Bridging human and AI perspectives: semantic annotation of generic nouns in German

Iván Arias-Arias[1], **Elena Martín-Cancela**[2]

[1]Instituto da Lingua Galega, Universidade de Santiago de Compostela, Praza da Universidade 4, 15782 Santiago de Compostela, Galicia

[2] Universidade da Coruña, Rúa Lisboa 7, Campus da Zapateira, 15008 A Coruña, Galicia

E-mail: ivanarias.arias@usc.gal, elena.martin@udc.es

Abstract

Generic nouns such as *Sache* and *Ding* pose a challenge for semantic annotation due to their referential underspecification and context-dependent meaning. Although frequently classified under categories like {artefact} or {object}, their actual referents often belong to abstract or cognitive domains, as in *Der Placeboeffekt ist eines der faszinierendsten Dinge in der Welt der Medizin*. Drawing on valency grammar, this study shows that these nouns activate different argument structures depending on their syntagmatic environment, reflecting semantic flexibility and combinatorial variability. Lexical databases such as GalNet or GermaNet frequently assign multiple synsets to these nouns, illustrating their ontological ambiguity. This paper examines whether large language models (LLMs) can replicate this nuanced classification. Using a gold standard corpus annotated by linguists, we implement a two-step prompting strategy —supplying LLMs with predefined semantic tags and contextual windows— to test their performance. The results underscore the limitations of current LLMs in dealing with the lexical underspecification of generic nouns, even when provided with an extended context window. These findings contribute to ongoing discussions on the automation of semantic tagging and point to meaningful ways in which AI systems can complement human expertise in natural language processing tasks.

Keywords: automatic semantic annotation; generic nouns; large language models; lexicological information systems; valency grammar

1. Introduction

Automatic semantic annotation of corpora is a fundamental challenge in natural language processing, especially when working with generic nouns like the German *Sache* and *Ding* ('thing'). The semantic classification of these nouns is complicated by their referential underspecification and context-dependent meanings.

Despite significant advances in natural language processing, the polysemy and inherent underspecification of certain nouns —known as "generic nouns", "general nouns" or "shell nouns"— present a considerable barrier to precise and coherent semantic classification. These nouns, often superficially classified as {artefact} or {object}, frequently refer to abstract or cognitive domains.

Historically, general nouns have received scant attention in linguistic research. Earlier studies (Gödeke et al., 2022) have emphasised their unique function within discourse, noting that they not only contribute to world-building in narratives but also convey crucial information about the narrative instance, the significance of what is reported, or even the real world. Unlike nouns with more concrete referents, the interpretation of general

nouns is highly context-dependent, complicating their formal description and automatic recognition.

From a linguistic perspective, the "borderline nature" of generic nouns —situated between lexical and grammatical items— imbues them with a particular cohesive function. Their interpretation often relies on reference to preceding elements in the text, like pronouns or correlates (cf. Mollica, 2010). According to corpus-based studies (Halliday & Hasan, 1976; Mahlberg, 2005), word meaning emerges from recurrent contextual patterns. This is especially evident with generic nouns, as their combinatorial behaviour plays a crucial role in constructing and specifying their actual meaning. Generic nouns can trigger different argument structures depending on their syntagmatic environment.

Against this backdrop, this study investigates the capability of large language models (LLMs) to perform context-sensitive semantic annotation of the German generic nouns *Sache* and *Ding*. Our research addresses the question of how reliably LLMs can replicate this nuanced classification of generic nouns. Lexical databases such as GalNet and GermaNet (see §3) already reflect this ontological ambiguity by assigning multiple synsets to these nouns, thereby highlighting the inherently intersubjective nature of semantic annotation.

To assess LLMs' capacity for capturing context-sensitive meaning and addressing the lexical underspecification of *Sache* and *Ding*, we implemented a two-step prompting strategy. This strategy entails supplying the LLMs with predefined semantic labels and context windows, and comparing their output with a gold standard previously annotated by expert linguists. We have adopted the annotation framework of the Portlex lexical ontology (see §3), which aligns with valency grammar principles and employs mutully exclusive, broadly applicable top-level categories for noun description.

Our findings in this study not only contribute to the discussion concerning the automation of semantic annotation for underspecified nouns, but also form part of a broader discourse on how artificial intelligence can address complex linguistic phenomena such as generalisations, complementing human expertise in NLP tasks.

This paper is structured as follows: section 2 explores general considerations regarding generic nouns, discussing their terminology and semantic particularities from the perspective of valency grammar. Section 3 examines their representation in lexical information systems. Section 4 outlines the annotation methodology, including corpus and LLM-based annotation framework. Sections 5 and 6 present and analyse the results obtained. Finally, section 7 discusses the implications of our findings and section 8 provides conclusions and avenues for future work.

2. Generic nouns: general considerations

The elusive nature of certain nouns has led to a variety of terms and approaches in linguistics. This section examines the various labels assigned to these nouns, their inherent discursive nature and their semantic combinatorial behaviour. In linguistic literature, several expressions are used to refer to the category of nouns under consideration —most notably "generic nouns", "general nouns" or "shell nouns". Although the nuances differ slightly, all these terms share a common feature: a non-specific referential nature and meaning that arises from context.

Halliday & Hasan (1976), who introduced the concept, locate these nouns at the intersection between lexical items (an open class) and grammatical items (a closed class). Their examples include "people, person, man, woman, child, boy, girl [human]", "thing, object [inanimate concrete count]", "business, affair, matter [inanimate abstract]" and "place [place]". They argue that such nouns possess "generalised reference" within major noun classes and fulfil a specific cohesive function, serving as correlates in discourse organisation —much like synonyms of earlier elements or even as implicit referents alongside pronouns. However, subsequent corpus-based studies have challenged the consistency of their interchangeability with pronouns. Halliday & Hasan (1976) further noted that the "generality" of meaning in such nouns makes them difficult to apprehend.

Mahlberg (2005), from a corpus-driven perspective, defines general nouns as those that are "relatively frequent" and that "fulfil multiple textual functions". She argues that meaning arises from co-occurrence patterns rather than from words in isolation; in this sense, frequency is integral to lexical meaning. Mahlberg emphasises the importance of local textual functions, explaining how these nouns contribute to textual cohesion. One identified role is the "support function", where the noun serves less to convey propositional content and more to structure discourse appropriately (cf. Mollica, 2010).

