Compositionality and Event Retrieval in Complement Coercion: A Study of Language Models in a Low-resource Setting

Matteo Radaelli
Norwegian University of Science and Technology
matteo.radaelli@ntnu.no

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuele.chersoni@polyu.edu.hk

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Giosuè Baggio
Norwegian University of Science and Technology
giosue.baggio@ntnu.no

Abstract

In sentences such as *John began the book*, the complement noun phrase, lexically denoting an entity, is interpreted as denoting an event. This is known in linguistics as *complement coercion*: the event associated with the verb is not overtly expressed but can be recovered from the meanings of other constituents, context and world knowledge. We investigate whether language models (LMs) can exploit sentence structure and compositional meaning to recover plausible events in complement coercion. For the first time, we tested different LMs in Norwegian, a low-resource language with high syntactic variation in coercion constructions across aspectual verbs. Results reveal that LMs struggle with retrieving plausible events and with ranking them above less plausible ones. Moreover, we found that LMs do not exploit the compositional properties of coercion sentences in their predictions.

1 Introduction

Sentences like *John began the book* are examples of complement coercion, originating from a type-mismatch between the required verb argument and the observed one (Pustejovsky, 1991, 1995): the aspectual verb (e.g., *begin*) semantically requires an event-denoting argument but is composed with an entity as its syntactic complement. Although the event is not overtly expressed, a plausible candidate can often be recovered by exploiting lexical and contextual information (Pustejovsky, 1991, 1995; Lapata and Lascarides, 2003): the sentence above can be interpreted as meaning *John began* {reading, writing, ...} *the book.*

Complement coercion has drawn attention as a potential violation of the Fregean principle of compositionality. Compositionality implies that all aspects of sentence meaning should derive from the meanings of the constituent parts and the way they are combined syntactically (Asher, 2015; Baggio et al., 2012). The interpretation of various covert elements results from enriched composition: semantic processes that exploit conceptual meaning, discourse context and world knowledge (Pustejovsky, 1991, 1995; Jackendoff, 1997; Baggio, 2018).

Language Models (LM) based on the Transformer architecture (Vaswani et al., 2017) have demonstrated remarkable capabilities in a wide range of NLP tasks, including natural language understanding. Despite their success, few studies have focused on enriched composition phenomena such as complement coercion (Gu, 2022; Ye et al., 2022). Some studies have investigated LM performance viewing complement coercion as an event retrieval task and demonstrating the challenges of recovering underlying semantic information from coercion sentences (Rambelli et al., 2020; Ye et al., 2022; Gietz and Beekhuizen, 2022; Gu, 2022; Im and Lee, 2024; Rambelli et al., 2024). However, most studies have been conducted in English, a language with low variability in the syntax of coercion constructions. As a consequence, little is known about the interplay of syntax and semantics in covert event retrieval in LMs: (how) do machines exploit compositional properties of coercion sentences to arrive at plausible interpretations?

The current study makes three contributions. First, to our knowledge, it is the first study of LMs on complement coercion that uses a language other than English (Norwegian) and that evaluates and compares different LMs (autoencoders and autoregressive models). Second, we investigate the interaction between different aspectual verbs Katsika et al. (2012) and post-verbal constituents in canonical syntactic constructions. Norwegian shows some variation in how complement coercion is syntactically realized, and therefore allows us to probe whether LMs are sensitive to syntactic and compositional semantic properties of these constructions across aspectual verbs. Finally, Norwegian is cur-

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 469–480
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

rently considered a low-resource language (Kummervold et al., 2022; Liu et al., 2024; Samuel et al., 2025), and we are releasing our evaluation dataset for complement coercion resolution in Norwegian. Complement coercion with aspectual verbs is statistically rare in Norwegian corpora (see below): recovering implicit events could be challenging for a 'data hungry' technology such as LMs.

2 Related Work

2.1 Complement Coercion in Norwegian

Complement coercion has been studied in several high-resource languages. Apart from English, we find studies on German (Rüd and Zarcone, 2011; Zarcone and Padó, 2011; Zarcone et al., 2012, 2014), French (Godard and Jayez, 1993; Pustejovsky and Bouillon, 1995), Dutch (Sweep, 2012), and Chinese (Hsu and Hsieh, 2013), while there has been little research on Scandinavian languages. Spalek (2015) analyzed the cessation verb *avslutte* (to conclude), comparing Norwegian, English, Spanish and German. Spalek concluded that coercion is limited to a reduced set of entities that can be combined with the verb, especially "information-content entities" (e.g., text) (Spalek, 2015, p. 531). Spalek and Sæbø (2019) argued that Norwegian speakers tend to combine dynamic verbs with specific particles that denote a particular stage of the event (e.g., *å stryke ferdig*, to finish ironing).

Radaelli and Baggio (2025) conducted a study on the Norwegian Colossal Corpus (NCC) (Kummervold et al., 2022), a large set of corpora that includes approximately 21M documents for a total of 7B tokens. The study examined a wider class of aspectual verbs than previous theoretical research: *begynne* (to begin), *starte* (to start), *fortsette* (to continue), *ende* (to end), and *avslutte* (to conclude). The authors found that the syntax of complement coercion in Norwegian can vary according to the aspectual verb. Initiation verbs are usually combined with PPs introduced by the prepositions *på* or *med*:

(1) *Gutten begynte | startet på | med boken.*
(The boy began | started [with] the book.)

