Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 16, No. 2
journal.magisz.org

A Case Study Evaluating Different Language Models for Norwegian Agriculture

Kristian Nikolai Jæger Hansen[1]

[1]*kristianikolai@gmail.com*

I N F O
Received: 31/05/2025
Accepted:23/07/2025
Available on-line: 29/09/2025
Responsible Editor: László Várallyai

Keywords:
n-gram models,
Artificial intelligens,
Integrated plant protection

A B S T R A C T

This article aims to start the development of language models for agronomic advisory service. This is achieved by evaluating various n-gram language models with different smoothers, including Modified Kneser-Ney, Add-k, and Absolute. The models were constructed using an earlier collected dataset on Norwegian agriculture. They were adapted to provide practical agronomic advice on integrated pest management. Model performance was measured using perplexity. The Add-k (k=0.1) scored perplexities of 1920, 7800, 13600, and 16200 for 2-, 3-, 4-, and 5-grams, respectively. In this study, the Modified Kneser-Ney (D=0.8, D=0.8, and D=0.8) performed best, achieving perplexities of 494, 363, 344 and 339 for the same orders. The best performing model, the Modified Kneser-Ney model (trigram) can predict sentences such as "Meadowgrass is a grass that grows in more or less dense lawns". However, before being practically useful, the models need further development.

1. Introduction

This study establishes a foundation for developing specialised language models in Norwegian agriculture. Such development is crucial because agricultural practices vary across regions and between countries due to differences in management approaches, agroclimatic conditions, and the prevalence of pests and weeds. These variations necessitate region-specific interventions to sustain productivity. Consequently, if artificial intelligence (AI) tools are to be applied as advisory tools, they must be trained or adapted using region-specific data to enhance accuracy and minimise the risk of generating misleading or incorrect information.

The general application of AI in agronomy has been discussed for its promising potential, spanning from soil sciences to supporting advisory services (e.g., Castro et al., 2024; Tzachor et al., 2023). However, with a few notable exceptions (e.g., Balaguer et al., 2024), the adoption and evaluation of language models in agriculture remains limited, often involving the use of a general language model without adaptation or development prior to evaluation (Yang et al., 2024). Adapting or creating a model from scratch for language models might improve the availability of answers to region-specific agronomic questions and overcome language barriers if constructed in languages other than English.

AI models adapted to a specific dataset can be constructed in different pathways, such as fine-tuning or Retrieval-augmented generation of existing large language models (LLMs), improving model accuracy (Balaguer et al., 2024). However, a third pathway also exists using pure statistical models such as *n*-gram models, which can be trained at a lower cost. Markow initially developed these models in 1913, and they are now described as Markow chains (e.g, bigrams and trigrams) that predict upcoming letters in text (Jurafsky & Martin, 2025).

This study examines the reliability of n-gram-based language models in domain-specific data describing Norwegian integrated pest management in agriculture. The article aims to initiate the development of language models for practical agronomic advice on integrated pest management in Norwegian.

2. Material and methods

2.1. Data description

The dataset used in this study was compiled by Bugaiova & Hansen (2025) and consists of 1,090 articles, 33,928 sentences, and 27,092 words. It describes integrated pest management in Norway, visualised with bigrams in Table 1. The dataset was retrieved from the webpage plantevernleksikonet.no (Plant Protection Lexicon), based on web scraping. The web page planteverleksikonet.no is maintained by the Norwegian Institute of Bioeconomy Research (NIBIO), reviewed and written by agronomic experts in Norwegian.

Table 1: for eight of the words in the whole dataset of the plantevernleksikonet_text_data.jason file of 33,928 sentences. Each cell shows the count of the column label word following the row label word before any smoothing. Bekjempelse av ugras og skadedyr, translated to; managment of weeds and pests.

	bekjempelse	av	ugras	og	skadedyr
bekjempelse	12	147	0	9	0
av	0	0	3	185	8
ugras	0	1	0	16	0
og	11	33	8	1	4
skadedyr	1	1	0	13	0

The text data were sourced from a JSON file containing all articles related to plantevernleksikonet.no. Each article included a title, summary, and content, future described in Bugaiova & Hansen (2025).

2.2 Text processing, tokenisation and software

Tokenisation, i.e. splitting text into words and assigning them values for more straightforward preprocessing, was performed by splitting the cleaned text into words based on whitespace. The entire corpus was thus converted into a sequence of tokens. The original dataset was divided into three datasets, a training set (70%), a validation set (10%), and a test set (20%), by a random sampling approach.

