed and takes a significant amount of time. This paper leverages experience working for the DoD as a member of a team that was assigned the task to level audios, videos, and text; during this process, differences in opinions among members occurred regarding whether a passage was ratable/unratable, or able/unable to fit in a specific level on the ILR scale. I argue that machine learning processes available in artificial intelligence (AI), specifically a natural language processing (NLP) platform such as Open AI’s ChatGPT or Google’s Bard, among other AI chatbots, offer human raters a tool that can assist them in increasing their efficiency while removing potential subjectivity from the leveling process.

Keywords: *artificial intelligence, AI chatbots, ChatGPT, leveling content, ILR scale*

INTRODUCTION

Artificial intelligence (AI) systems are producing communication responses that are increasingly similar to those produced by humans. Although they cannot yet replace humans, these systems can facilitate and help language professionals do complex tasks. One such complex task that AI can assist with is the leveling (i.e., assigning a rating or score) of language texts from authentic written sources according to the descriptions established by the Interagency Language Roundtable (ILR). The pilot study outlined in this paper provides a glimpse of AI’s potential to assist the language professional in assigning levels to texts in the DoD context. With time, AI will

APPLIED LANGUAGE LEARNING (ISSN 2164-0912) is the journal of the Defense Language Institute Foreign Language Center, in the U.S. Department of Defense. The views expressed herein are those of the author(s), not the Department of Defense or its elements. Further reproduction is not advisable. Whenever copyrighted materials are reproduced in this publication, copyright release has ordinarily been obtained only for use in this specific issue. Requests for reprints should be directed to the individual authors.

likely be able to support human raters as they determine ILR reading levels for authentic texts, bringing efficiency and reducing potential subjectivity.

BACKGROUND

Machine Learning and Natural Language Processing

AI chatbots such as OpenAI's ChatGPT or Google's Bard are becoming increasingly well-known and utilized across all fields. These tools use a chat-like interface in which the user can ask the application just about anything and the application uses AI to answer the question. These AI tools are developing rapidly and will continue to improve in quality and accuracy over time. The most common method AI chatbots use patterns of language to train is via machine learning algorithms (MLAs). MLAs are used to analyze a large corpus of annotated text data. This algorithm is then trained to use these patterns to predict the sentiment of new, unannotated text. Sentiment analysis, which involves analyzing the language used in text to determine its emotional tone, is a Natural Language Processing (NLP) technique.

For some languages such as Brazilian Portuguese, where high-quality training data exists, the corpora used to train MLAs can be fine-tuned to a specific domain or context, which further improves the ability of the AI chatbot to interpret sentiments in each domain. For example, the AI chatbot may take into consideration specific terms that trigger sarcasm such as, "*Pois é*," in Brazilian Portuguese which is equivalent to the term "sure" in English. Furthermore, it can use contextual factors, such as the tone of the overall text, the relationship between parties in a conversation, and the subject matter of the text to determine the emotional undertones of a message. Although limited by their finite nature in reference to an infinite world, AI chatbots may be considered more capable than limited in their ability to "read between the lines." Considering that humans are imperfect, can be irrational, and have diverse educational backgrounds and world experiences, one may argue that an AI chatbot's ability to "read between the lines" and provide consistent analysis makes it more capable than most humans because of the AI's ability to collect and process information from an infinite repository of online information.

AI chatbots use NLP platforms to understand and respond to human language. NLP platforms have been used to tackle minor tasks for language professionals and teachers of foreign languages because of NLP's capabilities with language recognition and manipulation. NLP is capable of differentiating text types to include some intentions from authors. Academics such as Guren (2022) have demonstrated that NLP platforms can be used to simplify difficult-to-understand technical texts and make them more comprehensible. This relates to AI rating levels because it confirms that NLP platforms can identify and recognize the level of difficulty and even styles in various texts. If NLP is used to manipulate Brazilian Portuguese and other languages when an adequate amount of data is present in a specific manner, it can be used to recognize the difficulty and assign an ILR level as well. Despite achieving great success with the lexical simplification process, Guren (2022) does not claim that AI is a solution or remedy for all situations. Notably, additional data and surveys were needed to explain technical terms and broaden the fields of study. This means NLP could have equal challenges with technical terms and idiomatic expressions as well, despite the major advances in NLP since 2019. In his survey of