These combinations appear with higher frequency in complement coercion sentences compared to other aspectual verbs. The continuation verb *fortsette* introduces coercion mainly with *med*-PPs and, to a lesser extent, directly with nominals:

(2) *Gutten fortsatte [med] boken.*
(The boy continued [with] the book.)

The cessation verb *avslutte* prefers direct objects, while *med*-prepositional phrases appear less often:

(3) *Gutten avlsuttet [med] boken.*
(The boy finished [with] the book.)

Not all aspectual verbs can trigger complement coercion (e.g., the verb *ende* was excluded), nor do aspectual verbs significantly differ in occurrence frequency in coercion constructions.

The corpus analysis confirms the findings of Spalek (2015) and Spalek and Sæbø (2019): complement coercion occurs with a restricted set of entity categories. Although a similar trend can be found also in other languages (e.g., see Verspoor (1997) for English and Rüd and Zarcone (2011) for German), Norwegian shows even less variability, reducing the set of entities primarily to everyday objects such as text, music, songs, food and drinks.

Considering Pustejovsky's Generative Lexicon perspective (Pustejovsky, 1995), the productivity of coercion can also be limited by the interaction of syntactic and semantic factors. If, on the one hand, entities admit either AGENTIVE or TELIC qualia readings, their combination with prepositions may further reduce the set of plausible event candidates. The preposition *med* appears to play a 'passe-partout' role, with greater flexibility in event interpretation, including not only default qualia readings but also contextual information, if present. The preposition *på*, on the other hand, tends to further constrain interpretations: the corpus data showed a stronger tendency to express AGENTIVE interpretations with entities that are created rather than used. Radaelli and Baggio (2025) also found that Norwegian speakers prefer to express similar concepts to complement coercion through a broad range of phrasal constructions (e.g., *å sette i gang*, to begin). The study concluded that complement coercion is a relatively low-frequency phenomenon, with around 1500 cases over 80,000 sentences with aspectual verbs and syntactic constructions compatible with coercion.

2.2 LM Approaches to Complement Coercion

Before the LM era, complement coercion interpretation was carried out via either probabilistic (Lapata and Lascarides, 2003; Shutova, 2009; Shutova et al., 2013) or distributional semantic modeling (Zarcone et al., 2012, 2013; Chersoni et al., 2017;

McGregor et al., 2017; Chersoni et al., 2021). In one of the first studies testing LMs on complement coercion, Rambelli et al. (2020) evaluated the events retrieved by pretrained models of the BERT and the GPT families. They found that LMs performed well, but not significantly better than the best distributional models.

Ye et al. (2022) argued that Transformer-based models can learn coercion interpretations via *dense paraphrasing* (DP): DP involves the reformulation of a given coercion sentence in a way that eventive information is revealed, ambiguity is removed and the original sentence meaning is preserved. They found that BERT struggles in interpreting coercion, but a fine-tuning with explicitly paraphrased sentences improved its performance.

Finally, Gu (2022) investigated the behavior of GPT-2 on complement coercion by analyzing surprisal estimates. The goal was to understand how LMs process coercion constructions at the VP. Significant surprisal effects were observed at the target region, aligning with psycholinguistic findings of increased processing costs at the complement (McElree et al., 2001; Traxler et al., 2002; Baggio et al., 2010, 2011, 2016).

3 Experimental Settings

3.1 Task Proposal

In previous research on complement coercion in LMs, evaluations typically compared a narrow set of high-likelihood predictions against a predefined set of gold standard outputs. In contrast, our study introduces a novel evaluation approach, based on a ranked prediction distribution of class-specific verbs, rather than just the most probable outputs: for every context-neutral sentence[1] s belonging to a set S, a given model m generates a set of top-k ranked output predictions $O = \{o_1 ... o_k\}$. We then evaluate each output with a mean average precision metric, allowing us to determine not only whether a model predicts covert events, but also to what extent LMs consistently predict plausible event interpretations in their rankings.

The distribution should reflect a re-ranking of tokens when the model is exposed to coercion sentences, providing evidence of its sensitivity to coercion. In cases where a LM is exposed to a sentence such as *The boy began the book*, we expect that the combination of the triplet <subject, aspectual verb, entity> would result in a re-ranking of candidate implicit events (see Figure 1): the ranking should reflect the interaction of the triplet composition, where plausible verbs (events) are collocated at the top of the rank as the most likely interpretations.

Instead of using a set of predefined events, our study will consider any event that meets the syntactic and semantic constraints of complement coercion as correct. According to Piñango and Deo (2016) and Spalek and Sæbø (2019), the covert event of a complement coercion sentence should be *telic*: combined with the subject and complement, it should establish a natural endpoint or goal state.

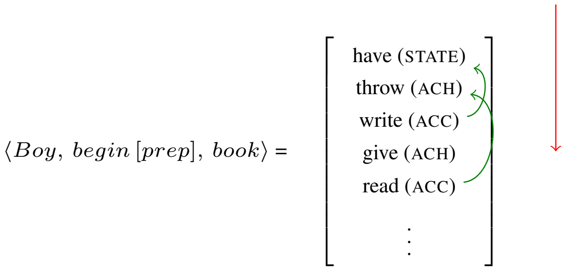

Figure 1: An example of re-ranking candidate events when the expressions in the given triplet are composed in a sentence. The matrix shows output verbs associated with their Aktionsart class such as state, achievements (ACH), and accomplishments (ACC).

