The Cross-linguistic Role of Animacy in Grammar Structures

Nina Gregorio[*,1] **Matteo Gay**[*,2] **Sharon Goldwater**[1] **Edoardo M. Ponti**[1]
[1]University of Edinburgh [2]University of Pavia
n.gregorio@sms.ed.ac.uk

Abstract

Animacy is a semantic feature of nominals and follows a hierarchy: personal pronouns > human > animate > inanimate. In several languages, animacy imposes hard constraints on grammar. While it has been argued that these constraints may emerge from universal soft tendencies, it has been difficult to provide empirical evidence for this conjecture due to the lack of data annotated with animacy classes. In this work, we first propose a method to reliably classify animacy classes of nominals in 11 languages from 5 families, leveraging multilingual large language models (LLMs) and word sense disambiguation datasets. Then, through this newly acquired data, we verify that animacy displays consistent cross-linguistic tendencies in terms of preferred morphosyntactic constructions, although not always in line with received wisdom: animacy in nouns correlates with the alignment role of agent, early positions in a clause, and syntactic pivot (e.g., for relativisation), but not necessarily with grammatical subjecthood. Furthermore, the behaviour of personal pronouns in the hierarchy is idiosyncratic as they are rarely plural and relativised, contrary to high-animacy nouns.

1 Introduction

Animacy is a semantic property of nominals, which is defined as the degree to which their referents are perceived to be alive and sentient (Comrie, 1989; Dahl, 2008). Linguists traditionally conceptualise it as a hierarchy (Corbett, 2010), with its simplest form structured as Human > Animate > Inanimate. Exploring its interaction with other linguistic scales (e.g., the Person Hierarchy and the Referentiality Hierarchy), Croft (2002) has subsequently proposed the 'Extended Animacy Hierarchy', which adds 1st and 2nd person pronouns as the most animate. Within language typology, animacy has long functioned as a categorical device to account for language-specific phenomena. When viewed from this explanatory perspective, it is referred to as 'linguistic' (Radanović et al., 2016) or 'grammatical' (de Swart and van Bergen, 2019) animacy: a language-internal category that is explicitly encoded at the morpho-syntactic level and whose presence varies across languages (most of which lack it entirely). For example, in Russian, animate and inanimate direct objects are differentiated morphologically.

In contrast, we here examine animacy at its *semantic* level—as an 'extra-linguistic conceptual property' (Comrie, 1989) inherent to nominal referents and, therefore, treatable as a variable present in all languages. Following previous insights from de Swart and van Bergen (2019) and Bayanati and Toivonen (2019), we propose analysing animacy informed by the soft vs hard constraint distinction (Bresnan et al., 2001): what serves as a hard constraint and is explicitly grammaticalised in some languages may exist as a soft constraint—i.e., a statistical preference or tendency—in others. Thus, departing from the qualitative, fragmentary accounts of animacy provided by previous linguistics studies (primarily concerned with hard constraints), we study the role of animacy as a semantic property in shaping grammar (Haley et al., 2025) and whether this reflects a universal, cross-linguistic soft tendency.

This pursuit has remained unattainable until now due to the lack of multilingual data annotated for animacy. Hence, our main contributions are twofold. Firstly, we devise a method to classify animacy cross-linguistically by fine-tuning Large Language Models (LLMs)—such as mBERT (Devlin et al., 2019) and Aya (Dang et al., 2024)—on multilingual word sense disambiguation data, where senses are mapped to the classes of human, animate, and inanimate. In 11 out of 12 languages, we achieved an almost perfect macro F1 score of over 97. In turn, this newly created dataset unlocks the opportunity

[*]Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7349–7363

to study how animacy classes correlate with several aspects of linguistic form in 11 languages, using data extracted from treebanks annotated according to the Universal Dependencies (UD) framework.

From these studies, it emerges that higher animacy in nouns correlates cross-linguistically with both aspects of morphology (in particular, plural number marking) and syntax. In this case, we find that it correlates with earlier positions in a clause, the alignment role of Agent, and being a syntactic pivot (in particular, for relativisation); however, subjects with other roles, i.e., Sole and Patient, are considerably less animate. These observations further problematise the very notion of 'subject': not only, as argued by (Evans and Levinson, 2009), the typical bundle of semantic (animacy), syntactic (pivot), and discourse (topicality) properties of subjects finds exceptions in several languages: we find that their cross-linguistic association with animacy is dubious. In addition, our analysis shows that 1st and 2nd person pronouns, posited to be on top of the Extended Animacy Hierarchy, behave idiosyncratically with respect to different grammatical constructions: they are rarely plural or relativised, but often agents and in early positions.

In general, we hope that our quantitative framework will facilitate a broader exploration of how linguistic meaning influences grammatical form. We release our datasets, models, and code at `https://github.com/Naiina/animacy-annotations`.

2 Motivation and Research Questions

Theoretical linguistics and language typology have extensively examined animacy as a grammatical and semantic feature, primarily through qualitative, example-based analyses that account for the strict constraints animacy imposes in specific languages and linguistic phenomena. To our knowledge, only two studies have quantitatively investigated *semantic* animacy, both focusing exclusively on word order interactions: Thuilier et al. (2021), in a monolingual study on French, and Asadpour (2023), in a multilingual study on Armenian, Mukri Kurdish, Northeastern Kurdish, Jewish Northeastern Neo-Aramaic, and Azeri Turkic.

Filling this gap, this study provides the first large-scale quantitative investigation of cross-linguistic correlations of animacy, analysing 11 languages across multiple dimensions, pertaining both to morphology (specifically, plural marking) and syntax (word order, grammatical and alignment role, and relativisation patterns). In this section, we review hard animacy constraints on these constructions in individual languages, before studying whether they are reflected as soft, cross-linguistic tendencies in Section 4.

2.1 Animacy and Number

Typologically, number marking is not uniformly distributed across noun classes. In some languages, animacy has been proposed as a key factor influencing this variation. Specifically, plural marking tends to be more prevalent for nouns higher on an Animacy Hierarchy (Corbett, 2000; Smith-Stark, 1974). In these languages, a hard constraint has been postulated as underlying the hierarchy, stating that if a tier of the hierarchy allows plurals, then all levels to its left on the scale must also have them (Corbett, 2000; Haspelmath, 2013). For example, Smith-Stark (1974) observed that, in Georgian, verbs agree in number with animate subjects but default to singular when the subject is inanimate. Similarly, in Marind (Papuan), plural agreement is obligatory for humans and some animals but absent for inanimates (Corbett, 2000). Experimental findings further suggest a concurrent processing advantage for plural inflection in animate nouns (Zanini et al., 2020), mirroring the biological salience of the referents. Building on these observations, we aim to quantitatively assess whether a cross-linguistic soft constraint favours the pluralisation of more animate entities over inanimate ones, in languages where animacy is not an active grammatical category in number systems.