The term "shell nouns" (Schmid, 2000) overlaps with the notion of generic nouns: these can "encapsulate" sections of discourse, invoking meaning based on contextual elements. Frequently, they integrate {factual} complements, thus prompting abstract interpretations such as {act} or {event}. In short, generic nouns form a particular subset characterised by their high frequency, inherently general or underspecified meaning, and capacity to perform a range of textual functions. These attributes make them both syntactically flexible and semantically rich.

2.1 On the semantic combinatorics of generic nouns

Valency grammar widely assumes that not only verbs and adjectives, but also nouns can open complement slots within the syntactic and semantic context in which the nominal predicate appears (cf. Engel, 2004; Hölzner, 2007; Domínguez Vázquez, 2011; Wöllstein & Dudenredaktion, 2022). The concept of valency has evolved into a multidimensional framework (Hölzner, 2007; Domínguez Vázquez, 2011) in which several linguistic levels —logical-semantic, morphological, syntactic and pragmatic— play a significant role. As in other theoretical approaches (cf. Mel'čuk, 2015), the perspective adopted in the present study holds that semantic actants are integral to the inherent meaning of the predicate, even if they are not always overtly realised at the surface syntactic level (Hölzner, 2007; Sántáné-Túri, 2020).

Noun complements may provide essential information and can even lead to different interpretations of the core meaning of the nominal predicate (compare, for instance, *the question of the inflation* vs. *the question of the participants*) (cf. Arias-Arias, 2025; Valcárcel Riveiro & Pino Serrano, 2023). This phenomenon is particularly evident in the case of generic nouns such as *Sache* and *Ding* ('thing'), which constitute the focus of the present study. Special attention will therefore be paid to their combinatorics, as it is through their semantic arguments that the actual meaning of the predicate is effectively constructed and specified.

In light of the debate as to whether only deverbal nouns can function as predicates or whether distinct classes of nominal predicates should be differentiated (Hölzner, 2007; Sántáné-Túri, 2020), the classification of generic nouns remains an open issue, particularly given their lack of a corresponding verbal base. It is also noteworthy that such nouns are not included in major reference works on noun valency, such as the valency dictionary by Sommerfeldt & Schreiber (1983). This paper argues that generic nouns are indeed capable of opening valency slots and therefore functioning as predicates; however, this frequently occurs beyond the boundaries of the noun phrase itself —that is, at a transphrastic level (cf. Hölzner, 2007).

The valency patterns of generic nouns (here, *Sache* and *Ding*) directly influence the activation of one meaning or another. The analysis will begin with cases of active valency; namely, valency that is directly governed by the noun predicate (Domínguez Vázquez, 2022) and realised within the noun phrase (intraphrastically). In such instances, only a limited number of argument slots appear:[1]

- A person or entity who is responsible for something (as quasi-*Experiencer* or quasi-*Ferens*,[2] (cf. Domínguez Vázquez, 2011): *Sache der Polizei* ('a matter of the police'). In this case, the predicate assumes an abstract meaning and the syntactic realisation implies the zero realisation of the determiner.[3] The argument shall be ontologically described in this case as {human} or {institution}.
- As element that is classified as belonging to a higer-level category —referred to as *Klassifikativ* in Domínguez Vázquez (2011)— is exemplified by expressions such as *staatspolitische Dinge* ('political matters'), which can be paraphrased as 'matters belonging to the domain of politics'.[4] Another syntactic realisation of this classification argument, also interpreted as a noun complement (cf. Domínguez Vázquez, 2011), is found in the form of a subordinate clause, as in *die Sache, dass die Frist kurz bevorsteht* ('the fact that the deadline is approaching'). In constructions such as the appositive *Dinge wie Ausstellungen* ('things like exhibitions'), the complement also serves a classification function, with the generic noun (*Ding*) being further specified by its complement. In such cases, the interpretation of the predicate also tends to be abstract (*Sache* or *Ding* as {act} or {event}), and the categorial meaning of the argument (Engel, 2004) may vary considerably, ranging from {institution} to {intellectual}.

When considering transphrastic realisations of active valency, we may observe the occurrence of what Sántáné-Túri (2020: 74) describes as a quasi-argument: a type of complement

[1] Unless otherwise specified, all language examples used in this study are drawn from the *Deutsches Referenzkorpus* (DeReKo).

[2] Domínguez Vázquez (2011: 172) defines the *Experiencer* as the semantic relational role that denotes an entity affected —either physically or mentally— by the corresponding event. A related case is that of the *Ferens* also described by Domínguez Vázquez (2011: 172) as the entity that is affected, though not subjected to a constant or ongoing influence by the event. These roles appear to adequatly capture the semantic combinatorics of the selected generic nouns when describing the first argument position, as opposed to the role of *Agens*.

[3] The realisation with a determiner and plural form (*seine Sachen* 'his things') typically refers to the concept of personal belongings, whereas the singular form (*das Ding* 'the thing') tends to denote a more concrete and individuated referent.

[4] This phrase also gives rise to an alternative interpretation, which corresponds to the paraphrase *Sachen der Staatspolitik* ('matters of state policy').

which functions as an appositive to the subject and contributes to the activation of its inherent meaning.[5] This is frequently the case with subordinate clauses that function as subject or object complements of the main verb and which correlate anaphorically or cataphorically with the given generic noun (Kolhatkar & Hirst, 2014). In this regard, Schmid (2000) and Kolhatkar & Hirst (2014) emphasise that generic nouns belonging to the 'thing' category often take {factual} elements as complements. This supports that the meaning activated in such contexts tends to be abstract in nature, frequently involving semantic categories such as {act} and {event}, among others.