Given the above requirements, events predicted by LMs should be evaluated considering their Aktionsart (lexical aspect) class by using Vendler's classification system (Vendler, 1967). We identify the class of **accomplishment** verbs as our ground truth in this task, as they denote dynamic and durative actions with a specific endpoint, aligning with the telicity criterion by Spalek and Sæbø (2019). As there is no predefined set of implicit events for interpreting a coercion sentence, we consider as compositionally plausible candidates all predicted verbs that belong to the accomplishment class. In case a model predicts events weakly associated with a specific coercion triplet (e.g., *begin the book* → *eat*, see Lascarides and Copestake 1998), this does not necessarily indicate low performance: the output can count as correct, if the retrieved event is an accomplishment. It is possible to construct contexts where even apparently deviant events are plausible, so long as they are accomplishments: e.g., *The goat began* {eating} *the book*.

[1] Context-neutral, canonical coercion sentences include the subject, the aspectual verb and its complement, with unmarked word order and no additional sentence context.

3.2 Dataset

We created a new dataset of sentence pairs with (a) a context-neutral sentence with a coercion triplet and variable syntactic structure (*på*-NP, *med*-NP, NP) and (b) a sentence prompting event resolution:

(a) `Kim {VERB-FIN} {PREP|Ø} {ENTITY-DEF}.`

(b) `Det som Kim {VERB-FIN} å gjøre, var å [MASK]. (What Kim {VERB-FIN} to do, was [MASK].)`

Each placeholder in brackets is replaced with the relevant lexical item. The template encompasses a combination of the following elements:

- 90 entities ({ENTITY-DEF}) were carefully selected to represent real artifacts, avoiding abstract and ambiguous concepts. In addition, following Piñango and Deo (2016, p. 387), we used entities that can be semantically interpreted as "incremental theme arguments of the implicit event", a crucial element in coercion configurations. We included entities that never occurred in coercion sentences in the NCC corpus study, ensuring that the models, especially those trained exclusively on NCC, are exposed to sentences not seen during pretraining. Six distinct entity categories were used: *food*, *text*, *clothing*, *everyday objects*, *construction/housing*, and *entertainment*. All nouns were only used in definite form.
- Four aspectual verbs ({VERB-FIN}), namely ***begynne*** (begin), ***starte*** (start), ***fortsette*** (continue), and ***avslutte*** (finish) were composed with each entity. The verb was always presented in the simple past form (*preteritum*) in both sentences in a pair.
- Three syntactic constructions were used ({PREP|Ø}): the complement is either introduced by a PP with the prepositions **på** or **med** followed by the NP denoting an entity, or only by the latter NP.
- The same subject was used for every sentence, with a neutral name (*Kim*) to avoid gender and other biases that may affect the results.
- In all pairs, the prompt (b) included the [MASK] token the model has to predict.

A total of 1080 sentence pairs in standard written Bokmål form were used with each model.

Model	# Par.	Tr. Data
MBERT CASED/UNCASED	178M	3.3B*
NB-BERT-BASE	178M	7B
NB-BERT-LARGE	355M	7B
NORBERT	111M	1.9B
NORBERT2	125M	15B
NORBERT3-base	123M	25B
NORBERT3-large	353M	25B
NORBERT3-SMALL	40M	25B
NORBERT3-XS	15M	25B
NORBLOOM-7B-SCRATCH	7B	26.7B
NORGPT-369M	369M	25B
NORGPT-3B	3B	25B
NORGPT-3B-CONTINUE	3B	25B
NORLLAMA-3B	3B	26.7B
NORMISTRAL-7B-SCRATCH	7B	26.7B
NORMISTRAL-7B-WARM	7B	26.7B

Table 1: Tested LMs with approximate information on number of parameters (*#Par.*) and training data (*Tr. Data*). *mBERT was trained on 114 languages.

3.3 Models

We evaluated a total of 17 different pre-trained Norwegian LMs varying in architecture, parameter size and training data. The models belong to two broad families: BERT-like autoencoder models, and autoregressive models such as GPT-2 (Radford et al., 2019), LLAMA-2 (Touvron et al., 2023), Mistral (Jiang et al., 2023) and Bloom (Scao et al., 2023) (Table 1). All models are available on Hugginface[2].

3.4 Baseline Model

To assess event retrieval in complement coercion by LMs, it is necessary to use a baseline model, here provided by the NCC, an open-source corpus used for training most LMs in Norwegian [3]. For each entity in the dataset, we extracted the most likely verbs (events) associated with the entity. The extracted verbs were determined on the basis of the Pointwise Mutual Information (PMI) score, a metric evaluating the association strength between two words $w1$ and $w2$ (Church and Hanks, 1990):

$$\mathrm{PMI}(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1) \cdot P(w_2)}$$

In our study, the score was calculated through the joint probability between each sentence predicate (event) and its object (entity) in the entire

[2] https://huggingface.co/

[3] The National Library of Norway, the maintainer of the NCC corpus, in December 2024 has updated its distribution policies, limiting access to some subcorpora. This has caused a decrease in the corpus size from approximately 7B to 4.5B tokens. For further information, see https://huggingface.co/datasets/NbAiLab/NCC

corpus. The PMI score provides a measure of general lexical association between verbs and entities. A comparison with this baseline allows us to understand whether an LM genuinely interprets covert events based on coercion settings or simply mirrors statistical co-occurrence patterns observed during pre-training.