The dataset was imported using the jsonlite package in R, and individual articles were structured into a tibble format using the dplyr package. The work presented in this article was developed and applied within the programming language R (R Core Team, 2021) using the RStudio IDE (RStudio Team, 2020). The packages jsonlite, stringr, dplyr, kgrams, ggplot2 were used to construct and evaluate the models.

2.3 Language models

The n-gram models are statistical language models. The *n*-gram models uses the n-1 word to predict the next word, thus assigning a probability distribution over all possible sentences in a text. The probability of a sentence is decomposed into a product of word continuation probabilities, where each probability represents the likelihood of a word occurring given its preceding context. In *n*-gram models, an important simplifying assumption is made: the continuation probability of a word depends only on the last $n-1$words, thereby also assigning probabilities to entire sequences of the model, as seen in formula 1.

$$P(W_n|W_{1:n-1}) \approx P(w_n|w_{n-1}) \quad (1)$$

This simplification retains sufficient context for many practical applications while reducing computational complexity.

Different smoothing techniques were applied to address the inherent sparsity in *n*-gram counts. Smoothing redistributes probability mass from frequent to infrequent *n*-grams, thereby mitigating the impact of rare events on the overall model. Thus, smoothing algorithms provide a way to estimate probabilities for unforeseen events in the training data. Hence, it is necessary when the dataset varies and has unexpected words. The Modified Kneser-Ney (MKN) smoothing models, referred to as MKN (D=0.4), MKN (D=0.5), and MKN (D=0.8), with a uniform discount at D1, D2, and D3 across all *n*-

gram orders. This approach uses absolute discounting by subtracting a fixed value from the probabilities of lower order terms to omit *n*-grams with lower frequencies, proposed by Reinhard Kneser, Ute Essen and Hermann Ney in 1994 (Ney et al., 1994). The article also explored Absolute Discounting, which, after this, is only named Absolute, with discount values of 0.3, 0.5, and 0.8. This approach subtracts a constant from observed counts while redistributing the removed probability mass (Ney et al., 1994). In addition, Laplace smoothing, defined as Add-k smoothing, was tested using constant values of k=0.1, k=0.5, and k=0.75 to address zero-frequency issues by adding a fixed value to each *n*-gram count.

2.4 Model evaluation

The perplexity was calculated on the test dataset to measure how well the models were trained. The perplexity measures how well the model predicts unseen data, and lower perplexity values indicate a better model fit (Formula 2).

$$perplexity\ (W) = \sqrt[N]{\frac{1}{P(W_1 W_2 \ldots W_N)}} \quad (2)$$

The best-performing model was then used to generate sentences of domain-relevant Norwegian terms by adhering to the statistical patterns learned from the training corpus. Text prediction was done by quiring on specific words and building sentences with a defined length of *n*=13.

3. Results, discussion and limitations

3.1 Model performance

The MKN (D=0.8, D=0.8 and D=0.8) achieve the lowest perplexities across 2- to 5-gram models around 494, to 339. By contrast, Add-k (k=0.1, k=0.5, and k=0.75) smoothing performed relatively poorly, with perplexities of 1920, 3250 and 4150 with the bigram models. However, while >*n*=3, all the Add-k models exceeded 10,000 in perplexity (Figure 1 and Table 2). The performance of Add-k, i.e. having a high perplexity, is in line with Church & Gale (1991). The overall ranking of the different smoothing techniques in this study is in line with the findings in Chen & Goodman (1999), who found that MKN were the most efficient smoothing technique on text data from the United States from the brown Corpus and Wall Street Journal (Chen & Goodman, 1999).

3.2 Sentence predictions

The best-performing model, MKN as a trigram model, randomly generated sentences by sampling tokens according to the model's probability distribution. This process produced the following Norwegian sentences, based on the words "tunrapp", "Hundekjeks" og "Kveke" (English and latin name: medowgras i.e. *Poa annua* L., cowparsly i.e. *Anthriscus sylvestris* L. and quackgras i.e. *Elymus repens* L.): (1) "Tunrapp er et gras som vokser i mer eller mindre tette tuer strået", (2) "hundekjeks er vanlig i hele landet til opp i snaufjellet til 1230m" and (3) "kveke bekjempes med glyfosat utenom selve kulturen eller med kvekemidler i tofrøbladete kulturer". When the sentences is directly translated to English, they become (1) "Medowgrass is a gras that grows in more or less dense lawns straw", (2) "Cowparsley are common throughout the country, even in mountain eras up to 1200 m", and (3) "Quackgras is controlled with glyphosate outside of the crop itself, or with quackgras-specific control agents in dicotyledons crops". These sentences have a partly unnatural flow and some agronomic errors.