From the perspective of the passive valency (Domínguez Vázquez, 2022) of generic nouns, it should first be noted that certain phraseological structures are highly lexicalised and, to varying degrees, presuppose the presence of a noun complement, whose dependence on the predicate may range from weak to strong. This is particularly evident in expressions such as *guter Dinge sein* ('to be in a good mood') or *in der Natur der Sache liegen* ('to lie in the nature of the matter'), where the categorial classification of the predicate transcends its internal lexical features and assumes a phraseological function. An example such as *Es liegt in der Natur der Sache, dass Forscher Grenzen überschreiten müssen, um zu neuen Erkenntnissen zu kommen* ('It lies in the nature of the matter that scientists must overcome boundaries to arrive at new discoveries') illustrates that it is the subordinate clause which determines the interpretation of the generic noun *Sache*.

To sum up, this section has shown that both the active and passive semantic combinatorics of generic nouns are crucial for their semantic interpretation and tagging, as it is these contextual realisations that activate the actual meaning of the predicate. The structures containing the noun predicates *Sache* or *Ding* are semantically underspecified (cf. Pustejovsky, 1995; Pustejovsky & Batiukova, 2019), which contributes to their flexibility and allows them to occur in a wide range of contexts. In the absence of such contextual anchoring, these nouns can be considered semantically underspecified —or even empty— and their meaning is actualised through their usage in discourse.

3. Description of generic nouns in lexical information systems

Automating the semantic tagging of generic nouns necessarily entails the consultation of various lexical information systems. These resources serve as reference frameworks for the annotation process, helping to identify potential (in)consistencies in the ontological categorisation of lexical items. In order to determine which categories are most suitable for the semantic-ontological description of the selected units, three lexical resources with ontological structuring are examined:

- GalNet, a multilingual lexical network based on the principles of EuroWordNet (Vossen, 1998). It includes data for several languages, including German, and connects synsets across languages via the Interlingual Index (ILI), facilitating cross-linguistic comparison (Solla Portela & Gómez Guinovart, 2015). Its ontological classification is enhanced through the integration of external models such as the Suggested Upper Merged Ontology (SUMO) and WordNet Domains.

[5] As illustrated by the examples above, this phenomenon also manifests within noun phrase frames, and it may be argued taht all occurrences involving *Sache* and *Ding* as predicates function as quasi-arguments.

- GermaNet, a comprehensive lexical-semantic network for German that includes nouns, verbs and adjectives. Particularly robust in its semantic field classification, GermaNet assigns each noun to at least one top-level category, which functions as its primary semantic classifier. While polysemous lexical entries are typically associated with multiple synsets and categories, the developers strive to limit polysemy to maintain representational clarity (Hamp & Feldweg, 1997; Hinrichs et al., 2020a).
- Portlex ontology (Domínguez Vázquez et al., 2021), a lexical ontology designed for multilingual applicability, developed from a bottom-up perspective grounded in corpus analysis and guided by the principles of valency grammar. Its lexical entries were generated using a hybrid method that draws on data extracted from GalNet via combined APIs. The ontology focuses on nouns and reduces the number of upper categories to enhance interoperability in information extraction tasks (Martín Gascueña, 2023).

These three lexical databases are consulted for two main purposes: first, to examine the ontological variability in the classification of generic nouns (see table 1); and second, to develop a consistent annotation framework that can be applied both by human annotators and by the tested LLMs, which will be presented with a predefined set of semantic tags from which to choose. The analysis focuses exclusively on the top-level categories provided by each resource, as these already reveal substantial ontological variation in the way generic nouns are conceptualised and classified.

GalNet	**GermaNet**	**Portlex lexical ontology**
Sache: • {physical entity} • {content} • {act} *Ding*: • {entity} • {physical entity} • {physical entity abstraction}	*Sache*: • {communication} • {top} • {cognition} *Ding*: • {act} • {communication} • {human} • {top} • {group}	*Sache/Ding*: • {intellectual} • {material} • {dynamic situation} • {static situation}

Table 1: Classification of *Sache* and *Ding* in different lexical resources

The ontological classification derived from the consultation of different lexical resources highlights the intersubjective nature of semantic annotation, which is further shaped by the specific goals of each project. In the case at hand, GalNet and GermaNet differ significantly in their classification of the two selected nouns despite their potential near-synonymy. This divergence is particularly evident in the abstract domain: while GalNet may classify a noun under {act}, GermaNet may instead opt for a category such as {cognition} or {communication}. Notably, GermaNet attributes the category {top} to both lexical items —a higher-level category that encompasses broader conceptual groupings.

In contrast, the Portlex lexical ontology opts for a more constrained categorisation, limiting itself to four upper-level categories to which these nouns may be assigned depending on their valency structure. The valency-based principles that underlie this resource, combined with its detailed classification of nominal structures, make the Portlex lexical ontology a particularly suitable annotation framework in this context. Moreover, its layered descriptive model allows for the future expansion of semantic granularity, adapting to the evolving needs of the annotation process.

4. Methodology: corpus and annotation framework with LLMs

The main objective of the present paper is the semantic annotation of generic nouns through the application of LLMs. To achieve this, consideration must be given to the range of contexts in which these nouns occur and the interpretative variations they allow. Kolhatkar et al. (2013) have already pointed out the challenges of annotating shell nouns along with their antecedents, as this process requires a deep understanding of the text and the cohesive relations it entails. In our study, although the consistent annotation of antecedents is not the primary goal, the cataphoric and anaphoric relations between the generic noun and the element that semantically completes or specifies it (an argument or quasi-argument) are crucial for achieving coherent and accurate semantic annotation.

4.1 The Portlex lexical ontology as annotation framework

For the semantic annotation of generic nouns, we relied on the ontological framework provided by the Portlex lexical ontology.[6] Specifically, we employed four top-level ontological classes, selected because they could plausibly apply to the two lexical units under study. Annotation was restricted to this upper level of the ontology, avoiding more fine-grained subcategories to maintain consistency and feasibility in both manual and automatic tagging. The ontological categories used were {material}, {intellectual}, {dynamic situation}, and {static situation}. These were chosen for their broad applicability and mutual exclusivity, which is conceptualised in part through relations of negation; in other words, a noun not classifiable as {material} might instead be considered {intellectual} etc. This logical framing allows for a simplified decision-making process when tagging ambiguous or polysemous nouns.