3.5 Evaluation and Annotation

We used two common evaluation metrics. One is mean average precision (mAP) (see Manning et al. (2009, from p. 159) and Kotlerman et al. (2010)):

$$\text{mAP} = \frac{1}{S}\sum_{s=1}^{S} \text{AP}(s)$$

It consists of the weighted means of average precision (AP) scores across all sentences (S):

$$\text{AP(q)} = \sum_{k=1}^{5} P(k) \cdot \Delta R(k)$$

where P is the precision value calculated at the cut-off rank k and $\Delta R(k)$ is the change in recall (R) from rank $k-1$ to k. mAP provides the ranking direction of models when complement coercion occurs. A high mAP value indicates a model that mostly considers accomplishment verbs in the prediction list, collocating them at the top, whereas a low mAP value suggests a failure in prioritizing accomplishment verbs as completions.

The second metric is the mean top-ranked accuracy (A1) in all sentences, considering only the most likely prediction in the ranking. This metric allows us to study what types of verb (events) the models consider as the most salient ones.

4 Results and Task Discussion

Table 2 shows the performance results of all LMs in the covert event retrieval task in Norwegian, with mAP and A1 scores. Model performance varies according to the interplay of two main factors: model framework and model size (number of parameters and training data). The NORBERT3 family shows relatively high performance compared to other BERT-like frameworks, with NORBERT3-BASE and NORBERT3-LARGE outperforming the baseline on both measures. Larger LMs outperform NB-BERT models and the previous generations of NORBERT models, which showed poorer performances, possibly due to less training data available. Models like NORBERT3-XS performed less well probably due to their reduced parameter size, despite the same amount of training data.

Model	mAP	A1
NCC (baseline)	0.59	0.47
NORGPT-369M	0.56	0.54
NORGPT-3B	0.48	0.42
NORGPT-3B-CONTINUE	0.46	0.42
NORLLAMA-3B	0.71	0.67
BERT-BASE-MULTILINGUAL-CASED	0.07	0.00
BERT-BASE-MULTILINGUAL-UNCASED	0.27	0.22
NB-BERT-BASE	0.38	0.33
NB-BERT-LARGE	0.54	0.47
NORBERT	0.25	0.18
NORBERT2	0.44	0.34
NORBERT3-BASE	0.63	0.58
NORBERT3-LARGE	0.60	0.55
NORBERT3-SMALL	0.59	0.55
NORBERT3-XS	0.29	0.16
NORBLOOM-7B-SCRATCH	0.46	0.34
NORMISTRAL-7B-SCRATCH	0.38	0.29
NORMISTRAL-7B-WARM	0.63	0.54

Table 2: Mean average precision (mAP) and top-rank accuracy (A1) results in the covert event retrieval task in Norwegian. NORLLAMA-3B is the best performing model overall.

Almost all GPT-2-based models, as well as NORBLOOM-7B-SCRATCH and NORMISTRAL-7B-SCRATCH performed poorly, ranking below the baseline, despite their size. NORMISTRAL-7B-WARM outperforms the baseline in both cases, compared to the version trained from scratch: pretraining on the English vanilla version and successive pretraining on Norwegian data may have given the model an advantage, allowing for the transfer of rich representations from English text. Finally, NORLLAMA-3B can be considered as the most capable model among those tested here. Its success could be attributed perhaps to its large training corpus, with more than 25B training tokens in Norwegian and other Scandinavian languages.

Language models generally struggle to perform the completion task. Overall low mAP scores suggest difficulties in generating plausible accomplishments among high-ranked candidate mask replacements. This is confirmed when cross-analyzing A1 scores: even the best model, NORLLAMA-3B fails to reach a 70% level of accuracy, indicating that non-accomplishments and other implausible verbs are predicted as candidate interpretations at the top of the list. Similarly, the top-10 ranked models achieve an A1 score ranging from 0.42 to 0.58: they have around 50% chance of failing to rank accomplishments at the top.

We will now turn to an analysis of model performance taking into account both mAP and A1

NorLlama-3B		mAP	A1
verb	prep		
avslutte	ø	0.66	0.61
	med	0.75	0.69
	på	0.64	0.53
begynne	ø	0.75	0.72
	med	0.79	0.81
	på	0.73	0.71
fortsette	ø	0.64	0.57
	med	0.64	0.56
	på	0.59	0.46
starte	ø	0.79	0.80
	med	0.81	0.83
	på	0.76	0.76

Table 3: Mean average precision (mAP) and top-rank accuracy (A1) results for NorLlama-3B categorized by aspectual verbs (*begynne*, *starte*, *fortsette* and *avslutte*) and syntactic composition (introduced by prepositions, *på* or *med*, or by a nominal, *ø*) in coercion sentences.

scores subdivided according to aspectual verbs and their syntactic structures in coercion sentences. For the sake of simplicity, we will consider the best performing model NorLlama-3B. The results are shown in Table 3.

Consistently high mAP scores are found with initiation verbs. The verb *starte* shows high mAP scores reaching 0.81 precision when entity arguments in coercion sentences are introduced by the preposition *med*, 0.79 with nominals and 0.76 with the preposition *på*. The verb *begynne* was associated with worse performance, while showing a similar trend as *starte*. Sentences with entity arguments introduced by *med* reached 0.79 precision, 0.75 with nominals and 0.73 with *på*. The two remaining aspectual verbs showed similar results, and arguments with *med* as preposition obtained higher precision scores. In sentences with *fortsette*, both nominals and *med*-prepositional phrases reached the same score (0.64).

