Gombrowicz

Are fronted sentence elements a hallmark of Gombrowicz' Translatlantyk?

One typical feature of Gombrowicz' style is the length of *sentence front*, as we call the sequence from the beginning of a sentence element (subject, object, or adverbial) to the finite verb of its main predicate. This feature is supposed to be particularly prominent in his novel *Transatlantyk* and its Swedish translation.

For either language, we compare documents of different genres considering their distribution of fronted subjects, objects, indirect objects, prepositional noun groups, adverbs and subordinate adverbial clauses and their distance from the main predicate.

We use the Polish and the Swedish part of InterCorp version 13ud, which has been automatically labeled with Universal Dependencies. This labeling enables an extraction of the given sentence elements. At the same time, we use the genre classification provided in the corpus metadata.

Technical limitations

We consider only documents with at least 50 sentences.

We discarded a few documents where the number of subjects exceeded the number of sentences - due to evident parsing errors. All these documents were discussion transcripts from EUROPARL.

KWICs (KeyWords In Context)

We retrieved KWICs matching sequences starting with a token labeled as the sentence element in question (e.g. adverbial), followed by a deliberate number of tokens and ending with a token labeled as the sentence root, all within one sentence. We considered only sentences where the root was represented by a finite verb form. Note that, in the UD notation, this excludes copula predicates!

This is the query for subject:

```
1:[deprel = "nsubj"] []* 2:[deprel = "root" & upos = "(VERB|AUX)" &
feats="VerbForm=Fin"] & 1.head = 2.id within <s/>
```

Retrieved statistics

We retrieved statistics for each combination of language and sentence element separately, querying InterCorp with the KonText API. The KonText API returns useful metadata and statistics along with the KWICs. From the information provided by the Kontext API, we used `text_id`, `txttype`, and KWIC length. We also retrieved the number of sentences in each text with a separate query and merged them with the query results.

As a next step, we aggregated this data for each row to represent one text with the following columns:

- `text_id`
- `txttype`: genre (fiction, nonfiction, children's reads, religious, discussion transcripts)
- `original`: is this text a language original?
- `count` of the sentences matching the query
- `sent_count`: count of all sentences in the text
- `count_pmsent`: relative incidence of the sentences matching the query per (a hypothetical) one million sentences in the same text, henceforth 'incidence'.
- aggregates of the KWIC lengths (distances between the focused sentence element and the finite main predicate)
  - `mean_klen` mean
  - `sd_klen` standard deviation
  - `deprel` sentence element focused in the query

Analysis

We compare the incidence of the individual sentence elements in the fronted position in Gombrowicz' Transatlantyk to their incidence in other documents in the given language subcorpus of InterCorp, grouped by text type.

First we consider the plain incidence, visualizing it as boxplots of each text type and a horizontal line with the incidence value of Transatlantyk. Then we examine the distance between the given fronted sentence element and the finite main predicate in each text compared to its incidence, using a scatterplot with point ranges.

How to interpret the boxplots

The Y-axis represents the measured value, here the incidence of the given fronted element. The boxes with whiskers represent the observation points. The thick line inside each box represents the median value. One half of the observed values is equal or lower than median and the other half is higher than median. The lower box edge represents a value at or below which we observe 25% of all observed values. Accordingly, the upper box edge represents a value which accommodates 75% of all observed values. Hence, the box captures one half of the values. Note that the line does not have to be centered. When it is for instance closer to the upper edge, this means that there are more observations with values between the first quartile and the median than there are values between the median and the third quartile. The whiskers reaching out of the boxes denote intervals of values below the first quartile and above the third quartile that are still populated densely enough not to be called extreme values. Extreme values (outliers) are represented by points beyond the whiskers. The length of the whiskers is computed as the lower box edge minus 1.5 * the box height and the upper box edge plus 1.5 * the box height.

How to interpret the scatterplots

A scatterplot represents the observation points by two variables, here the incidence of the fronted sentence element in the given text (X-axis) and the mean of all distances between the fronted sentence elements and the finite main predicates in each text (Y-axis). This plot also displays the standard deviations of these values as bars. The standard deviation is a measure of dispersion; that is, how the data points typically differ from the mean value. Gombrowicz' Transatlantyk is marked with a different color.

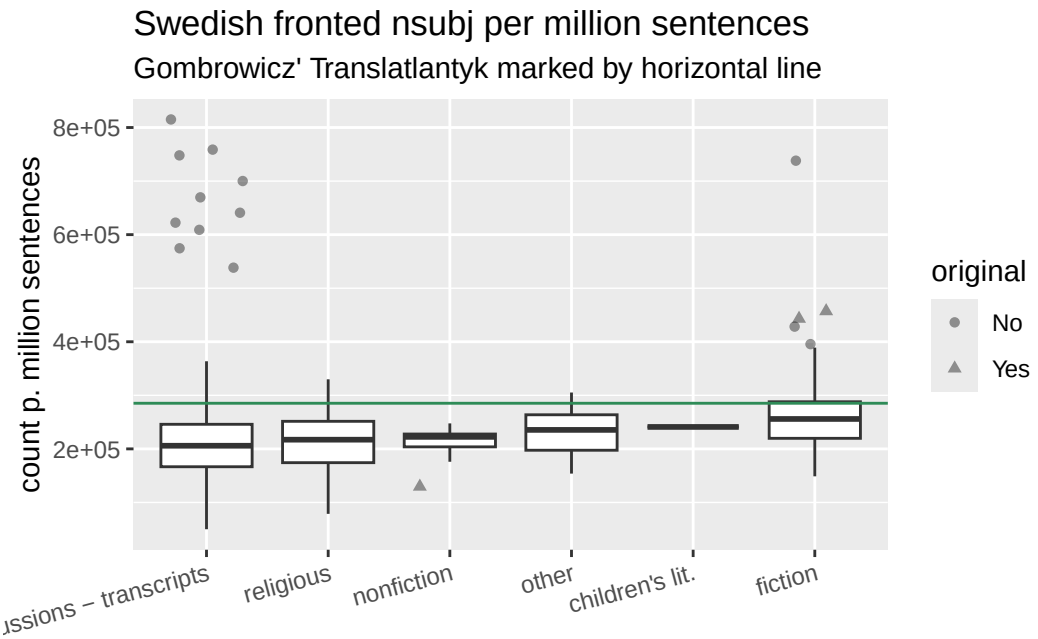
Fronted Swedish subject (`nsubj`)