2.2 Animacy and Word Order

Moreover, animate entities tend to appear earlier in sentences than inanimate ones (Tomlin, 1986; Kempen and Harbusch, 2004; Branigan et al., 2008). In multiple languages, word order constraints reflect animacy-based hierarchies. In Navajo, arguments must be ordered by animacy, with animate referents always preceding less animate ones (Hale, 1973; Croft, 2002). Similarly, in Shona and Sesotho, conjoined noun phrases must follow an animacy-based order, with human referents appearing first, followed by non-human animates, and finally inanimates (Hawkinson and Hyman, 1974; Morolong and Hyman, 1977). Cognitive studies have focused on determining whether this preference for animate-first ordering is a structural byproduct of grammatical roles assignment (*indirect hypothesis,* McDonald et al., 1993) or an independent effect (*direct*

hypothesis, Feleki and Branigan, 1999). Without delving into the intricacies of the debate, we aim here to empirically test whether higher positions on the animacy hierarchy tend to correlate with earlier placement in argument ordering within clauses.

2.3 Animacy, Grammatical Role and Voice

The degree of animacy constrains also the choice of the subject under certain circumstances (Itagaki and Prideaux, 1985; Yamamoto, 1999). In Japanese, Lakota, and Jakaltek, for instance, active transitive verbs typically require an animate subject (Kuno, 1973; Van Valin, 1997; Craig, 1977). Similarly, in Korean, inanimate subjects are ungrammatical in most passive constructions (Song, 1987; Palmer, 1994). Being sentient and alive is generally regarded as a prerequisite for agency, which in turn is considered a core property of subjecthood (notably, in UD's guidelines, subjects are explicitly defined as the nominal *proto-agents* of a clause). Thus, by analysing the distribution of animacy across dependency relations (subject, object, oblique) and alignment roles (Agent, Sole, Patient) in both active and passive verbs, we will try to disentangle the association of animacy with semantic roles and grammatical roles. This will allow us to verify if a systematic preference for animate entities as subjects exists (Evans and Levinson, 2009).

2.4 Animacy and Relativisation

Another property of subjects is to act as syntactic pivots, such as the head of relative clauses. Experimental studies have consistently shown that subject relative clauses are processed more easily than object relative clauses (Holmes and O'Regan, 1981; Ford, 1983). This finding mirrors the proposal, long established in linguistic typology, of the Accessibility Hierarchy Hypothesis (Keenan and Comrie, 1977), which posits that subjects are the most accessible elements for relativization, followed by direct objects, obliques, and possessors. In Malagasy (Western Malayo-Polynesian), for example, only subjects can be relativised (Keenan and Comrie, 1977). Given the association between animacy and syntactic subjecthood (Section 2.3), it follows naturally that animate entities should be relativised more frequently than inanimate ones.

3 Neural Animacy Classification

To conduct our cross-linguistic study, we must obtain animacy information for nominals in multiple languages. After reviewing previous work (Section 3.1), we show how to extract animacy classes from word sense disambiguation datasets (Section 3.2). We then fine-tune pre-trained multilingual large language models as animacy classifiers (Section 3.3) and use them to annotate dependency treebanks automatically (Section 3.4).

3.1 Related Work

Following a distinction proposed by Bjerva (2014), the early literature on automatic animacy annotation falls into: (i) approaches exploiting corpus frequencies, such as co-occurrences with certain verbs, syntactic patterns, and pronoun coreferences (e.g., on Norwegian, Øvrelid, 2005; on Japanese and English, Baker and Brew, 2010); and (ii) approaches relying on lexico-semantic resources, such as WordNet (Miller, 1994) — whose 'unique beginners' have often been used as proxies for animacy classes (Orasan and Evans, 2007; Bloem and Bouma, 2013; Baker and Brew, 2010). Nevertheless, these approaches are restricted to type-level classification, lacking the ability to resolve ambiguous nouns that depend on contextual interpretation.

In recent years, the limitations of such feature-engineering-heavy approaches have prompted a shift towards neural methods (Zhu et al., 2019; Kiyomaru and Kurohashi, 2021). Nonetheless, these have remained focused on monolingual settings (on German, Klenner and Göhring, 2022; Tepei and Bloem, 2024, on Romanian) with multilingual studies largely limited to bilingual English-Japanese cases (Baker and Brew, 2010; Kiyomaru and Kurohashi, 2021). Furthermore, several of these required human annotation of animacy classes, often through crowdsourcing (e.g., Klenner and Göhring, 2022). Motivated by these limitations, our work seeks to expand beyond monolingual paradigms and minimise reliance on resource-intensive manual annotations.

3.2 Creating an Animacy-labelled Dataset

The scarcity of readily available animacy-annotated datasets limited prior works to narrow linguistic diversity and resource-intensive manual annotations. To address these obstacles, we derive a new animacy-annotated dataset from XL-WSD (Pasini et al., 2021), currently the largest multilingual dataset for Word Sense Disambiguation (WSD).

Choice of Base Dataset XL-WSD is a sense-annotated dataset providing gold-standard evalu-

ation and testing data for 18 languages across 6 language families, and silver training data for 15 languages. It consists of sentences where words of certain lexical classes (nouns, verbs, adjectives) are sense-annotated with a BabelNet synset (Navigli and Ponzetto, 2012)—that is, a multilingual set of synonyms that share the same meaning. Since the polysemy of nouns can span across animacy classes (e.g., "*bat*" is animate as an animal but inanimate as a sports implement), we can reasonably frame animacy annotation as a coarse-grained form of WSD. Procedurally, this implies the possibility of deriving an animacy-annotated dataset through a deterministic algorithm that maps each BabelNet synset to its corresponding animacy class.

Extraction Procedure We proceeded as follows. First, we manually selected, through a top-down inspection of WordNet 3.0's (Miller, 1994) directed graph, the highest-level synsets that could serve as significant proxies for animacy classes:

Human synsets: PERSON.N.01, OPERATOR.N.02, TEACHER.N.02, KIN.N.02, PEOPLE.N.01, ENEMY.N.01;

Animate synsets: LIVING_THING.N.01, BIOLOGICAL_GROUP.N.01, SPIRITUAL_BEING.N.01, IMAGINARY_BEING.N.01

Then, for each noun-tagged instance in XL-WSD, we mapped its annotated BabelNet synset to one from WordNet 3.0, then extracted its hypernym path,[1] i.e., the chain of synsets linking it with the root synset, ENTITY.N.01. Finally, we labelled the noun as Human or Animate if its hypernym path included a human or animate synset, respectively; otherwise, we labelled it as Inanimate. We excluded nouns with the synset GROUP.N.01 in their hypernym paths due to their high heterogeneity with respect to animacy gradation (e.g., "*state*", "*team*", "*traffic*").