A1 scores show a similar trend. The model performs better when coercion sentences are introduced by *starte*, with 0.83 of A1 accuracy when the entity NP is introduced by *med*-PP, 0.80 without a preposition and 0.76 by *på*-PP. The verb *begynne* also serves as a trigger for complement coercion, with an A1 score of 0.81 with *med*-prepositional phrases, 0.72 with simple nominals and 0.71 with verb argument phrases introduced by *på*.

Two key observations are suggested by this analysis. First, different aspectual verbs are associated with differences in model performance. Our results indicate that the model can recover the implicit meaning more easily with initiation verbs in coercion sentences. This is consistent with the corpus analysis of Radaelli and Baggio (2025), which showed that among all aspectual verbs, initiation verbs feature more frequently in coercion sentences. Second, we only find weak differences in performance as a function of the syntax of post-verbal constituents. This suggests that the type of syntactic structures in complement coercion sentences plays only a minor role in the model's process of recovery of implicit meaning.

4.1 Sentence surprisal

Previous studies (see above) indicated that LMs struggle to consistently retrieve covert events in complement coercion sentences. To understand the reasons behind these prediction difficulties, one can study the model's behavior when it is exposed to complement coercion sentences. We conducted a further analysis that complements the previous ranking results by computing surprisal estimates for coercion sentences. Surprisal is used in NLP and psycholinguistic studies to quantify effort during sentence processing (Hale, 2001; Levy, 2008; Smith and Levy, 2013; Salicchi et al., 2023; Oh and Schuler, 2023; Shain et al., 2024):

$$S(w_i) = -\log_2 P(w_i \mid w_1, \ldots, w_{i-1})$$

Surprisal measures how unexpected a given word (w_i) is, given its left context ($w_1 ... w_{i-1}$). Higher surprisal values indicate greater processing difficulty, as upcoming words are less predictable.

We will use surprisal to assess whether models expect an entity-denoting noun in coercion triples. Specifically, we will compare surprisal estimates for complement coercion sentences (e.g., *Kim begynte på boken*, Kim began (on) the book) with their overt counterparts (e.g., *Kim leste boken*, Kim read the book). The events were selected considering the highest PMI scores between each accomplishment and its associated entity. In total, we examined 2,160 sentences, using the same sentences from the previous task (1,080 coercion, 1,080 overt) combining all aspectual verbs, all entities, and the same three different syntactic structures. To compute surprisal estimates, we used log-probabilities provided by model logits. As coercion and overt sentences may differ in length, we computed surprisal for sentences as the mean of each word's surprisal:

$$S_{mean}(s) = \frac{1}{N}\sum_{i=1}^{N} S(w_i)$$

where N is the number of tokens in a sentence s. Here too, we tested surprisal for NorLlama-3B as a high performance model in this task. For the calculation of surprisal estimates, we used the tool *minicons* on Python. The data were analyzed using Wilcoxon signed-rank tests to compare the surprisal values between coercion and overt sentences. We hypothesized that the model would show higher surprisal values for coercion sentences than for overt ones. The results confirmed the hypothesis, showing a statistical difference in surprisal ($W = 367176, p < 0.001$). This suggests a tendency of the model to assign prediction logits with lower probabilities for coercion sentences.

In order to analyze the extent to which syntactic structure can influence surprisal in coercion sentences, we compared two regression models. As a baseline, we ran a model on surprisal using only sentence type (coercion *vs* explicit) and sequence length as predictors. The second model also included syntactic structure as a predictor (with *på*-PPs, *med*-PPs and direct nominals as levels). The baseline model ($R^2 = 0.173$) revealed that coercion sentences significantly increased surprisal. Moreover, sequence length negatively correlated with surprisal, meaning that longer sentences led to lower surprisal values. The second model ($R^2 = 0.181$) shows a significant positive trend in the coercion condition, as the baseline model. On the other hand, sequence length shows in this case a positive effect on surprisal. Sentences with *med*-prepositional phrases demonstrate lowest surprisal, while sentences with *på* exhibit slightly higher surprisal, but still lower than in the nominal conditions. Comparing the variance of the two models ($\Delta R^2 = +0.008$), we find small improvements attributable to syntax. Prepositions therefore reduce surprisal in comparison to sentences with direct nominals, where *med*-sentences led to lower surprisal, followed by *på* (Figure 2).

5 Error Analysis

To study model errors, a relatively straightforward approach is to examine the overall prediction distribution of events and their Aktionsart. For practical reasons, the analysis is restricted again to the best performing model, NorLlama-3B. The analysis revealed the following findings. First, among 5,400 predictions made in 1,080 coercion sentences (divided by 90 different entities and presented with 4 different aspectual verbs and 3 post-verbal syntactic constructions), the model predicted 68 unique events. This small set suggests, on the one hand, that the model tends to predict events by avoiding many unrelated or random outputs. However, the low variation of events also suggests a tendency to reuse the same verbs across many entities.

	Baseline model **Coefficient (β)**	**Model With Syntax** **Coefficient (β)**
Intercept (Nominals)	9.7983 ($p < 0.001$)	2.0463 ($p < 0.001$)
Coercion	0.9429 ($p < 0.001$)	1.4946 ($p < 0.001$)
Sequence Length	-1.0612 ($p < 0.001$)	1.3388 ($p < 0.001$)
Explicit	—	0.5517 ($p < 0.001$)
Med	—	-2.5418 ($p < 0.001$)
På	—	-2.2583 ($p < 0.001$)

Table 4: Effects of syntax on sentence surprisal.