automated text simplification tools, Shardlow (2014) discusses how these tools can identify complex texts and generate substitutions that are determined to be simpler, focusing on lexical and syntactic simplification as well as text simplification and machine translation in languages other than English. Similarly, Bott et al. (2012) demonstrated how structural simplification operations can be accomplished for Spanish. In English, popular sites such as Grammarly and Rephrasely use NLP to make suggestions for students' papers. Again, to successfully substitute, the NLP must first recognize what is classified as a particular style or difficulty. This is what the NLP must do to rate the language level of the excerpt as well. These systems can provide summaries of a writer's tone or style and can also make suggestions on how to revise a paragraph to make it sound more "academic, human, simple, clear, elaborate, creative, formal, cohesive, confident, smart, or thoughtful" (Rephrasely.com). Although it is not possible to truly understand what an author intends, we can argue that most would agree on some specific meaning such as when trying to classify them as creative, formal, etc.

AI and NLP tools are also increasingly able to identify and analyze the inferential meaning of words in a text. Shaik et al. (2022) discuss how AI and NLP use feature extraction, feature selection, topic modeling, and text evaluation to analyze texts (in this case, student feedback from a course), including how these analyses can identify and interpret challenging language features such as sarcasm or ambiguity. Similarly, Nenkov et al. (2016) demonstrated how the integration of NLP platforms with tutoring systems can be used to improve learning management systems (LMSs) by providing language feedback via a chatbot. This confirms that NLP is capable of differentiating text types and, to some degree, interpreting authors' intentions.

Classification of complex texts is a critical step in leveling and one that can be supported by AI as well. Dzikovska et al. (2014) discuss the potential of NLP techniques for students in STEM areas such as electricity and electronics, fields that are useful for military electronics training and therefore relevant for DoD texts and related tasks. Dzikovska et al.'s research, which focused on developing an intelligent tutoring system, utilized automatic text classification and information extraction. It shows that automatic text classification can assist language professionals to classify texts, which is one of the critical decisions content developers and ILR text-level raters make daily.

Thus, the above studies provide support that NLP platforms can successfully differentiate between complex and simple texts, can simplify language, and to some degree can interpret subtleties in human communication such as sarcasm. They are also increasingly being utilized for their potential to make teaching and learning languages more efficient. In fact, the Department of Defense has created, along with MIT, an online tool for leveling texts. The Auto-ILR was developed in 2017 and automatically levels and classifies texts for DoD curricular contexts. It aids teachers and students as they assign levels to texts, making their teaching and studying processes more efficient. At this time, however, this tool does not appear to be accessible to the public. As can be seen, AI and tools such as ChatGPT or Google Bard have the potential to support DoD language professionals in terms of consistency and efficiency as they assign ILR levels to authentic texts.

Text Leveling in the Department of Defense

Text leveling in the Department of Defense is performed by human language professionals. While overall the process is accurate and consistent, there are some complications due to staffing shortages, inter-rater reliability, and other considerations. For example, some of the language professionals doing the rating are not highly proficient in the language for which they provide ILR levels. These professionals use tools to translate between English and the target language to determine the ILR level. Given the complexities and subtleties of language, this is not an ideal solution. Another challenge in rating passages objectively arises when a team working to assign a level to a text disagrees on what the appropriate rating should be, which reflects the natural complexity of authentic language. A team leader ultimately makes the final rating decision, which introduces subjectivity; depending on the individual and their training and background with the language and project, this may be significant. However rare this occurrence may be, it does highlight the potential challenges introduced by human subjectivity and the opportunity for AI to assist human language experts as they make leveling decisions.

OPPORTUNITIES FOR AI AS A TOOL TO ASSIGN ILR LEVELS

The Interagency Language Roundtable Scale

The intricacies of the ILR scale are important in this pilot study. For the sake of national security and organizational confidentiality, this paper will focus only on the information given on the ILR website (<https://www.govtilr.org/>). The ILR scale is briefly described in the table below:

Table 1

The Interagency Language Roundtable Scale

Scale Number	Description
0	No proficiency
0+	Memorized Proficiency
1	Elementary Proficiency
1+	Elementary Proficiency, Plus
2	Limited Working Proficiency
2+	Limited Working Proficiency, Plus
3	General Professional Proficiency
3+	General Professional Proficiency, Plus
4	Advanced Professional Proficiency
4+	Advanced Professional Proficiency, Plus
5	Functionally Native Proficiency

Each level has a more detailed description. To demonstrate the subtle differences between levels, we will analyze the different definitions for Levels 2 and 2+ Reading. A more complete

description of ILR level 2 Reading is found in Figure 1. (note that this reading descriptor is from the most recent ILR scale).