Finally, given the highly skewed distribution of animacy labels toward the Inanimate class (with an average of 88.4% Inanimate nouns and only 1.9% Animate nouns in the test sets across languages), we rebalanced our dataset by adjusting label distributions within each set per language. Specifically, we removed sentences with only Inanimate labels originating from silver-quality WordNet glosses and examples (the largest source of sentences in the XL-WSD training data) or from other sources (SemCor, development, and test data) and selectively reintroduced them if necessary to achieve an approximate 2:1 ratio of Inanimate to (Animate + Human) labels. Finally, we grouped sentences by the number of animacy-labelled nouns and assigned them to the training (75%), test (15%), and validation (10%) sets while ensuring that each split maintained the same 2:1 target ratio of animacy labels. This resulted in the final label distribution across languages shown in 1. A more detailed breakdown of the dataset composition is available in Table 5 in the Appendix.

Set	N	A	H
Train	61.3	17.2	21.4
Dev	58.0	14.6	27.3
Test	60.0	13.8	26.1

Table 1: Distribution of animacy labels across data sets (%); N = Inanimate, A = Animate, H = Human.

3.3 Models

Afterwards, we train a model from our animacy-labelled dataset to classify the animacy of nominals in new sentences automatically. Formally, we cast this as a task where, given a sentence $\mathbf{s} = [w_1, \ldots, w_n]$ as context and a target word $w_t \in \mathbf{s}$, the model is required to assign one of the three predefined animacy classes to w_t: namely, Inanimate, Animate, Human. With this goal, we conducted experiments using two modular and parameter-efficient fine-tuning techniques (Pfeiffer et al., 2023) on two distinct open-access multilingual models.

- Lottery Ticket Sparse Fine-Tuning (LT-SFT; Ansell et al., 2022), leveraging pre-trained language SFT adapters readily available, on Multilingual BERT (mBERT; Pires et al., 2019), an encoder language model;
- LoRA (Hu et al., 2022) instruction-tuning on Aya Expanse 8B, a decoder language model.

We considered these models for two main reasons. First, they are suitable for different subsets of languages. Hence, combining them increases our multilingual coverage. Second, this allows us to compare fine-tuning performances of models with different architectures (encoder vs. decoder) and sizes (110M vs. 8B parameters).

[1] nltk.org/howto/wordnet.html

ISO code	Language	Family	Inanimates (N)	Non-human animates (A)	Humans (H)	1st/2nd pers. pronouns (P)
DE	German	IE, Germanic	38,099	1220	6,854	855
EN	English	IE, Germanic	22,375	849	3,085	5,881
ES	Spanish	IE, Romance	67,672	927	12,138	1,098
ET	Estonian	Uralic, Finnic	74,080	2,543	11,519	3,000
EU	Basque	Basque	14,831	189	2,756	12
FR	French	IE, Romance	5,623	122	969	115
IT	Italian	IE, Romance	44,177	792	7,704	660
JA	Japanese	Japanese	41,329	807	5,673	-
NL	Dutch	IE, Germanic	30,075	367	6,278	472
SL	Slovenian	IE, Slavic	45,595	5,846	1,977	920
ZH	Chinese	Sino-Tibetan	22,658	650	3,650	79

Table 2: Languages included in our study and their family. For each language, we report the number of nominals annotated in each category of the Extended Animacy Hierarchy, arranged in order of increasing animacy.

3.4 Training and Prediction

To align with the distinct training regimes of the two models, we framed animacy annotation as a token classification task for mBERT—given its simple pre-training on unlabelled data—and as a conditional generation task for Aya—given its supervised fine-tuning for instructions and conversation. In the first case, we trained a task-specific SFT on mBERT, augmented with a standard single-layer classifier head, in two settings: i) monolingual, using English as the sole source language, and ii) multilingual, with a different source language in each batch (and the associated language-specific adapter activated during the forward pass). In contrast, for Aya, where we do not need to compose separate task- and language-specific adapters, we applied LoRA fine-tuning directly on mixed-language batches, such that the adapter parameters and updates are shared across languages. We report the full hyperparameter configuration in Appendix B.

We present the test set results in Table 3 for the 12 languages that have gold-standard annotation in XL-WSD and are also covered by either mBERT or Aya. From Table 3, it emerges that multi-source fine-tuning is crucial for both mBERT and Aya. These achieve almost perfect F1 scores, except Aya lags slightly behind in EN and ZH, whereas mBERT in JA and KO.

After training, we performed inference using UD sentences as context (s) and UD tokens POS-tagged as NOUN as target words (w_t). For each language, we selected the best-performing model based on its coverage and evaluation metrics found in Table 3. Finally, we annotated a fourth animacy class on UD words labelled as 1st and 2nd person pronouns.

	mBERT		**Aya**	
	single	multi	zero-shot	fine-tuned
DE	95.4	**98.7**	47.0	**98.7**
EN	97.8	**98.1**	59.0	94.5
ET	90.8	**98.7**	-	-
ES	95.5	97.7	53.0	**97.9**
EU	91.7	**98.0**	-	-
FR	96.5	**99.0**	46.2	98.8
IT	-	-	57.3	**98.3**
JA	88.3	95.2	48.8	**97.6**
KO	48.5	65.2	53.2	**77.9**
NL	-	-	45.9	**98.0**
SL	-	-	41.7	**97.1**
ZH	82.9	**91.9**	60.2	83.3

Table 3: Macro F1 scores for animacy classification on the test sets of 12 languages. The best-performing model is shown in bold.

This annotation pipeline yields a unified dataset that integrates animacy classification with morpho-syntactic annotations derived from the treebanks. The number of annotated nominals in each language from the UD-derived datasets is listed in Table 2. We use this data for the linguistic study in Section 4. Crucially, none of those languages have hard constraints on animacy for the properties we investigate.