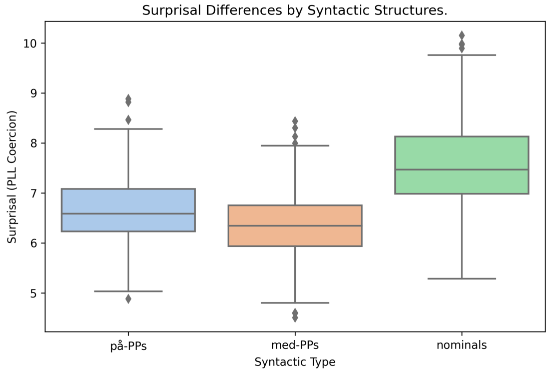

Figure 2: Surprisal values across syntactic structures.

Second, the distribution of events as predicted by the model is skewed and follows a Zipfian law, with the first most frequently predicted events in the ranking accounting for a substantial proportion of the total distribution, while the frequency of verbs ranked at lower positions rapidly decreases. Table 5 presents the distribution of the first 10 most predicted events across all coercion instances, including both their absolute and relative frequencies based on the total 5,400 predictions (5 predictions per instance). The most frequent verbs predicted by the model are *lage* (make), followed by *sette* (put/set), *ta* (take), *få*, and *gi* (give), which together amount to almost 67% of the total predictions. The remaining verbs have lower frequencies with a considerable subset of events that occur only once. This long-tail behavior further strengthens the hypothesis of a biased tendency of the model towards a very limited set of events.

Third, based on a close qualitative examination

Verb	Freq. (Rel._freq)
lage (make)	969 (0.18)
sette (put/set)	768 (0.14)
ta (take)	762 (0.14)
få (get/receive)	739 (0.14)
gi (give)	387 (0.07)
skrive (write)	175 (0.03)
spille (play)	155 (0.03)
legge (lay/put down)	113 (0.02)
male (paint)	109 (0.02)
dele (share/divide)	96 (0.02)
gå (go / walk)	96 (0.02)
sy (sew)	94 (0.02)
blande (mix)	84 (0.02)
strikke (knit)	83 (0.02)
synge (sing)	81 (0.02)

Table 5: Top predicted events made by NORLLAMA-3B, showing both absolute frequency and relative frequency out of a total of 5,400 model outputs.

of the restricted set of predicted events at the top of the ranking, we can notice the following patterns. The most frequently output events are usually non-accomplishment verbs: the only acceptable accomplishment verb is *lage*, which is the most frequent. Yet, this verb is polysemous and can be combined with a wide range of entities, denoting an action of creating or producing something, e.g., *lage pizza* (make pizza), *lage skulptur* (make a sculpture) and *lage sang* (compose a song). In contrast, other frequently predicted events, like *ta* (take), *få* (get) and *gi* (give), are either implausible in many coercion instances or typically denote achievements, and are therefore not acceptable in coercion constructions.

Despite a quite positive performance overall of NORLLAMA-3B in interpreting coercion items, the strong presence of semantically inappropriate verbs in the ranking may be due to their high frequency in the corpora used during pretraining. Since the task was designed to constrain the model to retrieve only infinitival verbs, the prediction of verbs that are not plausible accomplishments suggests that the model may rely more on the co-occurrence frequency between a verb and its nominal object during training, rather than on the semantic compatibility between the event and the entity, even in contexts in which a more compositionally appropriate event would be expected and could be retrieved.

6 General Discussion and Conclusion

The analyses carried out in the present study clearly show that complement coercion remains an open challenge for LMs in low-resource languages such as Norwegian. We investigated the extent to which LMs could recover implicit events in complement coercion sentences. If models recognize these as coercion constructions, that require event retrieval, they should be able to distribute verb (event) predictions in such a way that accomplishments are ranked as the most probable covert events.

However, the outcomes of the event retrieval task indicate that LMs still have difficulties recovering viable implicit events. In particular, A1 scores are consistently low across models, suggesting a failure to retrieve potential accomplishment verbs as the most likely event predictions in the task. Moreover, the mAP scores confirmed the models' limitations, as they fail to systematically and consistently rank accomplishment verbs higher. Only few models could outperform the baseline, whose predictions are based on simple statistical calculations on the NCC corpus frequency: this is significant, considering that such models were trained on corpora 3.5 times larger than the baseline. The results also highlight performance differences across models:

- NORLLAMA-3B outperformed all the models that were tested here; its success may be due to its new improved architecture and training optimization (e.g., SwiGLU activation function, Grouped query attention (GQA) mechanism, rotary positional embeddings), combined with a large amount of training data.
- On the other hand, even the largest GPT-class models could not perform the task efficiently. The traditional autoregressive GPT-2 may lack an architecture that can capture covert information like covert events.
- Even LMs such as NORBLOOM-7B-SCRATCH and NORMISTRAL-7B-SCRATCH performed poorly for their size. Their low performance could be due to training carried out exclusively on Norwegian data, especially compared to the best performing NORMISTRAL-7B-WARM with a pre-training phase that also included English data.
- The NORBERT3 family, in particular the base and large versions, could attain moderate performance levels despite their reduced number

of parameters. The BERT architecture then seems to be well-suited for learning and storing world knowledge and relational knowledge between words during pretraining, making them effective in cloze tasks (Petroni et al., 2019; Rogers et al., 2020). In addition, their customized autoencoder framework, incorporating the extended MLM pre-training task (Samuel et al., 2023), may have facilitated acquisition of syntactic and semantic information relevant for the present task. To this purpose, it should be noted that also in the complement coercion study of Rambelli et al. (2020) in English, a bidirectional architecture (RoBERTa) was the one showing the highest correlations with human production frequencies for the candidate covert event. However, vanilla architectures combined with less training data would drastically reduce performance as seen in the NB-BERT models.