Figure 1

ILR Level 2 Reading Descriptor

“Sufficient comprehension to read simple, authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context. Able to read with some misunderstandings straightforward, familiar, factual material, but in general insufficiently experienced with the language to draw inferences directly from the linguistic aspects of the text. Can locate and understand the main ideas and details in material written for the general reader. However, persons who have professional knowledge of a subject may be able to summarize or perform sorting and locating tasks with written texts that are well beyond their general proficiency level. The individual can read uncomplicated, but authentic prose on familiar subjects that are normally presented in a predictable sequence which aids the reader in understanding. Texts may include descriptions and narrations in contexts such as news items describing frequently occurring events, simple biographical information, social notices, formulaic business letters, and simple technical material written for the general reader. Generally the prose that can be read by the individual is predominantly in straightforward/high-frequency sentence patterns. The individual does not have a broad active vocabulary (that is, which he/she recognizes immediately on sight), but is able to use contextual and real-world cues to understand the text. Characteristically, however, the individual is quite slow in performing such a process. Is typically able to answer factual questions about authentic texts of the types described above” (govtilr.org).

The definition has a strong presence of subjectivity. For example, several terms in the definition are ambiguous. It’s difficult to determine what exactly is meant by the following terms: “familiar context,” “some misunderstandings,” “uncomplicated, but authentic prose,” “quite slow,” or “typically able.” One may find it difficult to pinpoint what is considered “familiar” or quantify “some.” This is particularly true for the professional language community because of a diverse set of backgrounds. This reveals an advantage of using AI when rating authentic texts. AI will define such terms consistently throughout any project.

Figure 2 shows the ILR description for the 2+ Reading level.

Figure 2

ILR Level 2+ Reading Descriptor

“Sufficient comprehension to understand most factual material in non-technical prose as well as some discussions on concrete topics related to special professional interests. Is markedly more proficient at reading materials on a familiar topic. Is able to separate the main ideas and details from lesser ones and uses that distinction to advance understanding. The individual is able to use linguistic context and real-world knowledge to make sensible guesses about unfamiliar material. Has a broad active reading vocabulary. The individual is able to get the gist of main and subsidiary ideas in texts which could only be read thoroughly by persons with much higher proficiencies. Weaknesses include slowness, uncertainty, inability to discern nuance and/or intentionally disguised meaning” (govtilr.org).

Again, the 2+ Reading level uses ambiguous terms. For example, the following terms are particularly ambiguous: “non-technical prose,” “familiar topic,” “real world knowledge,”

“sensible guesses,” “unfamiliar material,” “broad active reading vocabulary,” “slowness,” and “disguised meaning.” The term “slowness” may be a comprehension response time of three seconds for some readers while others may think it as a different value. The term may be defined in an entirely different fashion. Similarly, what is a “familiar topic” to one individual may be unfamiliar to another when cultural experiences and interests are taken into account. The use of AI has the potential to lessen this subjectivity, which is preferred in ILR ratings.

The definition of Reading Level 3 includes statements shown in Figure 3.

Figure 3

ILR Level 3 Reading Descriptor

“Text-types include news stories like wire service reports or international news items in major periodicals, routine correspondence, general reports, and technical material in his/her professional field; all of these may include hypothesis, argumentation and supported opinions. Misreading rare. Almost always able to interpret material correctly, relate ideas and “read between the lines,” (that is, understand the writers’ implicit intents in text of the above types).”
(govtilr.org)

A fundamental question for higher levels on the ILR scale is whether the AI chatbot itself can “read between the lines.” At this point, it appears that it can, and given the nature of AI and its potential to improve, it is likely that its ability to “read between the lines” will also be more consistent than with human levelers. An important step in the process of working with AI is to determine the degree to which the systems have access to the information they need. It therefore might be prudent to also query the AI chatbot about what it knows about the ILR, perhaps at specific levels, to ensure that the system can provide accurate information.

Limitations of AI Chatbots

AI chatbots are naturally limited by the input that they receive. They are only capable of learning from the input and language patterns that they are fed, and some language patterns are more complex and difficult than others. Content developers for the DoD or those who have personally taken a Defense Language Proficiency Test (DLPT) understand the importance of being able to “read between the lines” in the target language. For example, one may be asked when a writer is being indifferent, sarcastic, or provocative among other emotions and sentiments. To explore this issue further, I used Open AI’s ChatGPT to determine if it can understand various emotions expressed in text. It returned the following answer to the given question (see Figure 4).