To assess the classifier's performance on UD, native speakers of each language manually annotated 102 UD sentences, each containing a noun drawn

	Accuracy			F1
	H	A	N	
DE	94.12	80.65	100.00	91.9
EN	96.97	78.12	100.00	91.6
ET	82.35	57.58	77.42	72.3
ES	100.00	76.67	91.18	89.6
EU	93.94	66.67	100.00	86.4
FR	97.06	83.87	97.06	92.8
IT	100.00	85.29	97.06	94.1
JA	91.18	87.50	91.18	90.1
NL	94.12	38.24	100.00	74.5
SL	93.75	42.42	87.88	73.0
ZH	79.41	67.65	97.06	81.1

Table 4: Accuracy and F1 of our classifier on UD data according to manual annotations from native speakers.

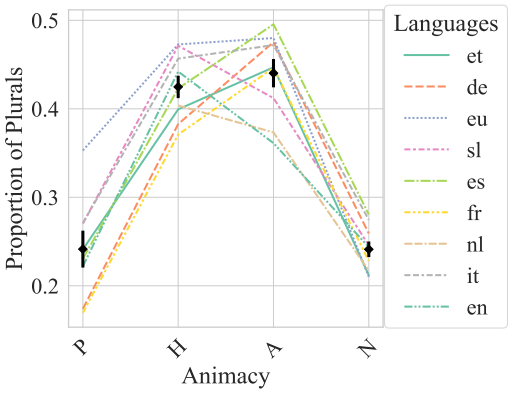


Figure 1: Proportion of plural nominals for each animacy class. The black diamonds with error bars indicate the mean and standard error across languages.

uniformly across animacy classes (34 sentences per class). The results, presented in Table 4, combined with the global confusion matrix reported in the Appendix C.2, confirm the high accuracy of our classifier on UD for human and inanimate classes. On the other hand, non-human animates have the lowest accuracy. This is because this class contains several borderline referents (e.g., plants or microorganisms) that humans struggle to reliably associate with animate or inanimate classes (Westbury, 2023; Radanović et al., 2016). Moreover, performance may be degraded compared with XL-WSD (Table 3) because of possible domain shifts. Therefore, it remains unclear whether the misclassifications are due to errors from humans or the neural classifier. In any case, the fact that the boundaries of this class are blurred could partly explain the mixed results for non-human animates in terms of correlations with morphosyntactic phenomena (see Section 4).

4 Linguistic Study

In this section, we address the research questions raised in Section 2 based on the dataset created via animacy classification in Section 3. We investigate whether the animacy hierarchy is correlated with morphological features, such as number marking (Section 4.1). Next, we examine how animacy interfaces with the order of the arguments in a clause (Section 4.2), grammatical and semantic roles (Section 4.3), and syntactic pivots in relative clauses (Section 4.4). These features are known to be correlated with each other: most notably subjects tend to be topical (so appear early), agentive, and pivots (Evans and Levinson, 2009). In our study, we aim to disentangle these different axes and determine if all levels of the hierarchy behave consistently for all these aspects of morphosyntax.

4.1 Number Marking

We aim to understand if plural number correlates with higher animacy. To this end, we computed the proportion of plural nominals for each animacy class, as shown in Figure 1. We excluded languages without number marking, such as Japanese and Chinese. For Basque, we included only the subset of nouns with number annotation in UD.

In all languages, the proportion of plural nouns referring to H and A is considerably higher than the proportion of plural nouns referring to P and N. Running a Welch's t-test, we find all pairwise differences between animacy levels to be significant with $p < 10^{-5}$, except for P vs N and H vs A.

This cross-linguistic tendency may be the origin of the phenomenon described in Section 2.1, such that in some languages number marking and verb agreement are restricted to entities that are high on the animacy scale (Corbett, 2000); however, Ps are an exception as they are rarely plural but highest on the scale for marking differentiation.

4.2 Order of the Arguments within Clauses

Given hard constraints on word order, we next investigate if there is a tendency for arguments of a predicate to appear earlier in each clause within a sentence if they are more animate. We divide each sentence from UD into clauses. A clause corresponds to the words spanned by a dependency

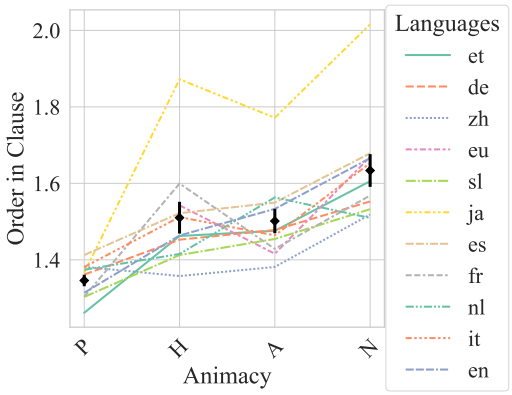
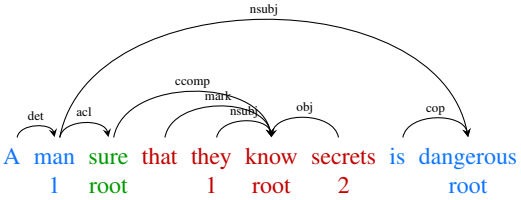


Figure 2: Average order of first and second person pronouns (P), humans (H), non-human animates(A) and inanimates (N) within a clause. The mean and standard error for each class are also shown.

subtree whose root is a predicate. More specifically, a word is identified as a predicate (i.e., the root of a subtree) if:

- It is the root of the whole sentence;
- It has a dependency relation of coordinate clause (annotated as CONJ in UD) with another root; or
- It has a dependency relation of clausal subject (CSUBJ), open clausal complement (XCOMP), clausal complement (CCOMP), clausal modifier of noun (ACL), or parataxis (PARATAXIS).

For each clause, we computed the average order of each animacy level (P, H, A, and N) if they were arguments of the predicate (including NSUBJ, OBJ, IOBJ, OBL). In this example:



"*A man is dangerous*", "*sure*", "*that they know secrets*" are the three clauses of the sentence, where ROOT identifies the root of their subtree. In the first clause (blue), the argument positions are ("*man*": 1), whereas in the third clause (red), the argument positions are ("*they*": 1, "*secrets*": 2).

As shown in Figure 2, for all the languages we investigated, H occurs on average earlier than N, and P in turn occurs earlier than H. Interestingly, the trend for non-human animates is less clear. In fact, running a Welch's t-test between animacy levels, we find that differences are all statistically significant with $p < 0.05$ except for H vs A and H vs N. According to Table 4 and Figure 6, A is the class with the least agreement between human and automatic annotations. This class is less intuitive for humans to label (see Appendix C) and reflects a continuum, which might explain why A is indistinguishable from H in terms of expected word order. Another factor could be that A is the class with the fewest examples in the dataset, which could induce distortions. Finally, the average order in Japanese tends to be higher across animacy levels, possibly due to being topic-first.

