Macrothink Institute™

**International Journal of Linguistics**
ISSN 1948-5425
2026, Vol. 18, No. 2

# Event and Agent Nominalisations Across Academic, Spoken, and Fiction Registers in Contemporary English: A Pilot Corpus Study of Clausal Density

Nickolas Komninos

Department of Languages and Literatures, Communication, Education and Society (DILL)

University of Udine, Italy

Palazzo Antonini, Via Petracco n. 8, Udine, Italy

Received: February 5, 2026 Accepted: March 9, 2026 Published: April 18, 2026

doi:10.5296/ijl.v18i2.23561 URL: https://doi.org/10.5296/ijl.v18i2.23561

**Abstract**

Nominalisation is a central device for packaging clausal information into noun phrases, thereby increasing clausal density and enabling the abstraction typical of academic prose. This article reports a pilot corpus study of nominalisation types across four English register samples drawn from the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC): Academic, Spoken, and Fiction. Using formal diagnostics from generative work on nominalisations (Picallo 1991; Baker & Vinokurova 2009) as an annotation guide, but adopting a corpus-register perspective for explanation (Biber et al. 1999; Nichols 1989), 500 tokens were manually classified as Event nominalisations, Agent nominalisations, or complex nominal constructions. The results show a robust register contrast within this stratified sample: Academic writing overwhelmingly favours eventive nominalisations (about 80%), while spoken and fiction samples favour agentive nominals and show more surface noun-phrase complexity. A chi-square test confirms that the association between register group (academic vs non-academic) and nominal type is large in the sample. The discussion interprets these patterns as register-conditioned preferences for conceptual reification and agent suppression, while also foregrounding the theoretical and methodological limits of using corpus distributions as evidence for formal grammatical architecture.

**Keywords:** Nominalisation, Event nominal, Agent nominal, Register variation, Corpus linguistics, Clausal density, Noun phrase complexity

## 1. Introduction

Nominalisation has long been recognised as a major resource for creating dense informational

style in writing. By converting predicates and relations into nominal expressions (e.g., analyse → analysis), speakers and writers can repackage clausal content as a noun phrase that can be modified, quantified, and embedded as an argument. This repackaging is closely connected to complex-sentence phenomena: rather than adding overt subordinate clauses, writers can compress propositional material into nominal structures and thereby increase clausal density.

The present article explores a specific question within this broad descriptive generalisation. Nominalisations are not uniform: some nominals primarily denote events or processes (e.g., the destruction of the particle), while others denote participants or agents (e.g., driver, researcher). Distinguishing these types matters because they differ in argument-structure potential and in how they support depersonalised, abstract exposition. The central claim investigated here is that academic discourse does not merely use 'more nominalisation' overall; rather, it preferentially selects nominal types that package events and processes and that support argument realisation inside the noun phrase.

Empirically, the article presents a small, manually annotated corpus sample (N = 500 tokens) drawn from two large English corpora, COCA and the BNC. The sample is balanced across Academic and non-academic registers (Spoken and Fiction). The study is explicitly pilot in nature: it is designed to test the feasibility of mapping theoretically motivated categories such as Event vs Agent nominalisations onto naturally occurring corpus tokens, and to provide an initial estimate of how strongly register conditions the distribution of these categories.

The remainder of the article is organised as follows. Section 2 introduces the background literature and, crucially, clarifies the theoretical position adopted here: formal generative analyses are used as diagnostic tools for categorisation, while register-functional work motivates the interpretation of distributional tendencies. Section 3 describes the corpora, sampling strategy, and annotation scheme. Section 4 reports the results, and Section 5 discusses their implications and limitations.

## 2. Background and Theoretical Positioning

### *2.1 Formal Diagnostics for Nominalisation Type*

Formal work on nominalisations has emphasised that nominal structures can preserve elements of clausal architecture. Within the DP hypothesis, nominals are analysed as functional projections headed by D, potentially containing internal functional structure that licenses arguments and modification (Picallo 1991). A key distinction for the present study is between process/eventive nominals and result/referential nominals: process nominals retain verbal argument-structure properties and readily combine with complements (often realised as of-phrases in English), whereas result nominals behave more like ordinary count nouns and are less argument-bearing.