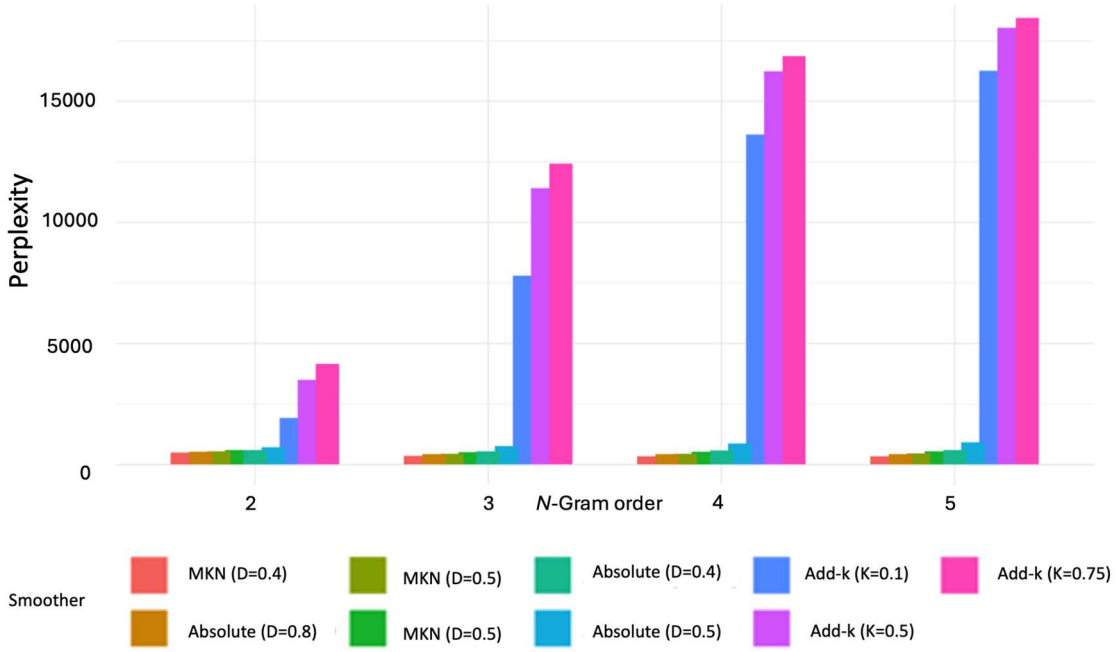


Figure 1: The perplexity of the evaluated language models trained with different smoothing techniques and evaluation across N-gram orders 2 to 5 on the test dataset. The evaluated models are MKN (D=0.4), MKN (D=0.5), MKN (D=0.8), Add-k (k=0.1), Add-k (k=0.5), Add-k (k=0.75), Absolute (D=0.3), Absolute (D=0.5), and Absolute (D=0.8).

Table 2: The test perplexity of language models trained with different smoothing techniques and evaluation across N-gram orders 2 to 5 in absolute values on the test dataset. The evaluated models are MKN (D=0.4), MKN (D=0.5), MKN (D=0.8), Add-k (k=0.1), Add-k (k=0.5), Add-k (k=0.75), Absolute (D=0.3), Absolute (D=0.5), and Absolute (D=0.8).

Smoother	n=2	n=3	n=4	N=5
MKN (D=0.4)	595	522	538	549
MKN (D=0.5)	554	454	455	459
MKN (D=0.8)	494	364	344	339
Add-k (k=0.1)	1920	7800	13600	16200
Add-k (k=0.5)	3500	11400	16200	18000
Add-k (k=0.75)	4150	12400	16900	18400
Absolute (D=0.3)	721	765	870	921
Absolute (D=0.5)	606	549	580	598
Absolute (D=0.8)	539	432	428	430

3.3 limitations

This study lays the foundation for developing specialised language models in Norwegian agriculture from scratch. However, to be implemented in practice, both model performance and the quality of generated sentences must be improved. A future pathway to increase the practical uasability of the models could be to introduce large-language models adapted to an agronomic context as seen in Balaguer et al. (2024).