To better understand how LMs process complement coercion sentences and investigate the causes behind their difficulties in event retrieval task, we compared surprisal estimates between coercion sentences and their overt event counterparts. Higher surprisal values for complement coercion sentences suggest that LMs generally find coercion constructions less predictable, which should be expected given their relative infrequency in Norwegian corpora. However, rare constructions in human language can still be interpreted compositionally by exploiting lexical meaning and syntactic structure, even when context is minimal or absent (Baggio, 2018, 2021). Overall, our results suggest that many LMs are largely unable to make productive use of the available compositional information to generate accomplishments as plausible event completions in complement coercion sentences. These results apply to Norwegian, but may well extend to other languages with similar characteristics, such as other Scandinavian or Germanic languages, and to other constructions infrequent in linguistic corpora.

Language models have been often argued to lack ‘common sense’, which makes them unsuitable as (general) problem solvers in real-world situations. Our results show that LMs may also have limited *linguistic common sense*, the ability to select and use all and only relevant (non)linguistic knowledge to interpret inputs to comprehension and learning (Lascarides and Copestake, 1998; Piñango and Deo, 2016; Baggio, 2018; Rambelli et al., 2024).

A more detailed analysis of the best performing model (NorLlama-3B) revealed only moderate variation in performance according to the specific aspectual verb used. Initiation verbs lead to better performance. Based on results of corpus studies, this may be due to stronger statistical associations between these aspectual verbs and (particular classes of) entity-denoting nominals. However, we could not find clear differences between different syntactic constructions within the same aspectual verbs, which suggests that models do not exploit differences in syntactic structure to recognize these as coercion constructions and accordingly attempt the retrieval of plausible accomplishments.

Linear regression models were also employed to assess whether coercion surprisal estimates were influenced by the syntactic structures proposed in the dataset. Results revealed weak differences in surprisal estimates, especially between coercion sentences with entity-denoting complements introduced by prepositions or directly by NPs, showing greater processing difficulties in the latter cases. This partially aligns with the results presented in table 3, where nominals led to lower scores, while *med*-PPs were associated with better performance. Furthermore, LM behavior aligns weakly with the NCC corpus study by Radaelli and Baggio (2025): the authors found that *med*-prepositional phrases occur more frequently in coercion constructions and allow greater flexibility in event interpretations.

Considering LM’s failure to exploit compositionality (lexical meaning and syntactic structure) with complement coercion sentences, future work should explore what other factors can impact LM’s performance in this type of task. There are at least two possible research directions. First, an analysis of the role of linguistic context as a factor in performance improvement: what aspects of sentence or discourse context can facilitate event retrieval? Second, an analysis of the extent to which LM’s performance is dependent on ontology: can event retrieval be facilitated by specific classes of entities, as is suggested by theoretical linguistic and corpus research?

7 Acknowledgments

EC was supported by a GRF grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 15612222). We thank three anonymous reviewers for their valuable suggestions.

References

Nicholas Asher. 2015. Types, meanings and coercions in lexical semantics. *Lingua*, 157:66–82.

Giosuè Baggio, Keith Stenning, and Michiel Van Lambalgen. 2016. Semantics and cognition. In Maria Aloni and Paul Dekker, editors, *The Cambridge Handbook of Formal Semantics*, pages 756–774. Cambridge University Press.

Giosuè Baggio, Michiel Van Lambalgen, and Peter Hagoort. 2011. The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, 26(9):1338–1367.

Giosuè Baggio, Michiel Van Lambalgen, and Peter Hagoort. 2012. The processing consequences of compositionality. In Markus Werning, Wolfram Hinzen, and Edouard Machery, editors, *The Oxford Handbook of Compositionality*, pages 655–672. Oxford University Press.

Giosuè Baggio. 2018. *Meaning in the Brain*. MIT Press.

Giosuè Baggio. 2021. Compositionality in a parallel architecture for language processing. *Cognitive Science*, 45(5):e12949.

Giosuè Baggio, Travis Choma, Michiel van Lambalgen, and Peter Hagoort. 2010. Coercion and Compositionality. *Journal of Cognitive Neuroscience*, 22(9):2131–2140.

Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Logical metonymy in a distributional model of sentence comprehension. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2021. Not all arguments are processed equally: A distributional model of argument complexity. *Language Resources and Evaluation*, 55(4):1–28.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Frederick G Gietz and Barend Beekhuizen. 2022. Remodelling complement coercion interpretation. *Proceedings of the Society for Computation in Linguistics 2022*, pages 158–170.

Daniele Godard and Jacques Jayez. 1993. Towards a proper treatment of coercion phenomena. In *EACL '93: Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, pages 168–177.

Yuling Gu. 2022. Measure more, question more: Experimental studies on transformer-based language models and complement coercion. *arXiv preprint arXiv:2212.10536*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *NAACL '01: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8.