Baker and Vinokurova (2009) develop a related cross-linguistic dichotomy between agent nominalisations and event nominalisations. Agent nominalisations typically denote a participant associated with an eventuality (e.g., -er nouns such as driver), are strongly referential, and show limited inheritance of higher verbal functional structure. Event nominalisations, by contrast, denote the event or process itself (e.g., many -tion/-ment

nominals and gerundive -ing forms), and they can license internal arguments and aspectually relevant modification. In Baker and Vinokurova's account, differences between agent and event nominalisations follow from the structural position where the nominaliser attaches in the verbal spine (the 'Bare VP constraint' is one formal device used to capture this contrast).

*2.2 Register-Functional Motivations for Nominal Style*

Register and genre studies provide a complementary perspective: rather than asking which structures are licensed by the grammar, they ask which structures are selected under particular communicative demands. Biber et al. (1999) show that registers differ systematically in their grammatical profiles, with informational prose characterised by dense noun phrases and relatively reduced clausal elaboration. From this viewpoint, nominalisation is not merely a derivational option but a stylistic resource that facilitates information packaging.

Nichols (1989) links nominalisation in scientific prose to depersonalised assertion: turning actions into nominal entities helps writers foreground processes and concepts rather than agents. In English academic writing, this motivation predicts a preference for eventive/process nominalisations that support argument realisation inside the noun phrase (e.g., the evaluation of the proposed mechanisms), as opposed to agentive nominalisations that foreground human participants (e.g., researcher, speaker).

*2.3 Theoretical Positioning and Evidential Scope*

The present study draws on both strands of work, but it does not treat them as interchangeable theoretical frameworks. Formal generative analyses of nominalisation are typically developed as competence-oriented accounts and are often sceptical of treating corpus frequency patterns as direct evidence for grammatical architecture. Conversely, corpus and functional approaches are explicitly usage-oriented and treat distributional tendencies as central evidence. Because these traditions can embody different evidential standards, an analysis that simply combines them without clarification risks becoming incoherent.

To avoid this problem, the study adopts a division of labour. The explanatory frame is register-functional: differences across Academic, Spoken, and Fiction samples are interpreted as register-conditioned preferences for particular information-packaging strategies. Formal analyses (Picallo 1991; Baker & Vinokurova 2009) are used primarily as a source of diagnostics for annotation - that is, as a principled way to distinguish eventive from agentive nominals and to identify argument-structure cues in naturally occurring tokens. In this role, formal theory informs operational definitions, but the pilot does not claim that frequency differences by themselves confirm or falsify a particular derivational analysis.

A further methodological motivation follows from the nature of the categories involved. 'Event nominal' and 'agent nominal' function here as comparative concepts in Haspelmath's (2010) sense: they are theoretically motivated abstractions that must be mapped onto corpus tokens, which often display indeterminacy and gradience. As Spike (2020) notes for indeterminate categories more generally, empirical classification often requires principled decision rules and an acknowledgment of borderline cases. For that reason, the study relies on a manually annotated stratified sample rather than attempting exhaustive corpus counts.

## 3. Data and Method

### *3.1 Corpora and Sampling Design*

The study uses register-labelled data from two widely used corpora: the Corpus of Contemporary American English (COCA; Davies 2008-) and the British National Corpus (BNC; BNC Consortium 2007). The focus is on contrasting an academic register with non-academic registers that differ in modality and communicative goals. Two academic samples were drawn (COCA Academic and BNC Academic) to check whether the predicted nominal style is consistent across the two corpora, while the non-academic comparison is represented by a spoken conversational sample (COCA Spoken) and a narrative written sample (BNC Fiction).

Because the pilot relies on manual annotation, the dataset is a stratified sample rather than an exhaustive extraction. A total of 500 tokens were collected, balanced across the four register samples (COCA Academic: 120; COCA Spoken: 120; BNC Academic: 130; BNC Fiction: 130). The balanced design supports direct comparison of proportions within the sample, but it does not license claims about absolute frequencies per million words in the underlying corpora.

### *3.2 Extraction Procedure*

Token extraction combined morphological and syntactic heuristics intended to capture frequent nominalisation strategies in English. Searches targeted common nominalising suffix families (-tion/-sion, -ment, -ing, -er/-or) and were constrained to nominal parts of speech using corpus interface tags (e.g., NN*). In addition, the extraction procedure deliberately sampled complex nominal constructions with heavy internal modification (e.g., N of N patterns and multiword noun sequences), since such constructions often function as alternatives to clausal elaboration.

The use of heuristic queries is a practical compromise. It makes manual annotation feasible but introduces biases toward frequent and easily retrievable forms. For this reason, all quantitative results are reported as distributional tendencies within the extracted sample rather than as population estimates for the corpora.