Incidence of fronted Swedish nominal subject (nsubj) per million sentences

In Swedish, the fronted position of the nominal subject is stylistically unmarked. At the same time, the subject cannot be dropped. The differences between texts are caused by the restrictions in our query, which excludes copula predicates and (probably less importantly) subjects formed by infinitives or subordinate finite clauses. The reasons are purely technical ones; one has to do with the sentence segmentation and the other one with the searchability of Universal Dependencies with the Corpus Query Language:

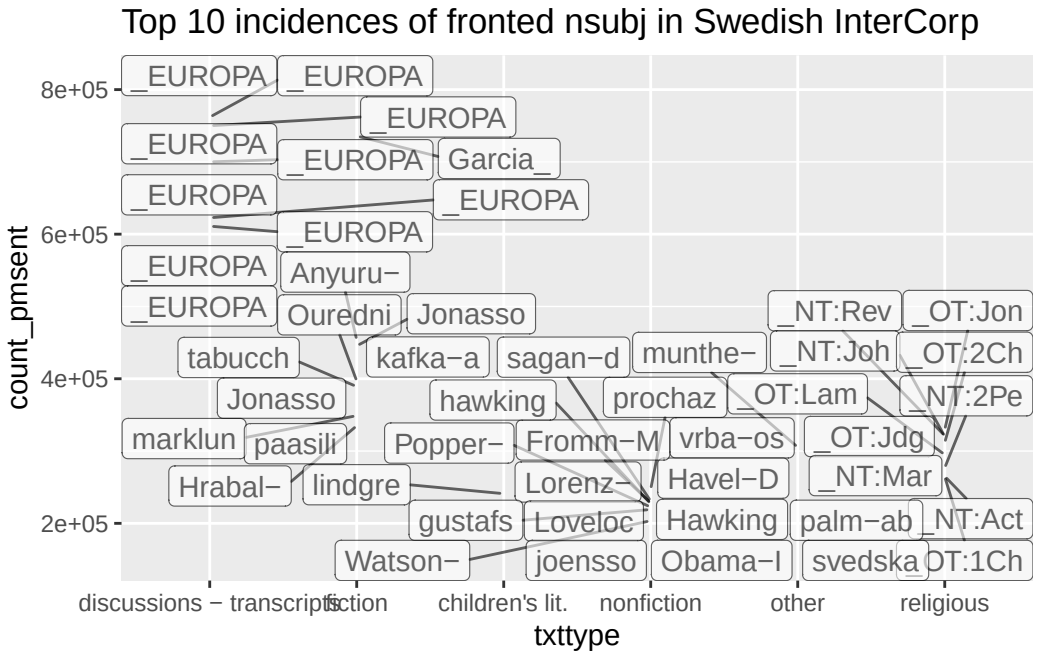
1. The sentences are not only segmented at full stops, but also at semicolons and colons. Subjects in clauses following these punctuation marks can be elided if they are expressed in the preceding clauses of the same sentence.

2. In the Universal Dependencies formalism, the actual roots of copula predicates are the predicate nouns or adjectives and the finite copula verb is located deeper in the subtree. It is certainly still possible to formulate a query in CQL that would capture the relation between the sentence element and this finite verb, but running it on a whole InterCorp language subcorpus exceeds the computational capacity of the API server.



The median incidence of fronted subjects is almost identical across all text types. At the same time, most documents with higher incidence occur in fiction. The incidence of fronted subjects in Swedish translation of Gombrowicz' Transatlantyk is roughly at the third quartile of fiction; that means that Transatlantyk is among the 25% of fiction texts with the highest incidence of fronted subjects.

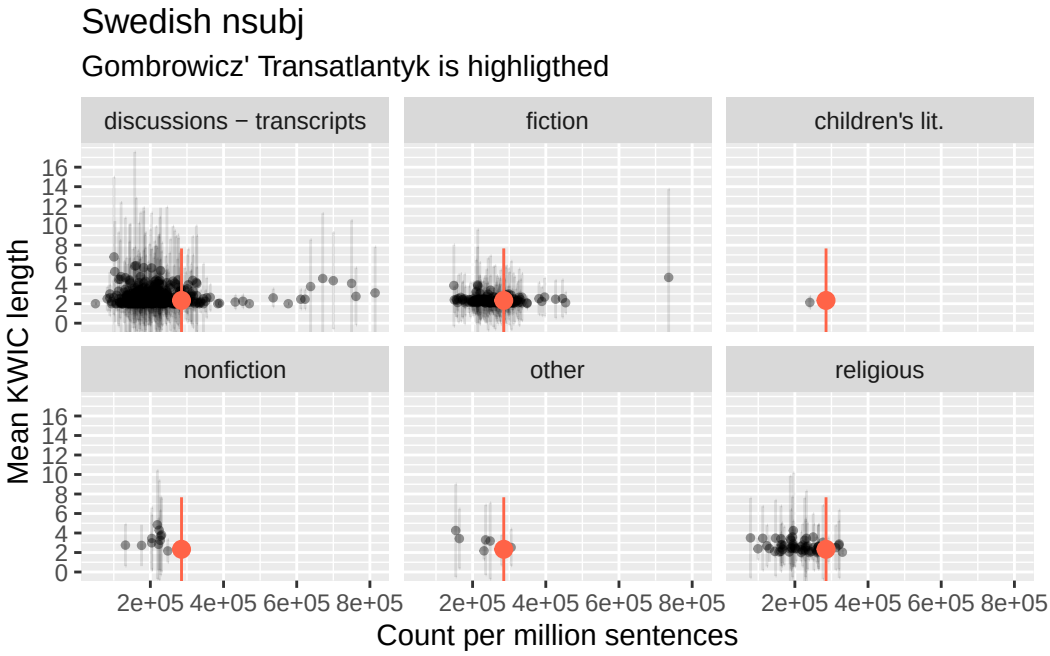
These would be the top five outliers for each text type:



Relative incidence of fronted subject in Swedish vs. its distance from the finite predicate

The distance from the mean predicate is aggregated as mean and standard deviation of cases observed within each text.