Figure 4

ChatGPT Example Demonstrating “Reading Between the Lines”

<i>User</i> “Given this text in Brazilian Portuguese, what is the tone and emotion that the author intends to convey in the following text: “Não interessa se foi A, B ou C, todo o episódio foi como uma facada nas minhas costas”
<i>ChatGPT</i> “Based on the given text in Brazilian Portuguese, the tone and emotion that the author intends to convey can be described as one of betrayal, disappointment, and possibly anger. The phrase “Não interessa se foi A, B ou C” suggests a sense of indifference or dismissal towards the specific individuals involved. However, the following statement “todo o episódio foi como uma facada nas minhas costas” (translated as “the whole episode was like a stab in the back”) conveys a strong feeling of betrayal, implying that the author feels deeply hurt and betrayed by the situation. The use of the metaphorical expression “facada nas minhas costas” emphasizes the emotional impact and the sense of personal betrayal experienced by the author. Overall, the author’s tone appears to be one of bitterness and resentment towards the events described.” (https://chat.openai.com/)

While this particular assessment by ChatGPT of the tone and emotion in this passage is consistent with my interpretation as a native Brazilian Portuguese speaker, it is important to keep in mind that the ability of ChatGPT and AI tools to provide accurate assessments will vary depending on the language and also on the type of input the AI tool has been trained on up to this point.

Pilot Study

For this pilot study, I compared the levels produced by human raters and Open AI’s ChatGPT chatbot for nine authentic texts available in the Defense Language Institute’s (DLI) Global Language Online Support System (GLOSS, <https://gloss.dliflc.edu/>). The texts chosen for this pilot study were written in Brazilian Portuguese and were all well under the maximum amount of text allowed in ChatGPT. GLOSS includes the official, human-rated ILR level for each of these texts. For the ChatGPT level, I entered each text into ChatGPT and asked for it to assign an ILR level using the following request:

“Based on the Interagency Language Roundtable scale, please rate the following passage in Brazilian Portuguese for the Department of Defense context: ‘(INSERT TEXT HERE)’”

Table 2 shows the title of the texts, the GLOSS human-leveled rating, the ChatGPT AI-leveled rating, and the difference between the two (i.e., Delta). Note that the GLOSS passages included in this pilot study could have been rated under a previous version of the ILR Reading scale.

Table 2

GLOSS and ChatGPT Ratings of Texts

Title of Authentic Text (year)	GLOSS Rating (human-rated)	ChatGPT Rating (AI-rated)	Delta
Dilma Appoints New Commanders of the Armed Forces (2015)	1	2+	1.5
142 Million Voters Can Check Polling Places (2014)	1	2	1
Brazilian Airlines in the Air (2013)	1	2	1
CIA Documents about the Dictatorship (2018)	2	2	0
The Kidnapping of Charles Burke Elbrick (2015)	2	2	0
Farming in Southern Brazil (2016)	2	2	0
The Controversy of Cracolândia (2017)	3	3	0
Scientific Experiment Tests Ethical Limits (2017)	3	3–4	0.5
The Unconscious Mind (2017)	3	4	1

The average delta between human- and machine-assigned ratings in this sample is 0.56. with ChatGPT rating texts at least half a level higher than the DoD human experts. Also, out of nine texts compared, only four have the same rated level in both human and AI ratings. Based on my experience in content creation for the DOD, the deltas seem to be an accurate representation of potential variations of ratings assigned by human experts.

One of the benefits of using ChatGPT for leveling is that it explains why it chose a specific level and provides specific examples. For example, when asked why the authentic text in Brazilian Portuguese entitled “The Unconscious Mind” was rated as Level 4 on the ILR scale, ChatGPT responded with these details after demonstrating it understood what Level 4 was defined as (see Figure 5).