Context-specific and verified data are often limitations in training or adapting language models and LLMs. Norwegian is considered a low-resource language, restricting the applicability of these models primarily to Norway, Denmark, and Sweden. From an agronomic perspective, even within this region, model reliability may be compromised by variations in agroclimatic conditions, management practices,

and agronomic policies. Expanding the dataset to include additional environmental factors, such as soil properties, fertilisation regimes, and climate conditions, could significantly enhance performance and adaptability. However, model performance should be improved before prioritising working on obtaining a higher data resolution. Parts of the issue with practical usability relate to the limitations of *n*-gram models, as discussed by Zhu et al. (2024), and might be solved using large language models (LLMs).

The models developed here rely on *n*-gram approaches with various smoothing techniques. Although these models offer interpretability, they have two limitations compared to artificial neural network-based large language models. First, the number of parameters grows exponentially with the *n*-gram order, which can lead to sparsity issues. Second, *n*-gram models do not generalise well to unseen word sequences, limiting their performance when encountering new vocabulary (Jurafsky & Martin, 2025).

Conclusion

The most effective *n*-gram model in this study was the Modified Kneser-Ney model with D=0.8, D=0.8 and D=0.8, which achieved perplexities of 494, 364, 344, and 339 for 2-, 3-, 4-, and 5-grams, respectively. This configuration proved optimal for the Norwegian integrated pest management dataset. In contrast, the worst-performing model was Add-k smoothing with k=0.1, recording a perplexity of 1920, 7800, 13,600 and 16,200 with for 2-, 3-, 4-, and 5-grams gram models.

The Modified Kneser-Ney model generated sentences by sampling words from the test corpus based on their probability distribution, producing content that closely resembled the original dataset. This study exemplifies the development of language models in an agronomic context. However, from a practical agronomic perspective and in terms of grammatical accuracy, the generated sentences still leave room for improvement.

Conflict of Interest

The authors declare no conflict of interest

ORCID

Kristian Nikolai Jæger Hansen: https://orcid.org/0000-0002-5875-7300

Supplemental Material

The dataset used in this article can be accessed through Bugaiova & Hansen (2025). The code for the work conducted is provided by request.

Refferences

Balaguer, A. *et al.* (2024) RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. *arXiv* [online]. Available at: http://arxiv.org/abs/2401.08406.

Bugaiova, O. & Hansen, K. (2025) A dataset dedicated to the training of large language models for agronomic management practices and production in Norwegian agriculture. *Data in Brief*, 59, 111326. doi:10.1016/j.dib.2025.111326.

Castro, J. P., Gasch, C. K. & Flores, P. (2024) Artificial intelligence in soil science: Where do we go now? *Agricultural and Environmental Letters*, 9(2), e20134. doi:10.1002/ael2.20134.

Chen, S. F. & Goodman, J. (1999) An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4), pp. 359–394. doi:10.1006/csla.1999.0128.

Church, K. W. & Gale, W. A. (1991) A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5(1), pp. 19–54. doi:10.1016/0885-2308(91)90016-J.

Jurafsky, D. & Martin, J. H. (2025) N-gram language models. In: *Speech and language processing* (3rd ed., draft of January 12, 2025). Available at: https://web.stanford.edu/~jurafsky/slp3/3.pdf.

Ney, H., Essen, U. & Kneser, R. (1994) On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8(1), pp. 1–38. doi:10.1006/csla.1994.1001.

R Core Team (2021) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

RStudio Team (2020) *RStudio: Integrated Development for R*. Boston: RStudio, PBC. Available at: http://www.rstudio.com/.

Tzachor, A. *et al.* (2023) Large language models and agricultural extension services. *Nature Food*, 4(11), pp. 941–948. doi:10.1038/s43016-023-00867-x.

Yang, S. *et al.* (2024) GPT-4 as Evaluator: Evaluating Large Language Models on Pest Management in Agriculture. *arXiv* [online]. Available at: http://arxiv.org/abs/2403.11858.

Zhu, H., Qin, S., Su, M., Lin, C., Li, A., & Gao, J. (2024) Harnessing Large Vision and Language Models in Agriculture: A Review.