The distance of subjects from the finite predicate in Gombrowicz' Transatlantyk is not exceptional either in terms of its mean or standard deviation.

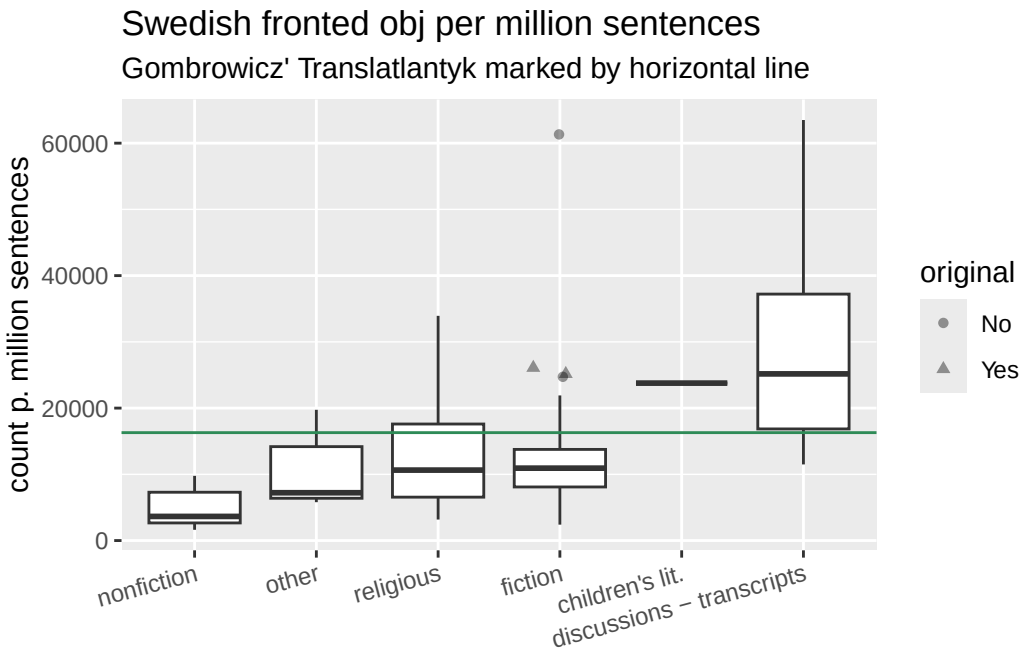


Fronted Swedish direct object (`obj`)

Incidence of fronted Swedish direct object (obj) per million sentences

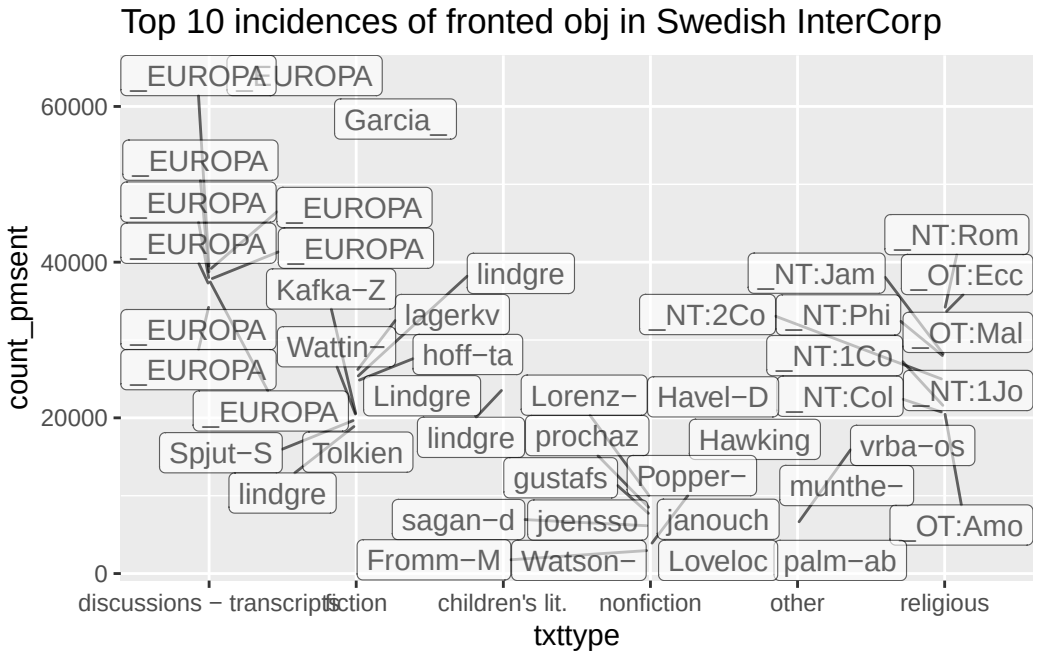
Unlike subject, fronted direct object is stylistically marked, and hence it is far less common: the top incidence values were reaching 800,000 per million, whereas the maximum incidence of object observed in this data set is thirteen times lower: around 60,000 p.m. Besides, we observe differences in the median values of fronted objects between most text types.

Fiction and religious texts (parts of the Old and New Testament) have a similar profile, just with less variation in fiction. Gombrowicz' *Transatlantyk* (16,297 fronted objects per million sentences) lies again in the fourth quartile in fiction, but is not an outlier.



These are the top fiction, "other", and "religious" texts with more fronted subjects than Translatlantyk: G. García-Márquez' *The Autumn of the Patriarch* (61,321), diverse chapters from the Old and the New Testament (33,929 - 20,027), P. Lagerkvist: *Barabbas* (25,954), B. Hoff: *The Tao of Pooh*, A. Lindgren: *Mardie* (25,146) and *Bullerbyn* (23,770). Above 20,000 are also T. Lindgren (*The Way of a Serpent*) and F. Kafka: *The Castle*.