The use of more granular ontological categories, while potentially yielding more detailed semantic representations, introduces significant complexity for both automatic tagging systems and LLMs and may thus reduce consistency and accuracy in annotation. To support both human annotators and LLM-based tagging, we provided concise glosses for each of the four categories. These glosses facilitate greater inter-annotator agreement (cf. Stefanowitsch, 2020) and serve as prompts for the task of semantic classification in LLMs. Their role is thus both descriptive and operational:

- {material}: refers to entities in the physical world that are tangible and observable; these can include both substances and discrete objects.

[6] This work is part of a broader project focused on the development of an automatic semantic annotator, which further justifies the use of the Portlex lexical ontology as the foundational framework for the annotation procedure.

- {intellectual}: includes abstract yet perceptible entities related to cognition, communication, or mental content —such as ideas, messages, or knowledge.
- {dynamic situation}: encompasses events, actions and processes that unfold over time and may involve agents (either human or non-human). These are inherently temporal and can include both natural occurrences and intentional acts.
- {static situation}: captures states, conditions, and properties —typically atemporal phenomena such as qualities, attributes, or relational configurations.

In the first phase of the annotation process, two expert linguists were tasked with annotating 30 concordances of the generic nouns *Sache* and *Ding* extracted from two different corpora: (a) the gold standard corpus developed within the framework of the ESMAS-ES$^+$ project,[7] which comprises multilingual texts sourced from the TED2020 repository available through the OPUS corpus platform; and (b) the German reference corpus DeReKo. Given that the ESMAS-ES$^+$ project's gold standard corpus contained only five occurrences of these generic nouns, the dataset was expanded with an additional 25 concordances extracted from DeReKo, resulting in a total of 30 concordances. This expansion was necessary to obtain a sufficient sample size for observing and analysing annotation patterns, particularly in the context of LLM-based semantic tagging.

All concordances included a three-sentence context window: the target sentence containing the generic noun, plus the immediately preceding and following sentences. Among the 30 concordances, two contained multiple occurrences of *Ding* or *Sache* within the same context window. In both cases, human annotators classified all instances identically, as they referred to the same referent, resulting in no semantic variation within these concordances. This structure ensured sufficient context for semantic disambiguation. In cases where there was no initial agreement between human annotators regarding the corresponding semantic category, consensus was reached through discussion and by consulting several lexical resources (e.g., monolingual German dictionaries such as Duden or DWDS, lexical databases and corpus examples) to determine the most suitable classification for each instance. Once this consensus was established, the human inter-annotator agreement provided a gold standard, which serves as a baseline for evaluating the performance of LLMs as semantic annotation systems.

4.2 LLMs: overview and prompting strategy

The LLMs selected for the present study were based on two main criteria: their widespread adoption in both research and applied contexts, and the frequency with which they are updated by their developers.[8] These updates often incorporate advancements in model architecture or training data, with the potential to improve performance over earlier versions.[9] Moreover, the selection was restricted to conversation-based language models

[7] Given that the primary goal of this research project is the development of a semantic automatic tagger for multilingual data, it is methodologically appropriate to extract concordances from the project's core corpus in order to inform and evaluate the annotation procedures. These concordances will serve as a baseline for comparison with the results produced by the automatic tagger in future phases of the project.

[8] Information on the usage and performance of various models is available at the following website: https://eleks.com/blog/best-llms-for-language-processing/.

[9] Note that these experiments were conducted between April and June 2025 and reflect the state of development of the selected LLMs at that point in time.

—frequently referred to as chatbots— because of their interactive capabilities, as they are designed to sustain multi-turn dialogues in real time. This interactivity allows researchers to iteratively refine their prompts based on the models' responses, which provide an intuitive and accessible interface for exploration (cf. Nasution & Onan, 2024; Yu et al., 2024; Petukhova & Kochmar, 2025). The interactional setup is particularly suitable for applications in lexicographic contexts (cf. Arias-Arias et al., 2024; Alonso Ramos, 2023; Tarp & Nomdedeu-Rull, 2023), as well as other NLP tasks (Enis & Hopkins, 2024). The following LLMs were employed for automatic semantic annotation:[10]

- ChatGPT-4o is a variant of GPT-4 developed by OpenAI, optimised to deliver similar performance with lower computational requirements. It retains the core Transformer architecture, and it achieves competitive results on many standard NLP benchmarks, even surpassing GPT-4 in some multilingual comprehension evaluations (cf. Siddiky et al., 2025).
- Gemini 2.5 Pro is a generative LLM developed by Google. Its advanced understanding capabilities make it suitable for a wide range of tasks in NLP, "including text summarization, object recognition, content understanding, classification and extrapolation" (Islam & Ahmed, 2024).
- Claude 4 Sonnet[11]was developed by Anthropic and launched in May 2025. It features enhanced capabilities for following instructions, and the company claims it can achieve very high levels of accuracy in research and scientific discovery, thanks to a recently implemented problem-solving foundation.

The prompting strategy employed in this study follows a zero-shot approach, meaning that the language model is not given any explicit examples of the task it is asked to perform. Instead, the prompt includes a detailed description of the procedure and clear, step-by-step instructions, along with a definition of the four semantic categories from which the model must choose. Additionally, the prompt presents the 30 items to be annotated (see Appendix, §9). This setup allows us to evaluate the model's performance based solely on the general knowledge acquired during training, without any task-specific fine-tuning or in-context learning. Zero-shot prompting is used to assess how well a model can handle novel tasks based on its internal representations (cf. Islam & Ahmed, 2024; Yu et al., 2024; Bhattacharjee et al., 2024). The instructions are formulated in English, as this language is still considered to dominate the training data of most LLMs (cf. Zhong et al., 2024).

4.3 Statistical procedure for LLMs performance

The evaluation of the annotation framework will be carried out using three standard metrics —precision, recall, and F1-score— by comparing the model's output to the gold standard that was annotated by two expert linguists (cf. Nasution & Onan, 2024; Yu et al., 2024):

[10] Access to the different sessions and conversations with the LLMs is available through the following links: ChatGPT [https://chatgpt.com/c/6856d16d-bb54-8010-9883-cf14dbf84b58], Gemini [https://gemini.google.com/app/5e56c1c85d101537?hl=pt] and Claude [https://claude.ai/public/artifacts/12f4c7e7-0d5b-4e7e-8a73-c1c20e6fb8a8].