### *3.3 Annotation Scheme and Decision Rules*

Each token was assigned to one of three mutually exclusive categories:

(i) Event nominalisation (EN): a nominal that denotes an eventuality (process, action, or event) and shows event-oriented behaviour (e.g., compatibility with internal arguments or event-related modification).

(ii) Agent nominalisation (AN): a nominal that denotes a participant associated with an eventuality (typically human or animate), behaving like a referential count noun and not primarily denoting the event itself.

(iii) Complex nominal construction (CNC): a nominal expression whose complexity is realised primarily through multiword noun phrase structure (e.g., NN sequences, N of N, or heavily modified noun phrases) and which was not coded in this pilot for event vs agent type.

The event/agent distinction was guided by diagnostics from Picallo (1991) and Baker and Vinokurova (2009), including argument-realisation possibilities and the semantic contribution of nominalising morphology. Ambiguous cases (e.g., building as process vs result) were resolved by reference to local syntactic context (e.g., presence of complements, determiners, and modifiers). A small proportion of borderline cases remained; these were conservatively assigned to CNC when neither a clear eventive reading nor a clear agentive reading was supported by the local context.

Annotation was conducted by a single coder. Given the pilot scope, the study does not report formal inter-annotator agreement. The Discussion section therefore treats the category system as an initial operationalisation and identifies annotation reliability as a priority for follow-up work.

### *3.4 Quantitative Analysis*

The analysis proceeds in two steps. First, descriptive distributions are reported as counts and proportions by register. Second, the association between register group (Academic vs non-academic) and nominal category is assessed using a chi-square test of independence on the contingency table derived from the sample. Because the sample is not a random draw from each corpus, the inferential statistics are interpreted as exploratory indicators of the strength of association in the sample rather than as definitive population-level tests.

## 4. Results

Table 1 summarises the distribution of nominal categories across the four register samples. Because the dataset is balanced by design, the proportions are directly comparable within the sample.

Table 1. Distribution of nominal categories by register sample (counts and within-sample proportions).

| Register sample | N (tokens) | Event nominalisations (EN) | Agent nominalisations (AN) | Complex nominal constructions (CNC) |
|---|---|---|---|---|
| COCA Academic | 120 | 95 (79.2%) | 10 (8.3%) | 15 (12.5%) |
| BNC Academic | 130 | 105 (80.8%) | 15 (11.5%) | 10 (7.7%) |
| COCA Spoken | 120 | 40 (33.3%) | 50 (41.7%) | 30 (25.0%) |
| BNC Fiction | 130 | 45 (34.6%) | 60 (46.2%) | 25 (19.2%) |

Two patterns stand out. First, the two Academic samples converge closely: eventive nominalisations constitute about four-fifths of the extracted tokens in both COCA Academic (79.2%) and BNC Academic (80.8%). Second, both non-academic samples show the opposite profile: agent nominalisations constitute the largest category in COCA Spoken (41.7%) and BNC Fiction (46.2%), while eventive nominalisations account for only about one third.

To provide an exploratory measure of the strength of association between register and nominal category, the Academic samples were pooled and compared to the pooled non-academic samples. A chi-square test of independence on the resulting 2 x 3 contingency table is significant (chi-square = 111.17, df = 2, $p < .001$), with a large effect size (Cramer's V = 0.47). A four-register test likewise indicates a strong association in the sample (chi-square = 114.12, df = 6, $p < .001$; Cramer's V = 0.34).

Qualitatively, the Academic samples contain many eventive nominals with internal-argument realisation (typically as of-phrases), such as the measurement of kinetic energy and the application of these principles. By contrast, the non-academic samples contain many referential agentive nouns that designate human or institutional actors (e.g., speaker, owner, management), consistent with the participant-centred focus of conversation and narrative. Complex nominal constructions are especially common in the Spoken sample, where heavy noun phrase modification appears to function as a surface strategy for packaging information without the more abstract event-nominal style typical of academic prose.

## 5. Discussion

### *5.1 Register-Conditioned Preferences and Clausal Density*

The pilot results support the hypothesis that academic discourse favours eventive nominalisations. Within the sample, academic writing shows a near categorical preference for nominals that denote processes and events and that readily support internal modification and complementation. From a register perspective, these properties facilitate concept formation and information packaging: an event nominal can function as a discourse entity that can be referenced, modified, and linked to subsequent argumentation (Biber et al. 1999).