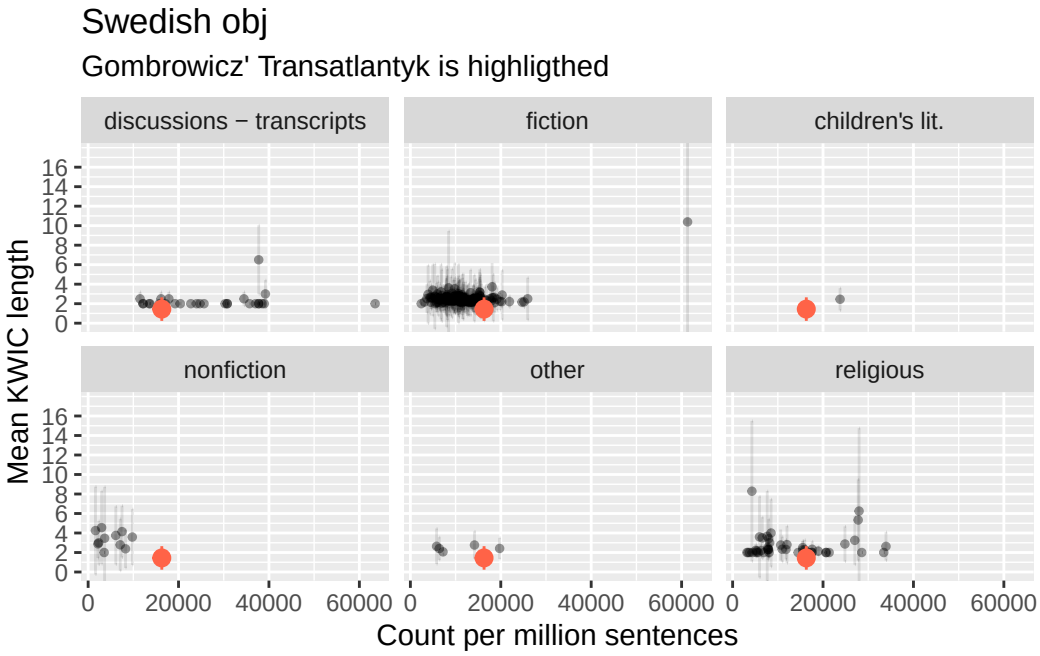
In the realm of nonfiction, sentence fronts loaded with direct objects prominently appear in the Swedish translations of the following authors: K. Lorenz (ethology), K. Popper (philosophy), S. Hawking (physics), C. Sagan (astronomy), E. Fromm (psychology), and J. Lovelock (environmental studies). We can also observe object-heavy sentence fronts in V. Havel's dramas, A. Munthe's *Book of San Michele* spiritual autobiography as well as in the literary testimony of the Auschwitz refugee Rudolf Vrba, the co-author of the famous Vrba-Wetzler report



Relative incidence of fronted object in Swedish vs. its distance from the finite predicate

The distance from the mean predicate is aggregated as mean and standard deviation of cases observed within each text.

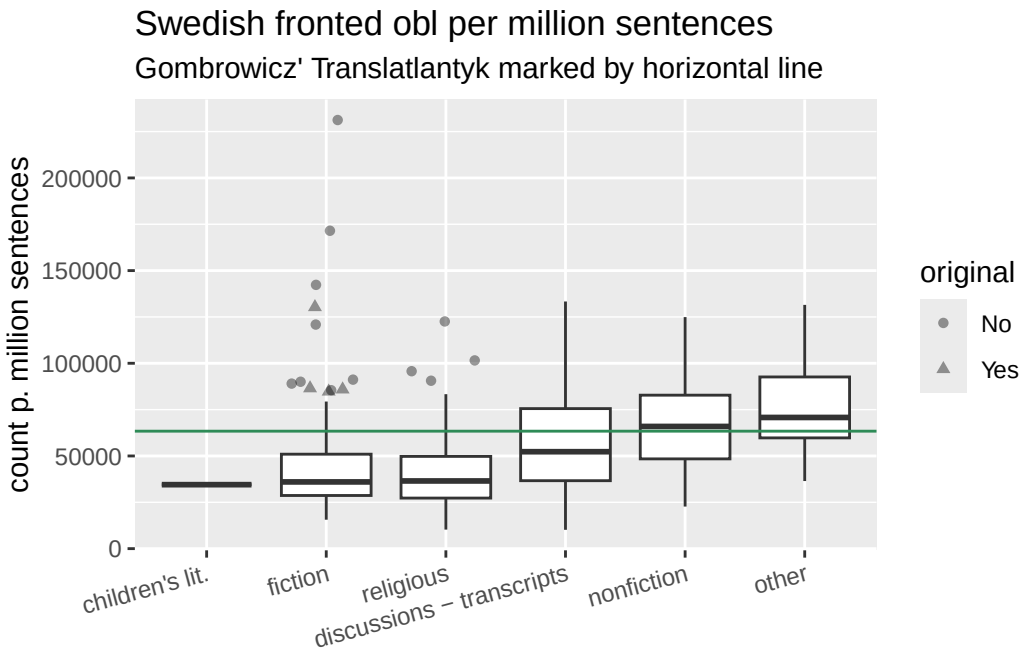
The distance of the fronted object from the mean predicate is typically between 0 and 2 tokens in *Translatlantyk* In general, other fiction texts vary more.