[11] For more information on the capabilities of the model, consult https://www.anthropic.com/news/claude-4.

- Precision measures the accuracy of the positive predictions made by a model. It is defined as the ratio of true positives to the total number of items the model predicted as positive (including true positives and false positives).
- Recall evaluates the model's ability to identify all relevant instances in the data. It is calculated as the ratio of true positives to the total number of actual positives.
- F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. The harmonic mean gives more weight to lower values, which means the F1-score will only be high if both precision and recall are reasonably high.

In line with the approach outlined by Nasution & Onan (2024), the evaluation also includes inter-annotator agreement measures to assess the consistency and reliability of the different annotations (cf. Stefanowitsch, 2020). To this end, Fleiss' Kappa is used: a statistical measure designed to quantify the degree of agreement among multiple annotators assigning categorical labels. This metric adjusts for agreement expected by chance, thus providing a robust estimate of actual consensus. High values of inter-annotator agreement indicate that annotation guidelines are well-defined and that the task is clearly understood by annotators, whereas low values might reflect ambiguity or subjectivity in the labelling process.

Cohen's Kappa will also be calculated to measure the agreement between each individual system and the gold standard. Unlike Fleiss' Kappa, which considers multiple annotators simultaneously, Cohen's version focuses on pairwise agreement, making it particularly suitable for evaluating how closely each language model aligns with the human annotation (cf. Nasution & Onan, 2024; Stefanowitsch, 2020). These agreement metrics not only validate the quality of the gold standard but also contextualise the performance of the LLMs in relation to human judgment.

5. How efficient are LLMs at annotating generic nouns?

As indicated above (see §4.3), this section provides a detailed discussion of the results of each statistical parameter (see table 2). In the case of ChatGPT, the precision value is notably low, suggesting that the system frequently fails to assign the correct semantic label to the lexical units under consideration. Equally important, however, is the observation that the recall is also low, indicating that the system not only mislabels items but also fails to recognise many relevant instances that should have been annotated. These results suggest that ChatGPT struggles both to identify relevant lexical units and to assign the appropriate semantic classes, potentially due to a mismatch between its underlying representations and the specific annotation schema used.

Gemini shows a substantial improvement over ChatGPT. The reported values point to a more balanced and competent performance: Gemini not only assigns more accurate semantic labels but also demonstrates a better ability to identify relevant lexical items. Nevertheless, the gap between precision and recall suggests that there is still room for improvement in terms of coverage and generalisation. Claude, in turn, outperforms both systems in terms of precision, indicating a higher level of accuracy when assigning semantic labels. However, its recall remains moderate, which implies that the system still fails to capture a significant number of instances. Despite this limitation, Claude achieves the highest F1-score among the three systems, reflecting the best overall performance.

System	Metric	Value
ChatGPT	Precision	0.0543
	Recall	0.208
	F1	0.0862
Gemini	Precision	0.487
	Recall	0.444
	F1	0.452
Claude	Precision	0.65
	Recall	0.471
	F1	0.511

Table 2: Performance comparison of AI systems

In essence, these results underscore the inherent trade-offs between precision and recall in LLMs. While Claude offers the most favourable balance, Gemini delivers a solid intermediate performance, and ChatGPT falls short in both accuracy and coverage, highlighting its limited reliability for this annotation task.

The confusion matrices presented in figure 1 offer a detailed view of the classification behaviour of the LLMs when compared to the gold standard.

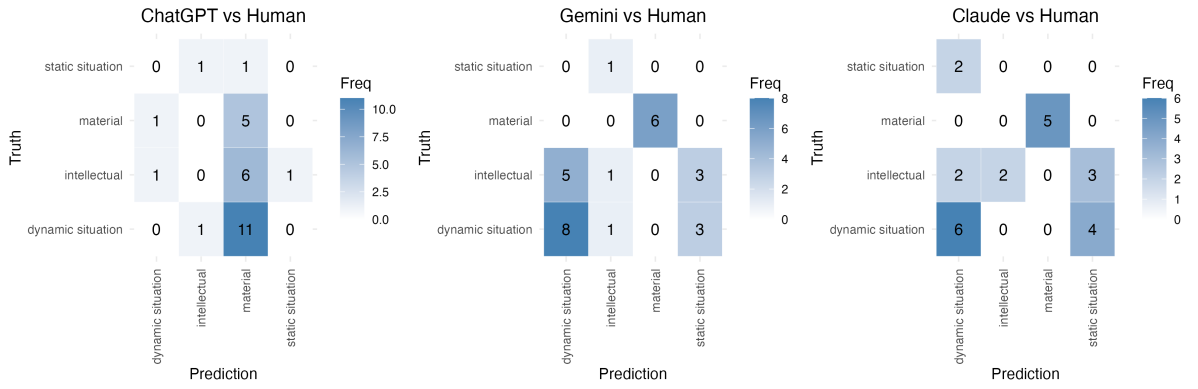

Figure 1: Predicted and true values of *Sache* and *Ding* across LLMs

ChatGPT shows a strong bias toward the {material} category, which is overused across all classes. Notably, 11 instances that should have been labelled as {dynamic situation} were incorrectly assigned to {material}, and the same misclassification appears in the {intellectual} row. The matrix indicates that the model struggles to distinguish between semantic classes. Gemini, in turn, performs more consistently. Most {material} lexical units are correctly classified, and there is relatively less confusion across other categories. However, {intellectual} and {dynamic situation} are occasionally misclassified as {static situation}, suggesting that the model may conflate abstract and situational categories. Claude, for its part, achieves the most balanced performance. Correct predictions appear for all classes, especially for {material} and {dynamic situation}, and the confusion is

more evenly distributed. Nevertheless, some {static situation} instances are misclassified as {intellectual} or {dynamic situation}, pointing to occasional difficulty in capturing fine-grained semantic distinctions.