Several linguistic hypotheses could explain these results. The speakers' egocentric perspective, or "human first" principle (Dahl, 2008; Meir et al., 2017) could explain why P and H tend to be mentioned first. Moreover, language production is an incremental process (Garrett, 1980; Dell and Reich, 1981; Kempen and Hoenkamp, 1987). Therefore, more accessible entities, especially P and H (Bondoc and Schafer, 2022) might appear earlier in clauses (Futrell, 2024).

4.3 Grammatical and Semantic Roles

In this subsection, we assess the two hypotheses from Section 2.3, which link animacy with agentic semantic roles (e.g., subjects of transitive verbs and oblique agents of passive verbs) and grammatical roles (subjects in general).

Hence, we computed the proportion of nominals of a certain animacy level for each of the following dependency relations and corresponding semantic roles (Agent, Sole, and Patient): subjects of an active verb (NSUBJ, treated as Agent when transitive and Sole when intransitive) direct object of an active verb (OBJ, a Patient), subjects of a passive verb (NSUBJ:PASS, a Patient), and oblique agent of a passive verb (OBL:AGENT, an Agent). Note that some languages (EU, JA, and SL) do not feature a passive voice or this is not annotated, in which case, the last two relations are excluded.

For each language, we computed the Pointwise Mutual Information (PMI) between animacy and each role, which is a measure of (positive or negative) association between these two variables. We report PMI rather than proportions as the uneven frequencies of different roles may otherwise obscure the trends. The results are shown in Figure 3: in every language of the corpus, H has high PMI with agent roles (either subjects of active transitive verbs, or oblique agents of passive sentences), and low PMI with patients (objects of active sentences or subjects of passive sentences). Vice versa,

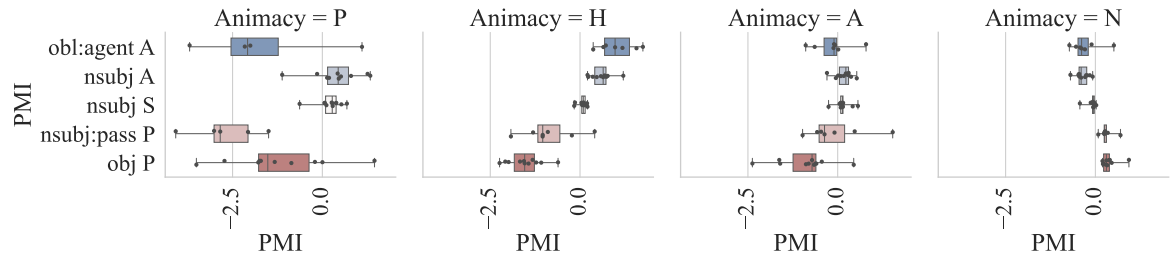


Figure 3: Each boxplot refers to an individual animacy level and indicates the PMI between an animacy class and a grammatical/semantic role across languages. Association is positive when PMI > 0 and negative when PMI < 0. Humans (H) associate with high agency roles, inanimates (N) with low agency.

the opposite (yet, milder) trend is observed for N. Therefore, we confirm a strong link between agentive semantic roles and animacy. P again behaves erratically, showing very inconsistent trends across languages, with a remarkably high variance for the roles of oblique Agent and object Patient.

Moreover, we found that subjects of passive verbs are distributed similarly to direct objects. In this sense, the link proposed within the literature (Bock et al., 1992; McDonald et al., 1993; Ferreira, 1994; Prat-Sala, 1997) between animacy and subjecthood is quite weak.

4.4 Relativisation

Finally, we study if animate entities tend to act more often than inanimate ones as syntactic pivots, and specifically as the head of relative clauses. To identify these, we select both i) nominals with a dependent with an ACL:REL relation, when this annotation is available (i.e., in EN, ET, FR, IT, NL, and ZH); or ii) heads of pronouns with PRONTYPE=REL annotations, in other languages. We excluded Japanese and Basque because neither annotation is available in UD, and their relative clauses (relying on strategies based on word order and morphology) are not amenable to being easily extracted automatically.

We plot the results in Figure 4. We observe that nouns referring to humans (H) are more relativised on average than nouns referring to inanimates (N) in all languages except Chinese. 1st and 2nd person pronouns are rarely relativised, as expected. To corroborate this, we ran a Welch's t-test and found statistically significant differences (with $p < 0.05$) for all animacy levels except for H vs A and A vs N: this is probably due to similar arguments presented in Section 4.2. This pattern is connected with the preference for relativising subjects over

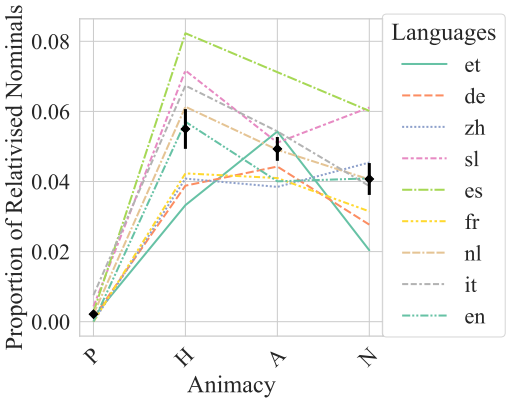


Figure 4: Proportion of nominals which are relativised by animacy class in 10 languages, along with the mean and standard error for each class.

other arguments discussed in Section 2.4. Since, as found in Section 4.3, subjects of active verbs tend to be more animate, it follows that relativised nouns are more frequently animate.

4.5 General Discussion

The findings from Section 4 hint at a broad connection between animacy and several grammatical features. Most notably, we empirically verified that there exists a correlation between high animacy in nouns and plural number marking (Section 4.1), early positions in a clause (Section 4.2), Agent roles such as subjects of a transitive active verb or obliques of a passive verb (Section 4.3), and syntactic pivots for relative clauses (Section 4.4). Nevertheless, the correlation with grammatical roles is weaker. In particular, animacy clearly declines in grammatical subjects of intransitive and passive verbs. This further complicates the 'universality' of the notion of subjecthood, which often conflates aspects related to syntax, semantics, and information

structure (Comrie, 1989; Aissen, 2001).

We also found that the cross-linguistic tendencies observed in *nouns* in Section 4 are ultimately reflected by the hard grammatical constraints in several languages identified in Section 2; however, this is not always the case for *1st and 2nd person pronouns*. While the Extended Animacy Hierarchy (Corbett, 2010) places them at the top of the hierarchy, they are rarely marked as plural or relativised.