Fronted noun in oblique case (`obl`) in Swedish

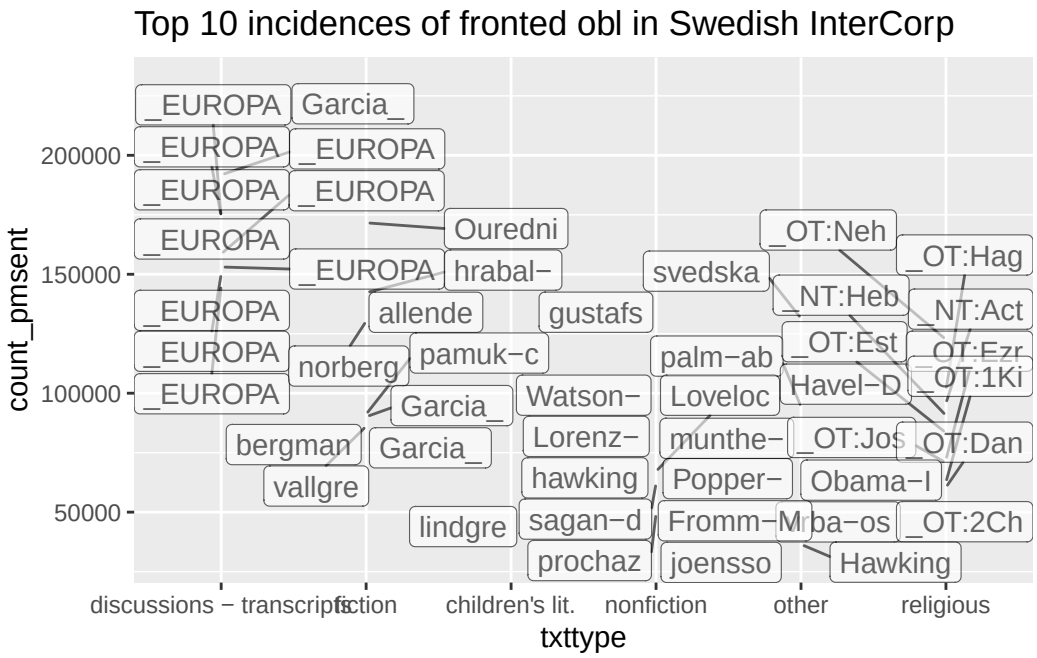
Incidence of fronted Swedish noun in oblique case (`obl`) per million sentences

The Swedish InterCorp dataset suggests that the fronted noun in oblique case (prepositional case) is typical of nonfiction and "other" texts rather than for fiction and religious literature. Again, the Swedish translation of Gombrowicz' *Transatlantyk* tends to use them abundantly (63,377 occurrences per million sentences), especially compared to other fiction and religious texts. In this respect, *Translatlantyk* is most similar to an average (median) non-fiction text.

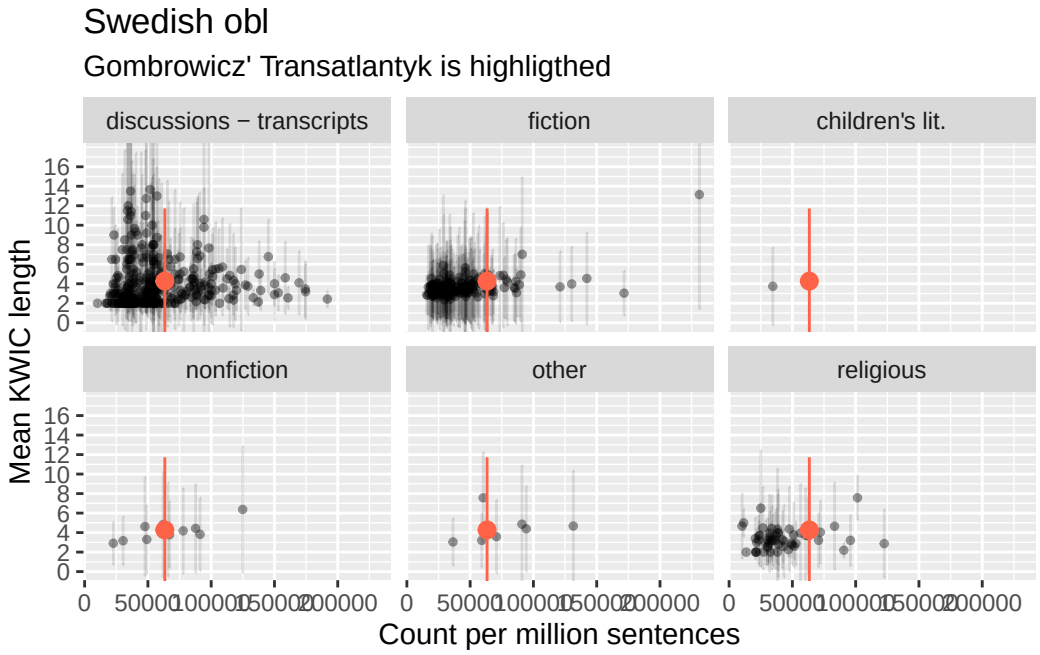


In the Swedish InterCorp sample, the generous use of fronted noun in oblique case is typical of G. García-Márquez (3 texts among the top ten per text type), the Czech author Hrabal and the French-Czech author Ouředník, Hrabal's prominent translator to French (*Europeana*, apparently inspired by Hrabal's distinct style), O. Pamuk (*The Black Book*), I. Bergman (*Laterna Magica*) , and I. Allende's memoirs (*Paula*).

In the non-fiction and "other" text types, the top documents are J. Norberg's *Progress: Ten Reasons to Look Forward to the Future* (misclassified as fiction)