To assess the overall consistency among the systems under evaluation, inter-annotator agreement was measured using Fleiss' and Cohen's Kappa (cf. Nasution & Onan, 2024). Fleiss' Kappa value for the three systems (ChatGPT, Gemini, and Claude) across 21 items[12] is $\kappa = 0.195$ ($p = 0.0151$). Although modest, this value is statistically significant and suggests a level of agreement that is higher than chance. According to commonly used interpretation thresholds (cf. Landis & Koch, 1977), this corresponds to "slight" to "fair" agreement, indicating that the systems share some patterns in how they assign semantic labels.

In contrast, the Cohen's κ values reveal more variation when each system is compared individually to the gold standard:

- ChatGPT yielded Cohen's $\kappa = -0.0662$ ($p = 0.331$), a non-significant and even negative value. This suggests that the agreement between ChatGPT and the human annotation is no better than chance, and potentially even worse.
- For Gemini, the agreement improves markedly, with $\kappa = 0.349$ ($p = 0.00084$), which is moderate and statistically significant. This indicates that Gemini can replicate human decisions to a meaningful degree.
- Claude shows the highest level of pairwise agreement, with $\kappa = 0.376$ ($p = 0.000597$), and is statistically significant. This value borders on the "moderate" agreement threshold and aligns with Claude's stronger performance on classification metrics.

These results suggest that Gemini, and especially Claude, produce annotations that are more consistent with human judgment, whereas ChatGPT's outputs diverge significantly, both in terms of precision and inter-annotator agreement. The relatively low Fleiss' κ further indicates limited consistency across systems, highlighting the need for more robust semantic alignment or for more precise and system-specific annotation guidelines tailored to the characteristics of each model's training data. The discrepancies observed across systems may stem from several factors. One possibility is the inconsistency of class definitions within the Portlex lexical ontology, which may result in ambiguous or insufficient glosses provided to the systems (see §4.1). Another contributing factor could be the conflict between the factual data included in the models' pretraining corpora and the input provided during the annotation task. These limitations and potential causes will be further examined in the discussion presented in §6.

6. Discussion of the results

The statistical analysis presented in §5 invites several qualitative considerations regarding the performance of the LLMs in the task at hand. First and foremost, it must be noted that the models often struggle to identify the lexical units to be annotated, even when the prompt explicitly defines the task. In other words, the LLMs were instructed to annotate

[12] Only the items that were identified and therefore annotated by all three systems are considered in these metrics.

occurrences of the nouns *Sache* and *Ding* in context, rather than annotating generic nouns in general, which might have yielded even less accurate results.

For example, ChatGPT frequently fails to recognise instances of *Dingen*, the dative plural form of *Ding*. As a result, such occurrences are not annotated at all, due to the system's inability to link inflected forms to their lemmas. In contrast, Claude and Gemini exhibit a different kind of limitation: they sometimes fail to annotate forms that do match the lemma exactly. In Claude's case, this seems to be related to the combinatorial or collocational patterns in which the noun appears. For instance, in expressions like *beschlossene Sache* ('settled matter') or *persönliche Sachen* ('personal belongings'), the model may infer idiomatic or metaphorical readings that prevent it from annotating the lexical unit as expected.

Furthermore, as previously mentioned, ChatGPT shows a strong bias toward the {material} class, which may reflect the nature of its training data. In many corpora, generic nouns like *Ding* or *Sache* may frequently appear in contexts where they denote physical objects, leading the model to favour this interpretation. However, this overrepresentation often ignores both the anaphoric function of these nouns in discourse and their context-dependent meaning activation. There are clear instances —annotated as {dynamic situation} by both the human annotator and the other systems— that ChatGPT misclassifies as {material}:

(1) Wer sich mit der **Sache** genauer befasst hat, weiß, dass Merz mit dem Verkauf der Kantonalbank den Kanton vor einem riesigen finanziellen Debakel bewahren konnte. ('Anyone who has taken a closer look at the **matter** knows that Merz was able to save the canton from a huge financial debacle with the sale of the Kantonalbank.')

(2) Skandalträchtig ist zum Beispiel, wenn ein Politiker sein Amt für persönliche Vorteilnahme missbraucht. In der Schweiz war es aufgrund der Konkordanzdemokratie lange Zeit möglich, solche **Dinge** unter den Teppich zu kehren, was künftig immer weniger der Fall sein wird. ('It is scandalous, for example, when a politician abuses his office for personal gain. In Switzerland, it was possible for a long time to sweep such **things** under the carpet due to the concordance democracy, which will be less and less the case in future.')

Gemini and Claude, in turn, appear to encounter greater difficulties when it comes to correctly classifying certain instances as belonging to the categories {dynamic situation}, {static situation}, or {intellectual}. It is worth noting that these are precisely the cases where human expert annotators also showed the greatest divergence during earlier phases of manual annotation. This was particularly evident in the distinction between {dynamic situation} and {static situation}, where the referent is sometimes ambiguous or open to multiple interpretations. The following examples serve to illustrate these cases more clearly:

(3) Es sei kaum **Sache** einer Gemeinde, privaten Wohnungsbau zu betreiben. ('It is hardly the **responsibility** of a municipality to build private housing.')

(4) Er ist einer der begabtesten Politiker der Gegenwart. Rechtlich sieht die **Sache** anders aus. Die laufende Strafuntersuchung kontrolliert nicht er, sondern sein Gegenspieler Starr, der einen beispiellosen Feldzug führt. ('He is one of the most talented politicians of our time. Legally, the **situation** is different. The ongoing criminal investigation is not controlled by him, but by his opponent Starr, who is waging an unprecedented campaign.')

The preceding cases demonstrate that the tags {intellectual} and {dynamic situation} assigned by human annotators are not always consistently recognized by the LLMs —and this is not without reason. In certain instances, such as example 3, there is no explicit referent; the noun merely invokes the notion of an agent's responsibility to respond to a given situation. In example 4, the generic noun functions cataphorically, referring to a situation that is specified in the following sentence. This kind of transphrastic realisation and long-distance dependency can present a challenge for LLMs, which may not resolve referential links beyond the sentence level. Nevertheless, both Gemini and Claude converge in classifying examples 3 and 4 as {static situation}.