Finally, while the tendencies we discovered are often clear-cut for humans and inanimates, the animate class seems more ambivalent. In fact, humans and automatic classifiers tend to disagree more on nouns from this class (see Figure 6 and Table 4), as it lies on a continuum without clear boundaries. This may partly explain why the effects on grammar structures of humans and animates are sometimes indistinguishable, especially for word order.

In future work, our animacy dataset may provide new insights into cognitive constraints on sentence production, most notably accessibility or the minimisation of surprisal (Futrell et al., 2020). Animate entities are known to be more accessible and influence the difficulty of processing parts of a sentence, e.g., relative clauses (Mak et al., 2002). Since language production is an incremental process (Garrett, 1980; Dell and Reich, 1981; Kempen and Hoenkamp, 1987), more accessible entities are more likely to appear early in sentences (Kathryn Bock and Irwin, 1980) and to be mapped into the grammatical role of subject (Branigan et al., 2008). This is consistent with our findings, which could be corroborated and expanded with cognitive and behavioural studies in the future.

5 Conclusions

We have proposed a method to automatically classify the semantic animacy of nominals in a set of 12 languages from 5 families. Relying on multilingual LLMs and word sense disambiguation datasets, we achieve a macro F1 score of 97 and above for 11 languages. This information unlocks new opportunities for the study of the role of animacy in grammatical constructions. We find that high animacy in nouns correlates with plural marking, early positions (connected with topicality), grammatical roles of Agent (subject of transitive verbs or oblique of passive verbs), and being a pivot for relative clauses. However, subjects of intransitive and passive verbs are substantially less animate. This indicates that animacy is more tightly connected to the semantic role of Agent than to the grammatical role of subject. Moreover, our results cast doubts over the hypothesis that hard animacy constraints in some languages reflect universal cross-linguistic tendencies: personal pronouns are believed to be the most animate entities, and yet they are infrequently marked as plural or relativised.

Acknowledgements

We would like to sincerely thank Coleman Haley for his valuable feedback, and all the anonymous reviewers for their comments. MG acknowledges the Institute for Advanced Study of Pavia and Ghislieri College for their financial support, which enabled his research stay at the University of Edinburgh. Computational resources for this work were provided by the Edinburgh International Data Facility (EIDF) and the Edinburgh Compute and Data Facility (ECDF).

Limitations

The limitations of our work lie in three key aspects: the evaluation of automatic animacy annotation in UD, the typological diversity of our sample of languages, and the impossibility of extracting information for all levels of the animacy hierarchy.

First, despite high macro F1 scores, the domain discrepancy between XL-WSD and UD data may degrade performance due to domain shift. Additionally—beyond manual annotation of subsamples by native speakers—we were not able to recruit human validators for the entirety of the UD animacy annotation, due to resource constraints.

Second, although our language sample spans five language families, it is skewed towards Indo-European languages (7 out of 11). The remaining four families are each represented by a single language. This imbalance reflects the typological distribution in XL-WSD (12 Indo-European out of 18), inherently limiting our approach as long as it depends on this dataset. In principle, the LLM animacy classifiers are suitable for all languages covered by their pretraining data; however, it would be impossible to validate their reliability in languages absent from XL-WSD.

Third, we could only extract information for 1st and 2nd person pronouns, but not for 3rd person ones. This is because in many languages this would require co-reference resolution to determine the animacy of the referent. We also excluded proper nouns from our analyses for a similar reason.

References

Judith Aissen. 2001. Markedness and subject choice in optimality theory. In *Optimality-Theoretic Syntax*. The MIT Press.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Hiwa Asadpour. 2023. A corpus analysis of the effects of definiteness and animacy on word order variation. *Languages*, 8(4).

Kirk Baker and Chris Brew. 2010. Multilingual animacy classification by sparse logistic regression. *Ohio State Dissertations in Linguistics (OSDL)*, 52:52–75.

Shiva Bayanati and Ida Toivonen. 2019. Humans, animals, things and animacy. *Open Linguistics*, 5(1):156–170.

Johannes Bjerva. 2014. Multi-class Animacy classification with semantic features. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–75, Gothenburg, Sweden. Association for Computational Linguistics.

Jelke Bloem and Gosse Bouma. 2013. Automatic animacy classification for Dutch. *Computational Linguistics in the Netherlands Journal*, 3:82–102.

Kathryn Bock, Helga Loebell, and Randal Morey. 1992. From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review*, 99(1):150–171.

Ivan Paul Bondoc and Amy J. Schafer. 2022. Differential effects of agency, animacy, and syntactic prominence on production and comprehension: Evidence from a verb-initial language. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 76(4):302–326.

Holly P. Branigan, Martin J. Pickering, and Mikihiro Tanaka. 2008. Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, 118(2):172–189. Animacy, Argument Structure, and Argument Encoding.

Joan Bresnan, Shipra Dingare, and Christopher D. Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG01 Conference*, Hong Kong. CSLI Publications.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Greville G. Corbett. 2000. *Number*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Greville G. Corbett. 2010. Implicational hierarchies. In *The Oxford Handbook of Linguistic Typology*. Oxford University Press.

Colette Grinevald Craig. 1977. *The Structure of Jacaltec*. University of Texas Press, Austin.

William Croft. 2002. *Typology and Universals*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.

Östen Dahl. 2008. Animacy and egophoricity: Grammar, ontology and phylogeny. *Lingua*, 118(2):141–150. Animacy, Argument Structure, and Argument Encoding.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*.

Peter de Swart and Geertje van Bergen. 2019. How animacy and verbal information influence V2 sentence processing: Evidence from eye movements. *Open Linguistics*, 5(1):630–649.

Gary S. Dell and Peter A. Reich. 1981. Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, 20(6):611–629.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–492.

Elina Feleki and Holly Branigan. 1999. Conceptual accessibility and serial order in Greek speech production. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society*, pages 96–101. Lawrence Erlbaum Associates.

Fernanda Ferreira. 1994. Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language*, 33(6):715–736.

Marilyn Ford. 1983. A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior*, 22(2):203–218.

Richard Futrell. 2024. An information-theoretic account of availability effects in language production. *Topics in Cognitive Science*, 16(1):38–53.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.

Merrill F. Garrett. 1980. Levels of processing in sentence production. In Brian Butterworth, editor, *Language Production, Vol. 1: Speech and Talk*, pages 177–220. Academic Press, London.