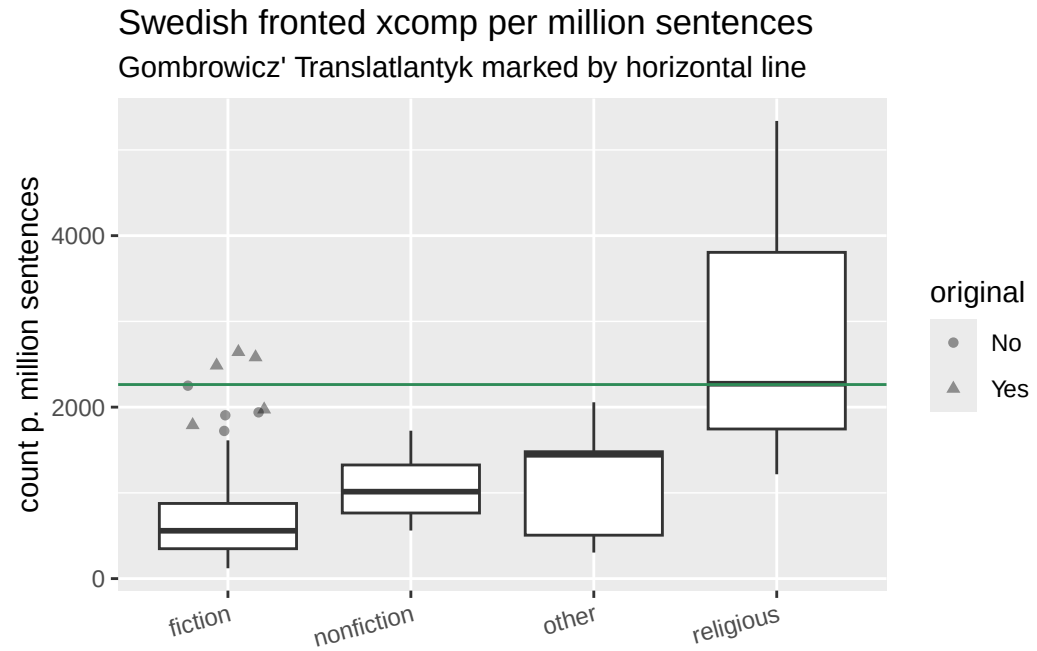


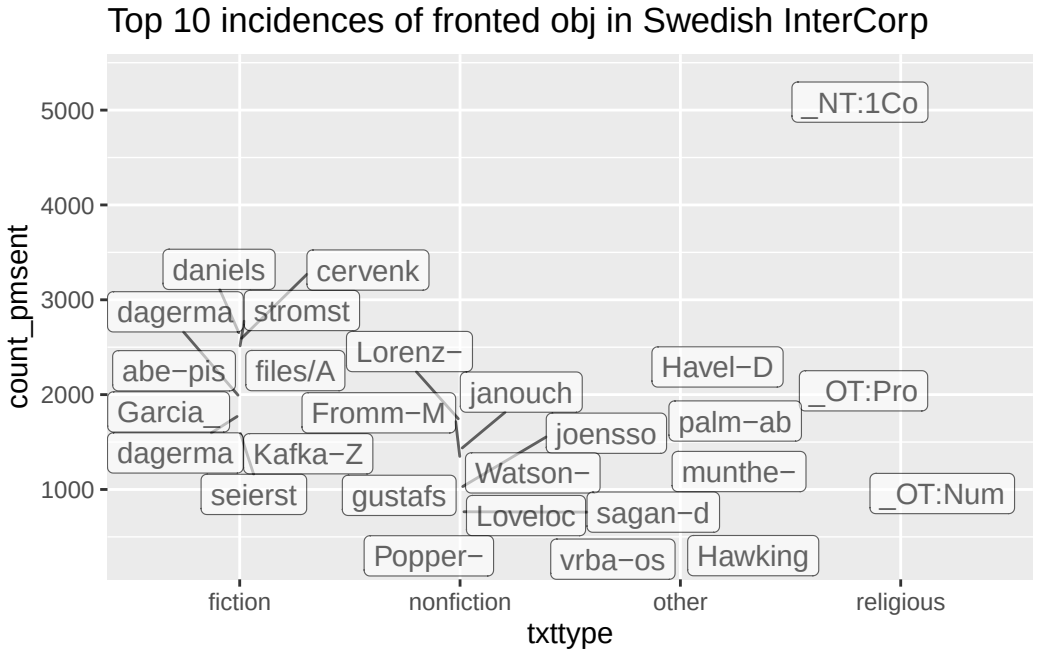
Relative incidence of fronted noun in oblique case in Swedish vs. its distance from the finite predicate



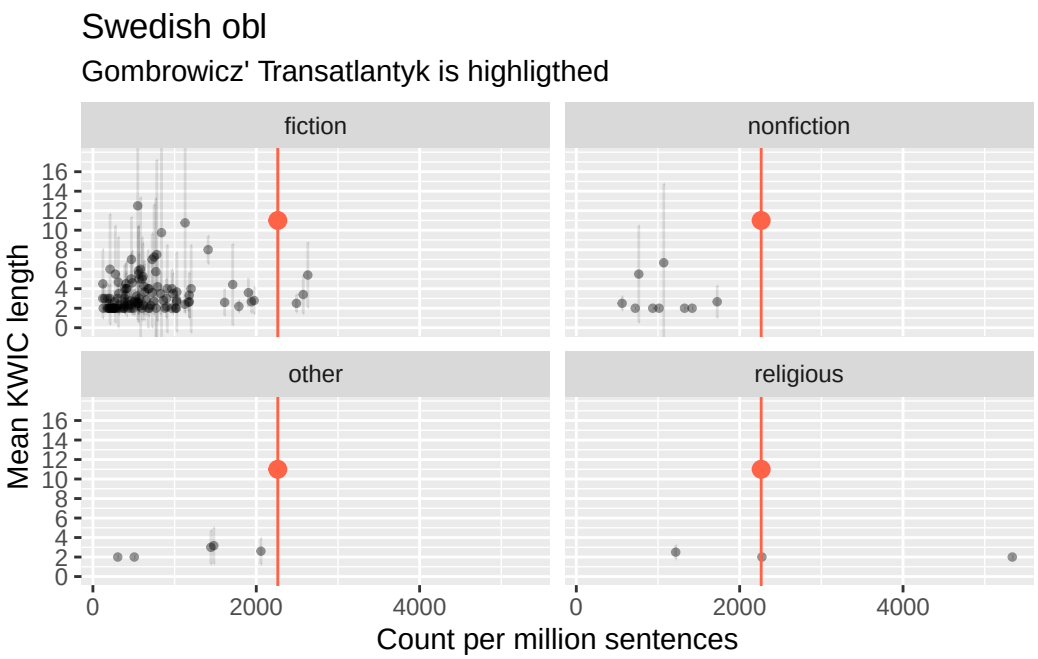
Fronted infinitive/participial clause (`xcomp`) in Swedish