The observed inconsistency in semantic tagging across the evaluated systems not only underscores the limitations of general-purpose language models in this type of task, but also suggests that providing more contextual or definitional information may prove insufficient to obtain truly reliable results. While different prompt strategies such as the inclusion of annotated occurrences can improve performance to some extent, the intrinsic variability in model outputs points to a more fundamental issue: these models are not explicitly optimised for semantic annotation tasks, which is in turn not one of their objectives. In this regard, fine-tuning a model specifically on semantic tagging data may be the only viable approach to achieving high-quality, consistent results. A task-specific model could internalise the fine-grained distinctions required by a lexical ontology, and thus be more sensitive to contextual nuances and disambiguation cues that general-purpose LLMs tend to overlook.

Moreover, the current results reinforce that LLMs, as generative tools, are not yet able to replace expert human annotators in semantic tagging workflows. Although their outputs may serve as a first-pass approximation, they often require post-editing and validation. From a workflow perspective, this dependence on human revision limits the efficiency gains expected from automated annotation and raises questions about the cost-benefit balance of using unadapted LLMs for such tasks. In short, true reliability and precision will likely only be achieved through the combination of task-specific fine-tuning and expert human oversight.

7. Concluding remarks

This paper has examined the challenges and possibilities of using LLMs for the semantic annotation of generic nouns, with a focus on referentially underspecified German nouns. Our

findings reveal both the potential and the limitations of current generative systems when applied to tasks that require fine-grained semantic disambiguation based on ontological and contextual cues.

While recent advances in NLP have made it possible to automate increasingly complex linguistic tasks, the automatic semantic annotation of nouns, and specifically of general nouns, continues to resist straightforward computational solutions. The results show that LLMs —especially Claude and, to a lesser extent, Gemini— can approximate expert annotations in a meaningful way, but still fall short of achieving consistent precision and recall. ChatGPT, in particular, demonstrates a strong bias towards the {material} category, reflecting a possible mismatch between the model's training data and the specific semantic categories defined in the Portlex lexical ontology.[13]

These findings underscore a broader point: the interpretation of generic nouns is inherently context-dependent and shaped by complex, often transphrasal, semantic relations. As such, the task of semantic annotation cannot be reduced to pattern recognition or lexical matching alone, but requires sensitivity to discourse structure, valency frames, and the cohesive function of generic nouns. LLMs should be viewed as supportive tools rather than autonomous annotators in semantic tagging workflows. Their usefulness lies in offering initial hypotheses or in assisting human experts with large-scale annotation projects. However, post-editing remains necessary.

Future work should explore the fine-tuning of LLMs on domain-specific datasets enriched with contextual semantic annotations. Such a step could substantially improve model performance and reduce the burden on human annotators. Moreover, the annotation framework presented here —based on the Portlex lexical ontology— can be further exploited to gain more granularity in the semantic description.

In sum, while the automation of semantic annotation for generic nouns remains a complex and unresolved problem, this study contributes to a growing body of work at the intersection of applied linguistics, corpus analysis and artificial intelligence. It demonstrates that interdisciplinary collaboration and methodological refinement are key to advancing our understanding in tackling one of the most elusive areas of lexical semantics.

8. Acknowledgements

This paper presents results from the ESMAS-ES$^+$ project (grant PID2022-137170OB-I00) funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU. Iván Arias-Arias acknowledges support from grant FPU21/00188 of the *Formación de Profesorado Universitario* programme, Spanish Ministry of Science, Innovation and Universities.

9. Appendix: detailed prompt used for interacting with LLMs

You are a linguist specialised in lexical semantics. Your task is to annotate 30 text examples that contain either the noun "Sache" or "Ding". You must assign one of the following four semantic classes to each instance of the noun: (1) material (refers to entities in the physical

[13] A particularly important observation is that LLM outputs may vary significantly across sessions, even when the same prompt is used. This underlines a fundamental instability in their responses and further limits their utility for tasks that demand reproducibility and consistency.

world that are tangible and observable; these can include both substances and discrete objects); (2) intellectual (includes abstract yet perceptible entities related to cognition, communication, or mental content —such as ideas, messages, or knowledge—); (3) dynamic situation (encompasses events, actions, and processes that unfold over time and may involve agents —either human or non-human. These are inherently temporal and can include both natural occurrences and intentional acts.); (4) static situation (captures states, conditions, and properties—typically atemporal phenomena such as qualities, attributes, or relational configurations.). The data is provided in TXT format, with each line containing a single text example. You must identify the noun "Sache" or "Ding" in each line (if present) and output your annotation by assigning the appropriate semantic class label to it. Please follow these rules: (1) Only annotate if "Sache" or "Ding" is present in the text. (2) Base your annotation on the contextual meaning of the noun in the sentence. (3) If there are multiple occurrences of "Sache" or "Ding", annotate based on the most prominent or semantically informative one. (4) Do not annotate any other words or tokens. Here are the 30 examples: [input TXT file].

10. References

Alonso Ramos, M. (2023). El papel de ChatGPT como lexicógrafo. In C. Garriga Escribano et al. (eds.) *Lligams: Textos dedicats a Maria Bargalló Escrivà*. Tarragona: Publicacions Universitat Rovira i Virgili, pp. 15–27.

Anthropic (2025). Claude 4.0 Sonnet. URL https://claude.ai/.

Arias-Arias, I. (2025). Nuevas vías para la desambiguación en frases nominales en alemán: fundamentos metodológico-lingüísticos para el desarrollo de una herramienta de anotación semántica (semi)automática. *Círculo de lingüística aplicada a la comunicación*. Forthcoming.

Arias-Arias, I., Domínguez Vázquez, M.J. & Valcárcel Riveiro, C. (2024). Der Effizienz- und Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz: kann ChatGPT die lexikographische Textsorte nachbilden? *Lexikos*, 34(1), pp. 51–76.

Berlin-Brandenburgische Akademie der Wissenschaften (2025). DWDS – Digitales Wörterbuch der deutschen Sprache. URL https://www.dwds.de/.

Bhattacharjee, A., Moraffah, R., Garland, J. & Liu, H. (2024). Zero-shot LLM-guided Counterfactual Generation: A Case Study on NLP Model Evaluation. 2405.04793v2.

Domínguez Vázquez, M.J. (2011). *Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens*. Munich: Iudicum.