Chan-Chia Hsu and Shu-Kai Hsieh. 2013. To Coerce or Not to Coerce: A Corpus-based Exploration of Some Complement Coercion Verbs in Chinese. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 13–20.

Seohyun Im and Chungmin Lee. 2024. What GPT-4 knows about aspectual coercion: Focused on "begin the book". In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 56–67.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Argyro Katsika, David Braze, Ashwini Deo, and Maria Mercedes Piñango. 2012. Complement Coercion: Distinguishing between type-shifting and pragmatic inferencing. *The Mental Lexicon*, 7(1):58–76.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860.

Maria Lapata and Alex Lascarides. 2003. A Probabilistic Account of Logical Metonymy. *Computational Linguistics*, 29(2):261–315.

Alex Lascarides and Ann Copestake. 1998. Pragmatics and word meaning. *Journal of linguistics*, 34(2):387–414.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Peng Liu, Lemei Zhang, Terje Farup, Even W Lauvrak, Jon Espen Ingvaldsen, Simen Eide, Jon Atle Gulla, and Zhirong Yang. 2024. NLEBench+NorGLM: A Comprehensive Empirical Analysis and Benchmark Dataset for Generative Language Models in Norwegian. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5543–5560.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge University Press.

Brian McElree, Matthew J Traxler, Martin J Pickering, Rachel E Seely, and Ray Jackendoff. 2001. Reading time evidence for enriched composition. *Cognition*, 78(1):B17–B25.

Stephen McGregor, Elisabetta Ježek, Matthew Purver, and Geraint Wiggins. 2017. A geometric method for detecting semantic coercion. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Maria Mercedes Piñango and Ashwini Deo. 2016. Reanalyzing the Complement Coercion Effect through a Generalized Lexical Semantics for Aspectual Verbs. *Journal of Semantics*, 33(2):359–408.

James Pustejovsky. 1991. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

James Pustejovsky and Pierrette Bouillon. 1995. Aspectual Coercion and Logical Polysemy. *Journal of Semantics*, 12(2):133–162.

Matteo Radaelli and Giosuè Baggio. 2025. Complement coercion with aspectual verbs is statistically infrequent in written norwegian. Forthcoming.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Giulia Rambelli, Emmanuele Chersoni, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2020. Comparing Probabilistic, Distributional and Transformer-based Models on Logical Metonymy Interpretation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 224–234.

Giulia Rambelli, Emmanuele Chersoni, Davide Testa, Philippe Blache, and Alessandro Lenci. 2024. Neural generative models and the parallel architecture of language: A critical review and outlook. *Topics in Cognitive Science*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Stefan Rüd and Alessandra Zarcone. 2011. Covert events and qualia structures for german verbs. In *Proceedings of the Metonymy 2011 Workshop*, pages 17–22.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – A Benchmark for Norwegian Language Models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633.

David Samuel, Vladislav Mikhailov, Erik Velldal, Lilja Øvrelid, Lucas Georges Gabriel Charpentier, and Andrey Kutuzov. 2025. Small Languages, Big Models: A Study of Continual Training on Languages of Norway. *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 573–608.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 373 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Ekaterina Shutova. 2009. Sense-based interpretation of logical metonymy using a statistical method. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 1–9.

Ekaterina Shutova, Jakub Kaplan, Simone Teufel, and Anna Korhonen. 2013. A computational model of logical metonymy. *ACM Transactions on Speech and Language Processing*, 10(3):1–28.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Alexandra Anna Spalek. 2015. The Influence of Context in Meaning: The Panorama of Complement Coercion. In *Modeling and Using Context*, pages 526–531. Springer International Publishing.

Alexandra Anna Spalek and Kjell Johan Sæbø. 2019. To Finish in German and Mainland Scandinavian: Telicity and Incrementality. *Journal of Semantics*, 36(2):349–375.

Josefien Sweep. 2012. Logical Metonymy in Dutch and German: Equivalents of Begin, Finish, and Enjoy. *International Journal of Lexicography*, 25(2):117–151.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*.

Matthew J Traxler, Martin J Pickering, and Brian McElree. 2002. Coercion in sentence processing: Evidence from eye-movements and self-paced reading. *Journal of Memory and Language*, 47(4):530–547.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press.

Cornelia Maria Verspoor. 1997. Conventionality-governed logical metonymy. In *Proceedings of the Second International Workshop on Computational Semantics*, pages 300–312. Citeseer.

Bingyang Ye, Jingxuan Tu, Elisabetta Jezek, and James Pustejovsky. 2022. Interpreting logical metonymy through dense paraphrasing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Alessandra Zarcone, Alessandro Lenci, Sebastian Padó, and Jason Utt. 2013. Fitting, not clashing! A distributional semantic model of logical metonymy. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pages 404–410.

Alessandra Zarcone and Sebastian Padó. 2011. Generalized Event Knowledge in Logical Metonymy Resolution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Alessandra Zarcone, Sebastian Padó, and Alessandro Lenci. 2014. Logical Metonymy Resolution in a Words-as-Cues Framework: Evidence From Self-Paced Reading and Probe Recognition. *Cognitive Science*, 38(5):973–996.

Alessandra Zarcone, Jason Utt, and Sebastian Padó. 2012. Modeling covert event retrieval in logical metonymy: probabilistic and distributional accounts. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 70–79.