Incidence of fronted Swedish infinitive/participial clause (`xcomp`) per million sentences



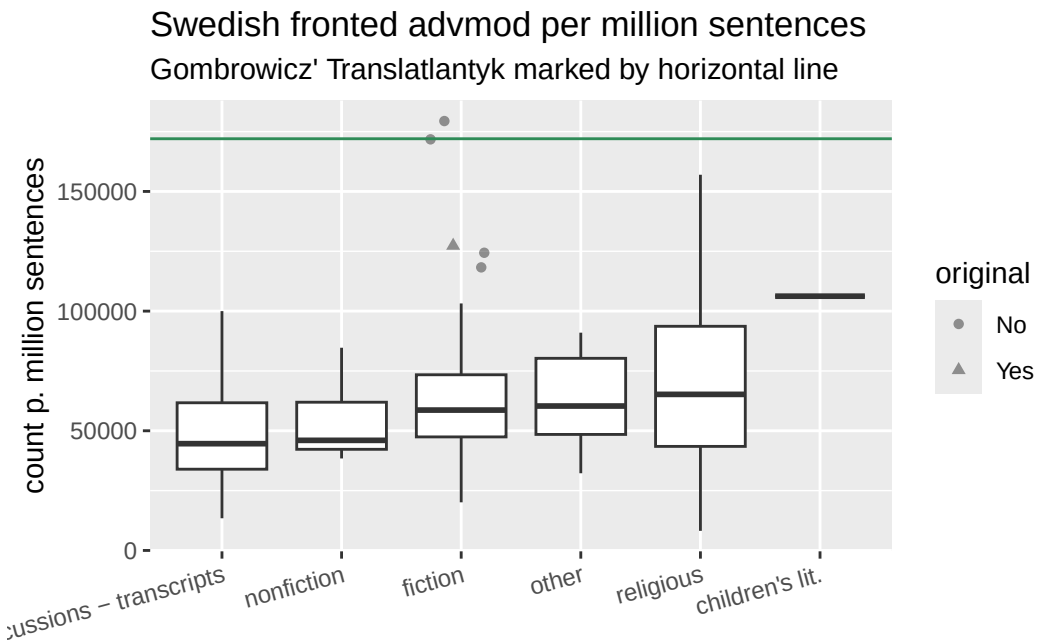


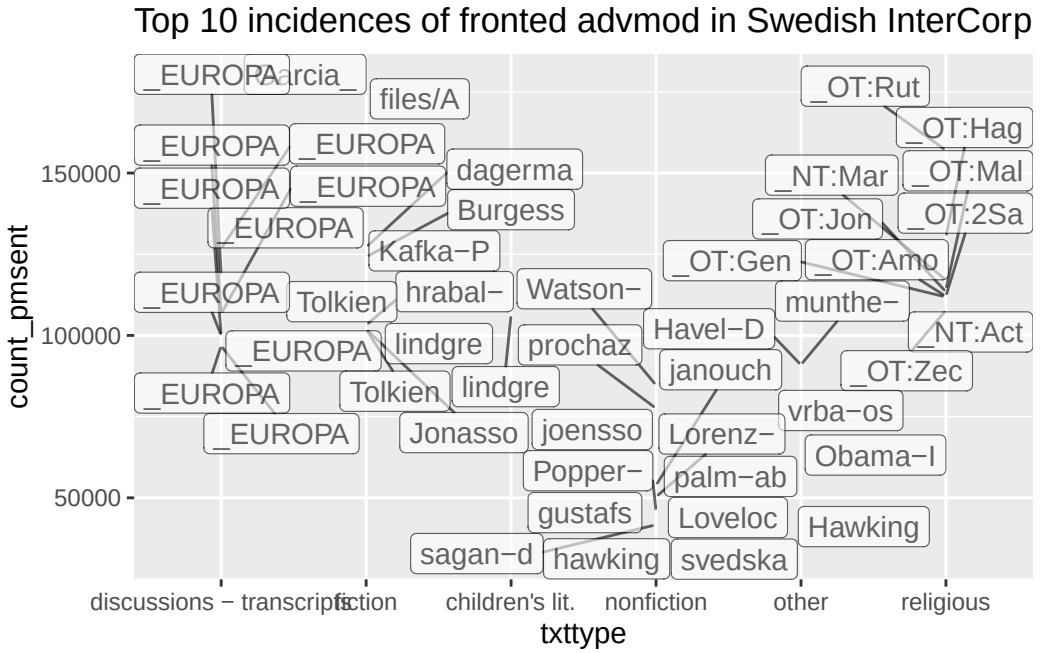
Relative incidence of fronted infinitive or participle clause (`xcomp`) in Swedish vs. its distance from the finite predicate



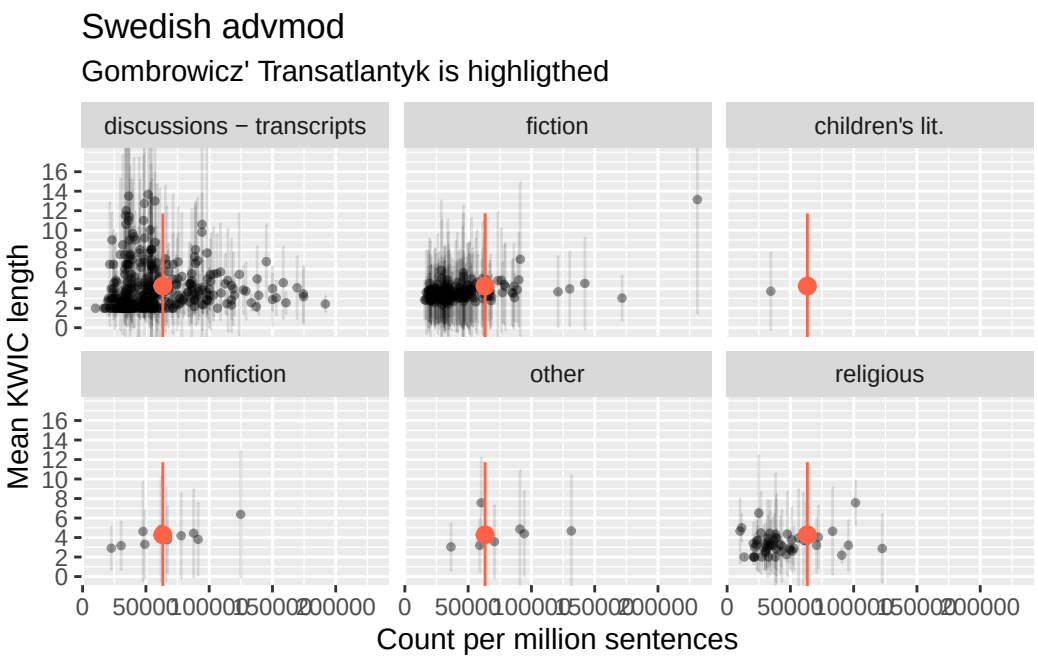
Fronted adverbial (adverb, `advmod`) in Swedish