Kenneth Hale. 1973. A note on subject-object inversion in Navajo. In Braj B. Kachru, Robert B. Lees, Yakov Malkiel, Angelina Pietrangeli, and Sol Saporta, editors, *Issues in Linguistics: Papers in Honor of Henry and Renée Kahane*, pages 300–309. University of Illinois Press, Urbana.

Coleman Haley, Sharon Goldwater, and Edoardo Ponti. 2025. A grounded typology of word classes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10380–10399, Albuquerque, New Mexico. Association for Computational Linguistics.

Martin Haspelmath. 2013. Occurrence of nominal plurality (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Ann Katherine Hawkinson and Larry Michael Hyman. 1974. Hierarchies of natural topic in Shona. *Studies in African Linguistics*, 5(2):147–170.

V.M. Holmes and J.K. O'Regan. 1981. Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior*, 20(4):417–430.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nobuya Itagaki and Gary D. Prideaux. 1985. Nominal properties as determinants of subject selection. *Lingua*, 66(2-3):135–149.

June Kathryn Bock and David E. Irwin. 1980. Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior*, 19(4):467–484.

Edward L. Keenan and Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8(1):63–99.

Gerard Kempen and Karin Harbusch. 2004. A corpus study into word order variation in german subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In Thomas Pechmann and Christopher Habel, editors, *Multidisciplinary Approaches to Language Production*, pages 173–182. De Gruyter Mouton, Berlin, New York.

Gerard Kempen and Edward Hoenkamp. 1987. An incremental procedural grammar for sentence formulation. *Cognitive science*, 11(2):201–258.

Hirokazu Kiyomaru and Sadao Kurohashi. 2021. Contextualized and generalized sentence representations by contrastive self-supervised learning: A case study on discourse relation analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5578–5584, Online. Association for Computational Linguistics.

Manfred Klenner and Anne Göhring. 2022. Animacy denoting German nouns: Annotation and classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1360–1364, Marseille, France. European Language Resources Association.

Susumu Kuno. 1973. Constraints on internal clauses and sentential subjects. *Linguistic Inquiry*, 4(3):363–385.

Willem M. Mak, Wietske Vonk, and Herbert Schriefers. 2002. The influence of animacy on relative clause processing. *Journal of Memory and Language*, 47(1):50–68.

Janet L. McDonald, Kathryn Bock, and Michael H. Kelly. 1993. Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive psychology*, 25(2):188–230.

Irit Meir, Mark Aronoff, Carl Börstell, So-One Hwang, Deniz Ilkbasaran, Itamar Kastner, Ryan Lepic, Adi Lifshitz Ben-Basat, Carol Padden, and Wendy Sandler. 2017. The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. *Cognition*, 158:189–207.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Malillo Morolong and Larry Hyman. 1977. Animacy, objects and clitics in Sesotho. *Studies in African Linguistics*, 8:199–218.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Constantin Orasan and Richard Evans. 2007. NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research*, 29(1):79–103.

Frank Robert Palmer. 1994. *Grammatical Roles and Relations*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Ponti. 2023. Modular deep learning. *Transactions on Machine Learning Research*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Mercè Prat-Sala. 1997. *The production of different word orders: A psycholinguistic and developmental approach*. Ph.d. thesis, The University of Edinburgh, Centre for Cognitive Science.

Jelena Radanović, Chris Westbury, and Petar Milin. 2016. Quantifying semantic animacy: How much are words alive? *Applied Psycholinguistics*, 37(6):1477–1499.

Thomas Cedric Smith-Stark. 1974. The plurality split. *Chicago Linguistic Society*, 10:657–661.

Nam Sun Song. 1987. Empathy-based affectedness and passivisation. *Transactions of the Philological Society*, 85(1):74–89.

Maria Tepei and Jelke Bloem. 2024. Automatic animacy classification for Romanian nouns. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1825–1831, Torino, Italia. ELRA and ICCL.

Juliette Thuilier, Margaret Grant, Benoît Crabbé, and Anne Abeillé. 2021. Word order in French: The role of animacy. *Glossa: A Journal of General Linguistics*, 6(1):55.

Russell S. Tomlin. 1986. *Basic word order*, 1st edition. Routledge.

Roberts Van Valin. 1997. *Syntax: Structure, meaning and function*. Cambridge Textbooks in Linguistics.

Chris Westbury. 2023. Why are human animacy judgments continuous rather than categorical? A computational modeling approach. *Frontiers in Psychology*, 14:1145289.

Mutsumi Yamamoto. 1999. *Animacy and reference: A cognitive approach to corpus linguistics*. Companion series. J. Benjamins Pub.

Chiara Zanini, Rosa Rugani, Dunia Giomo, Francesca Peressotti, and Francesca Franzon. 2020. Effects of animacy on the processing of morphological number: A cognitive inheritance? *Word Structure*, 13(1):22–44.

Yuanqing Zhu, Wei Song, Xianjun Liu, Lizhen Liu, and Xinlei Zhao. 2019. Improving anaphora resolution by animacy identification. In *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 48–51.

Lilja Øvrelid. 2005. Animacy classification based on morphosyntactic corpus frequencies: Some experiments with norwegian nouns. In *Proceedings of the Workshop on Exploring Syntactically Annotated Corpora*, pages 24–35.

A Statistics of the Animacy-labelled Dataset

Table 5: Label distributions across languages and splits. For each split, the raw number of sentences is provided. The number of labels per sentence varies based on the number of annotated target tokens in XL-WSD.