Domínguez Vázquez, M.J. (2022). Estructura argumental del nombre: generación automática. *Signos: estudios de lingüística*, 55(119), pp. 732–761.

Domínguez Vázquez, M.J., Valcárcel Riveiro, C. & Bardanca Outeiriño, D. (2021). Portlex lexical ontology. Ontología léxica. URL http://portlex.usc.gal/ontologia/.

Dudenredaktion (2025). Duden online. URL https://www.duden.de/.

Engel, U. (2004). *Deutsche Grammatik: Neubearbeitung*. Munich: Iudicum.

Enis, M. & Hopkins, M. (2024). From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. 2404.13813.

Google (2025). Gemini 2.5 Pro. URL https://gemini.google.com/.

Gómez Guinovart, X. & Solla Portela, M. (2019). GalNet. WordNet 3.0 do galego. URL https://ilg.usc.gal/galnet/.

Gödeke, L., Barth, F., Dönicke, T., Weimer, A.M., Varachkina, H., Gittel, B., Holler, A. & Sporleder, C. (2022). Generalisierungen als literarisches Phänomen. Charakterisierung,

Annotation und automatische Erkennung. *Zeitschrift für digitale Geisteswissenschaften*, 7.

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English: Grammar and Text*. London: Longman.

Hamp, B. & Feldweg, H. (1997). GermaNet: a Lexical-Semantic Net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. pp. 9–15.

Hinrichs, M., Lawrence, R. & Hinrichs, E. (2020a). Exploring and Visualizing Wordnet Data with GermaNet Rover. In *Proceedings of the CLARIN Annual Conference 2020*. Virtual Edition, pp. 32–26.

Hinrichs, M., Lawrence, R. & Hinrichs, E. (2020b). GermaNet Rover. URL https://weblicht.sfs.uni-tuebingen.de/rover/.

Hölzner, M. (2007). *Substantivvalenz. Korpusgestützte Untersuchungen zu Argumentrealisierungen deutscher Substantive*. Tübingen: de Gruyter.

Institut für Deutsche Sprache (2025). DeReKo - Deutsches Referenzkorpus. URL https://cosmas2.ids-mannheim.de/cosmas2-web/.

Islam, R. & Ahmed, F. (2024). Gemini—the most powerful LLM: Myth or Truth. In *5th Information Communication Technologies Conference (ICTC)*. Nanjing, China, pp. 303–308.

Kolhatkar, V. & Hirst, G. (2014). Resolving Shell Nouns. In A. Moschitti et al. (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 499–510.

Kolhatkar, V., Zinsmeister, H. & Hirst, G. (2013). Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In D. Yarowsky et al. (eds.) *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, pp. 300–310.

Landis, J.R. & Koch, G.G. (1977). The Measurement of Observer Agreement for Categorial Data. *Biometrics*, 33(1), pp. 159–174.

Mahlberg, M. (2005). *English general nouns: A corpus-theoretical approach*, volume 20 of *Studies in Corpus Linguistics*. Amsterdam: John Benjamins.

Martín Gascueña, R. (2023). Diseño de una ontología de semántica léxica para los proyectos MultiGenera y MultiComb. In M.J. Domínguez Vázquez & C. Valcárcel Riveiro (eds.) *Desarrollo de aplicaciones para la generación automática del lenguaje: los recursos del portal lexicográfico Portlex (RILEX: Revista sobre investigaciones léxicas)*. Jaén: Revistas Científicas de la Universidad de Jaén, pp. 77–106.

Mel'čuk, I. (2015). *Semantics: From meaning to text*, volume 3. Amsterdam: John Benjamins.

Mollica, F. (2010). *Korrelate im Deutschen und im Italienischen*. Frankfurt a.M.: Peter Lang.

Nasution, A.H. & Onan, A.A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access*.

OpenAI (2025). ChatGPT-4o. URL https://chatgpt.com/.

Petukhova, K. & Kochmar, E. (2025). Intent Matters: Enhancing AI Tutoring with Fine-Grained Pedagogical Intent Annotation. 2506.07626v1.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge MA: MIT Press.

Pustejovsky, J. & Batiukova, O. (2019). *The Lexicon*. Cambridge: Cambridge University Press.

Sántáné-Túri, A. (2020). *Die Selbständigkeit der Substantivvalenz*. Ph.D. thesis, University Szeged: SZTE Doktori Repozitórium.

Schmid, H.J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin: De Gruyter Mouton.

Siddiky, M.N.A., Rahman, M.E., Hossen, M.F.B., Rahman, M.R. & Jaman, M.S. (2025). Optimizing AI language models: A study of ChatGPT-4 vs. ChatGPT-4o. *Preprints.org*.

Solla Portela, M.A. & Gómez Guinovart, X. (2015). Galnet: o WordNet do galego. Aplicacións lexicolóxicas e terminolóxicas. *Revista galega de filoloxía*, 16, pp. 169–201.

Sommerfeldt, K.E. & Schreiber, H. (1983). *Wörterbuch zur Valenz und Distribution der Substantive*. Berlin: de Gruyter.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Berlin: Language Science Press.

Tarp, S. & Nomdedeu-Rull, A. (2023). Who has the last word? Lessons from using ChatGPT to develop an AI-based Spanish writing assistant. *Círculo de Lingüística Aplicado a la Comunicación*, 97, pp. 309–321.

Tiedemann, J. (2025). OPUS. Open Parallel Corpora. URL https://opus.nlpl.eu/.

Valcárcel Riveiro, C. & Pino Serrano, L. (2023). Application d'une méthodologie d'analyse des prédicats nominaux: l'exemple du lexème MORT1. *Çédille: revista de estudios franceses*, 24, pp. 557–589.

Vossen, P. (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht: Springer.

Wöllstein, A. & Dudenredaktion (2022). *Duden—Die Grammatik*. Mannheim: Dudenverlag.

Yu, D., Li, L., Su, H. & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. *International Journal of Corpus Linguistics*, 29(4), pp. 534–561.

Zhong, C., F., C., Liu, Q., Jiang, J., Wan, Z., Chu, C., Murawaki, Y. & Kuroshahi, S. (2024). Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think In? 2408.10811v1.