Figure 5

ChatGPT's Explanation of Rating The Unconscious Mind at ILR 4

"In the given text, the interviewees discuss the theories and contributions of Sigmund Freud, including his concept of the unconscious, the interpretation of dreams, and the importance of understanding the meaning behind the symptoms and behaviors. The interview also touches on the history and evolution of psychiatry and psychoanalysis. The conversation contains several characteristics that support a Level 4 rating: 1. Abstract and complex vocabulary: The text includes advanced vocabulary related to psychology and psychiatry, such as "neuroses," "associação livre," "psicoterapêutico," "Psiquiatria dinâmica," "processos psíquicos," "sintomas," and "delírio." These terms may not be familiar to the average person and require a higher level of proficiency to understand. 2. Long and complex sentences: The text includes several long and complex sentences that require advanced grammatical structures to understand. For example, "A interpretação dos sonhos, ela vai fazer uso da associação livre, que é um dispositivo a mais que a gente tem no atendimento." 3. Use of idiomatic expressions: The text includes several idiomatic expressions, such as "chamar uma pessoa por outro nome," which require a deep understanding of the language and its nuances. 4. Discussions of complex and abstract concepts: The text delves into complex and abstract topics related to psychology, including the unconscious, the interpretation of dreams, and the meaning behind symptoms and behaviors. These topics require a high level of intellectual engagement and critical thinking to understand fully." (<https://chat.openai.com/>)

This response highlights how AI can be a useful tool for DoD content developers and foreign language instructors because of its ability to highlight the language that would qualify a text as a specific level using the ILR definitions.

ChatGPT highlighted ambiguity issues in rating for the text entitled "Scientific Experiment Tests Ethical Limits" when it returned an ILR level range of 3–4 for this particular text. One can conclude that because ChatGPT produced a range, it shows the tool's ability to interpret the definitions of the various ILR levels; this range also highlights the general difficulty of assigning a level to a text. Given that having ILR level ranges for a given text is often not permitted, one can assume that outputs with score ranges are equivalent to what human levelers would determine "unratable texts."

CONCLUSION

NLP platforms such as AI chatbots show potential in their ability to support human language professionals who assign ILR levels to texts. Because of the limitations of AI, it is not recommended that these tools replace human levelers at this time, but rather that professionals explore ways to use AI as a type of "personal assistant" when making rating decisions. AI offers the potential benefits of increased efficiency and consistency, decreased subjectivity, and a starting point for further evaluation. It also provides extensive explanations of how a machine-assigned ILR rating was determined. The small pilot comparison between ChatGPT-assigned and human-assigned ratings presented in this article can be used to inform future research and development. For example, texts in languages other than Brazilian Portuguese can be compared to gather baseline information on similarities and differences in the ratings given. Also, research can be conducted on the specific needs of human language professionals and how AI might help meet those needs. AI developers may benefit from collaboration with DoD professionals to

provide input for AI algorithms that will enhance the training of AI in this area. Tools such as chatbots have potential implications for the classroom as well. Teachers can use them to determine the ILR level of a text they want to use in class, or to find an authentic text at a certain ILR level. Research is therefore warranted on how best to engage teachers in professional development for using AI in the classroom. Over time, opportunities for AI to contribute to the teaching and learning process will increase and be limited only by our imaginations.

REFERENCES

- Bott, S., Saggion, H., & Mille, S. (2012). *Text simplification tools for Spanish*. Proceedings of the eighth international conference on language resources and evaluation (LREC'12), Istanbul, Turkey. http://www.lrec-conf.org/proceedings/lrec2012/pdf/762_Paper.pdf
- Dzikovska, M., Steinhauser, N., Farrow, E., Moore, & J., Campbell, G. (2014). BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, 24(3). 284–332. <https://doi.org/10.1007/s40593-014-0017-9>
- Guren, E. G. (2022). *Exploring automatic text simplification of requirements*. [Masters thesis. University of Oslo]. https://www.duo.uio.no/bitstream/handle/10852/95635/8/thesis_final.pdf
- Nenkov, N., Dimitrov, G., Dyachenko, Y. & Koeva, K. (2016). Artificial intelligence technologies for personnel learning management systems. *2016 IEEE 8th International Conference on Intelligent Systems (IS)*. Sofia, Bulgaria, 2016, pp. 189–195, <https://doi.org/10.1109/IS.2016.7737420>
- Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1). 58–70. <https://doi.org/10.14569/SpecialIssue.2014.040109>
- Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L. (2022). A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access*, 10, 56720–56739. <https://doi.org/10.1109/ACCESS.2022.3177752>

AUTHOR

Jordan M. Eason, is a doctoral student at the University of Coimbra, Portugal. A former active duty Air Force officer and Language Enabled Airman (LEAP), Mr. Eason has worked as a Portuguese instructor contractor for ICA Languages on behalf of the Defense Language Institute, the US Air Force Academy, the University of Maryland, and the University of Michigan.