Lang	Split (Sentences)	A % (Raw)	H % (Raw)	N % (Raw)
de	Train (23,406)	19.14 (8,729)	22.57 (10,291)	58.28 (26,575)
	Dev (3,135)	14.82 (918)	30.63 (1,898)	54.55 (3,380)
	Test (4,676)	17.41 (1,581)	30.45 (2,765)	52.14 (4,734)
en	Train (43,818)	16.94 (29,074)	21.75 (37,339)	61.31 (105,244)
	Dev (5,873)	16.43 (3,848)	27.01 (6,325)	56.56 (13,246)
	Test (8,756)	11.89 (4,069)	26.45 (9,051)	61.66 (21,096)
es	Train (29,698)	13.84 (10,709)	20.06 (15,523)	66.10 (51,140)
	Dev (3,976)	9.39 (982)	23.83 (2,492)	66.78 (6,984)
	Test (5,931)	12.60 (1,940)	26.90 (4,143)	60.50 (9,318)
et	Train (22,460)	21.62 (8,309)	23.67 (9,097)	54.72 (21,034)
	Dev (3,006)	14.89 (777)	35.20 (1,837)	49.91 (2,605)
	Test (4,488)	24.90 (1,908)	28.81 (2,208)	46.29 (3,548)
eu	Train (26,075)	18.59 (9,579)	22.31 (11,492)	59.10 (30,450)
	Dev (3,488)	10.23 (711)	32.38 (2,250)	57.38 (3,987)
	Test (5,210)	14.37 (1,475)	29.24 (3,002)	56.40 (5,791)
fr	Train (25,958)	19.46 (11,191)	19.94 (11,468)	60.59 (34,843)
	Dev (3,476)	19.29 (1,505)	14.29 (1,115)	66.41 (5,181)
	Test (5,185)	17.02 (1,949)	22.16 (2,538)	60.82 (6,965)
it	Train (30,811)	14.89 (12,032)	19.32 (15,610)	65.79 (53,168)
	Dev (4,122)	18.54 (2,022)	31.27 (3,411)	50.20 (5,476)
	Test (6,157)	10.40 (1,676)	23.64 (3,810)	65.96 (10,630)
ja	Train (8,479)	9.49 (877)	33.51 (3,098)	57.00 (5,269)
	Dev (1,136)	9.76 (122)	36.80 (460)	53.44 (668)
	Test (1,694)	13.29 (245)	32.66 (602)	54.04 (996)
ko	Train (3,405)	0.68 (23)	5.43 (185)	93.89 (3,197)
	Dev (454)	0.88 (4)	6.39 (29)	92.73 (421)
	Test (681)	0.73 (5)	5.73 (39)	93.54 (637)
nl	Train (32,241)	18.26 (13,539)	21.83 (16,188)	59.91 (44,429)
	Dev (4,315)	13.28 (1,335)	27.12 (2,727)	59.60 (5,993)
	Test (6,441)	11.52 (1,702)	26.30 (3,884)	62.18 (9,184)
sl	Train (21,966)	21.53 (7,915)	24.40 (8,969)	54.06 (19,872)
	Dev (2,939)	13.88 (692)	28.95 (1,443)	57.16 (2,849)
	Test (4,390)	17.64 (1,293)	23.52 (1,724)	58.83 (4,312)
zh	Train (3,158)	2.99 (100)	8.47 (283)	88.54 (2,959)
	Dev (430)	4.76 (24)	8.33 (42)	86.90 (438)
	Test (629)	3.95 (26)	8.81 (58)	87.23 (574)

B Hyperparameter Configuration of the Neural Animacy Classifier

For LT-SFT, we followed hyperparameters from Ansell et al. (2022): 3 epochs of full fine-tuning and 10 epochs of sparse fine-tuning (batch size $= 8$, learning rate $= 5e-5$). For LoRA[2] (1 epoch, $\alpha = 64$, $r = 32$, batch size $= 6$, learning rate $= 3e-5$), we used mixed precision (bfloat16) with 4-bit quantization and accelerated training with FlashAttention (Dao, 2024).

C Animacy Annotation Experiment

C.1 Annotation guidelines

A pilot experiment showed that WordNet-derived animacy classes were not intuitive for humans. Despite precise guidelines and examples, annotators tend to misclassify non-animal animates, especially imaginary

[2]Implemented with PEFT: https://github.com/huggingface/peft.

In this experiment, you will read sentences and you will be asked to identify the animacy category of a word highlighted in bold. The categories are the following:
Human: Human beings, names of professions, groups of humans.
Imaginary being: Fictional or mythical creatures (e.g., witches, angels, giants).
Animal: All animals, excluding human beings.
Plant/bacteria: Entities like plants, flora, fungi, microorganisms, etc.
Inanimate: Non-living entities, objects, abstract ideas.
Unclear: Cases where the correct category cannot be confidently determined.

Here are some examples:
Please read carefully, as some examples might not be intuitive
Human:
The **team** scored a goal.
Being **doctor** is a stressful job.
Imaginary being
Believe me, she is a **witch**!
The painter depicted **angels** flying in the sky.
Animal:
This **species** is at risk of extinction!
The kid was riding an **elephant**!
Plant/bacteria:
The **grass** is green.
Microbes are found everywhere on Earth.
Strawberries are red.
Cells are the building blocks of life.
Inanimate
Justice had to be done.
The **wind** was surprisingly cold!

Figure 5: Guidelines for human animacy annotation on subsets of UD.

beings, trees, and bacteria. Moreover, most annotators reported that some examples were ambiguous. Therefore, we asked people to classify entities into one of the following classes: Human, imaginary being, animal, plant/bacteria, inanimate, unclear. We merged imaginary beings, animals, and plants/bacteria post hoc into a single category corresponding to animates to align them with our linguistic study.

For each of the three animacy categories (H, A, N), we extracted 34 sentences from UD. Sentences were randomly shuffled. Relying on PCIbex as an annotation platform, we asked one native speaker per language to annotate the animacy of a target word (displayed in bold) in each sentence, for each of the 102 sentences. The guidelines provided to our animacy annotators are shown in Figure 5. Note that provided guidelines are in English: to avoid misinterpretations, we recruited bilingual annotators with high English proficiency (among researchers from the University of Edinburgh).

C.2 Confusion matrix

We report the confusion matrix between humans and the automatic classifier in Figure 6, which shows generally strong agreement between the automatic classifier and human annotators in identifying animacy categories. Most notably: H (Human) and N (Inanimate) are correctly classified most of the time. On the other hand, A (Animate non-human) shows more confusion, particularly with 54 instances misclassified as H and 56 as N. This confirms what was observed during the pilot experiment: A is more ambiguous and less straightforward to classify. This likely reflects the fact that semantic animacy is a continuum, with animals being more animate (sometimes conflated with H) than plants and bacteria (sometimes conflated with N).

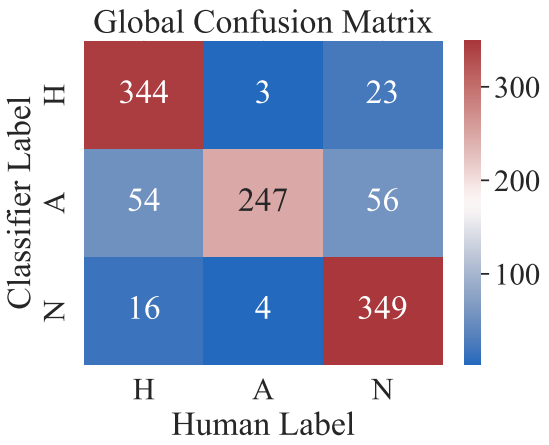


Figure 6: Confusion matrix comparing automatic and human animacy classification, aggregated across languages.