Incidence of fronted Swedish adverbial (adverb, `advmod`) per million sentences



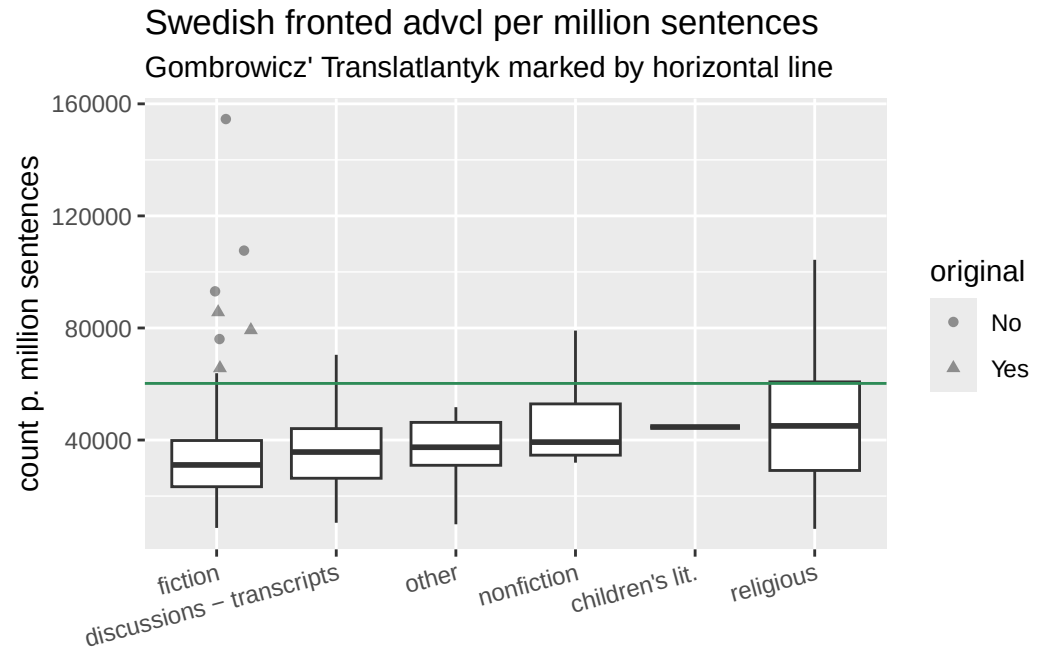


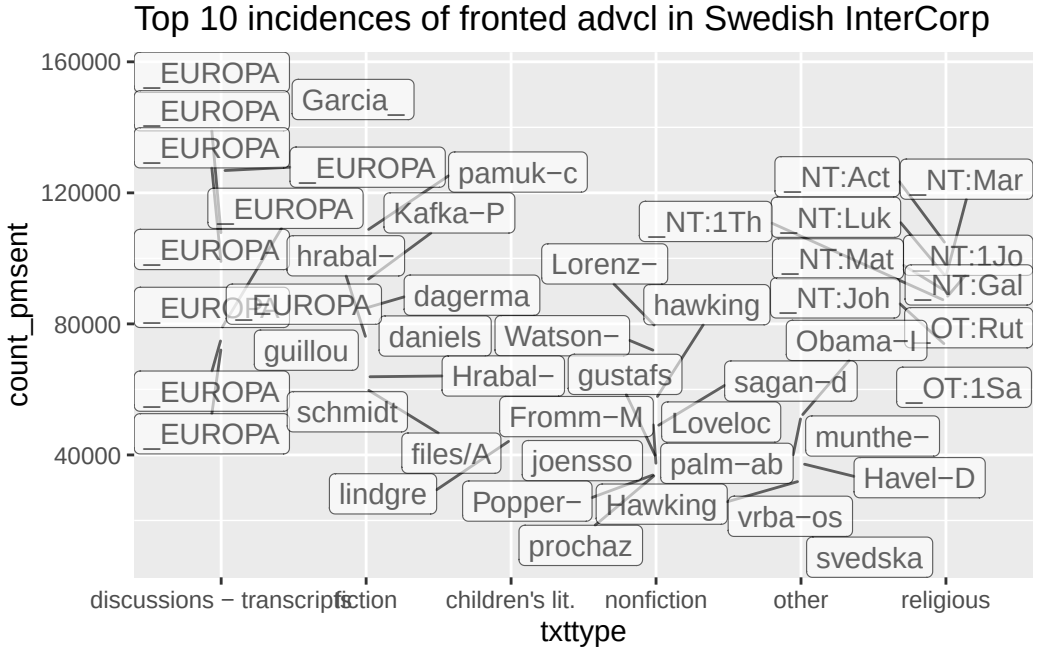
Relative incidence of fronted adverbial (adverb, `advmod`) in Swedish vs. its distance from the finite predicate



Fronted adverbial clause (`advcl`) in Swedish

Incidence of fronted Swedish adverbial clause (`advcl`) per million sentences





Relative incidence of fronted adverbial clause (`advcl`) in Swedish vs. its distance from the finite predicate