At the same time, eventive nominalisation aligns with the depersonalising tendency of scientific prose. Transforming a clause such as researchers analysed the data into a nominal expression (the analysis of the data) both reduces the prominence of the agent and foregrounds the process. Nichols (1989) notes that this type of nominal style supports impersonal assertion; the present distributional findings are consistent with this functional motivation, insofar as academic samples in the pilot show relatively few agent-denoting nominals.

### *5.2 Methodological and Theoretical Limitations*

Several limitations temper the interpretation. First, the dataset is a stratified sample collected via heuristic queries rather than a random draw; accordingly, the proportions reported here should be treated as estimates of tendencies within the extracted sample. Second, the study draws from two corpora that differ in variety, period, and sampling design (COCA vs BNC). The close convergence of the two Academic samples is suggestive, but a publication-scale

study would ideally control for corpus as a factor or rely on a single corpus with multiple registers.

Third, annotation faces genuine indeterminacy. Many nominal forms permit both eventive and referential interpretations, and corpus tokens can be underspecified for the intended reading. This is not a flaw of the data but a property of natural language; nevertheless, it implies that robust claims require an explicit codebook and inter-annotator reliability assessment. Finally, the CNC category is a coarse proxy for noun phrase complexity. Future work should treat complexity as a set of features (premodifier count, postmodifier type, embedding depth) rather than as a residual category.

### *5.3 Implications and Future Work*

Despite these limitations, the pilot illustrates a productive way to connect formal and functional perspectives without conflating their evidential roles. Formal analyses provide principled diagnostics for distinguishing nominal types and for identifying argument-structure behaviour (Picallo 1991; Baker & Vinokurova 2009). Corpus-register analysis then asks how often speakers and writers select these options in different communicative contexts. Under this division of labour, corpus data do not replace theory-internal arguments; instead, they reveal how theoretically characterised options are distributed in use.

A natural next step is a larger-scale study with transparent sampling and modelling. One extension is to extract a larger random sample from a single corpus with multiple registers, annotate event/agent type, and model the probability of eventive nominalisation as a function of register and morphological family. Another is to shift from type counts to structural diagnostics: how often do eventive nominals realise internal arguments (of-phrases), how often are agents expressed (by-phrases or genitives), and how do these features vary by register? Such analyses would directly connect the distributional profile to the mechanisms of clausal compression and depersonalisation.

## 6. Conclusion

This article reported a pilot investigation of nominalisation types across Academic, Spoken, and Fiction register samples from COCA and the BNC. Within a manually annotated stratified sample of 500 tokens, academic writing strongly favours eventive nominalisations, while non-academic registers favour agentive nominals and show more complex multiword noun phrase structures. The results are consistent with the view that academic discourse employs nominalisation as a strategy of clausal compression, concept formation, and depersonalised assertion.

Methodologically, the study highlights both the promise and the challenge of linking theoretically motivated nominal categories to corpus data. A careful division of labour between formal diagnostics and register-functional interpretation makes this link productive while respecting differences in evidential standards. Future work can build on the pilot by expanding the dataset, improving annotation reliability, and modelling argument-structure and complexity features directly.

## References

Baker, M. C., & Vinokurova, N. (2009). On agent nominalizations and why they are not like event nominalizations. *Language, 85*(3). 517-556.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

British National Corpus Consortium. (2007). *The British National Corpus,* version 3 (BNC XML Edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium.

Davies, M. (2008). The Corpus of Contemporary American English (COCA). Retrieved from https://www.english-corpora.org/coca/

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language, 86*(3). 663-687.

Nichols, J. (1989). Nominalization and assertion in scientific Russian prose. In J. Haiman, & S. A. Thompson (Eds.), *Clause Combining in Grammar and Discourse* (pp. 599-628). Amsterdam: Benjamins.

Picallo, M. C. (1991). Nominals and nominalizations in Catalan. *Probus, 3*(3), 279-316.

Spike, M. (2020). Fifty shades of grue: Indeterminate categories and induction in and out of the language sciences. *Linguistic Typology, 24*(3), 465-488.

## Appendix A. Coding Categories (Pilot)

Event nominalisation (EN): nominal head denotes an eventuality (process, action, event) and is compatible with internal argument realisation (often of-PP) or event-oriented modification (e.g., gradual, rapid, continuous).

Agent nominalisation (AN): nominal head denotes a participant (often human) associated with an eventuality and behaves as a referential count noun (typically pluralisable, readily used with determiners, limited argument structure).

Complex nominal construction (CNC): multiword noun phrase selected by extraction heuristics (e.g., N of N, noun-noun sequences, heavily modified NPs) that was not decomposed into event vs agent type in the pilot.

## Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/)