Comparable corpora in cross-linguistic research: Nominal number in English, Czech, and Greek

Konstantinos Diamantopoulos, Magda Ševčíková
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
{diamantopoulos, sevcikova}@ufal.mff.cuni.cz

Abstract

The paper examines the use of comparable corpora for contrastive research on the category of nominal number across three languages—English, Czech, and Greek. Two objectives are pursued: a cross-linguistic analysis of number and an assessment of the impact of automatic annotation on linguistic findings. For this study, corpora of comparable size and composition were compiled for the three languages from the Leipzig Corpora Collection. The data were automatically annotated using two open-access tools, Stanza and UDPipe, producing six datasets (two per language), each containing about 5 million sentences and 100 million tokens. Although derived from the same source, the paired datasets for each language differ in sentence and word segmentation, in the number of nouns identified, and in the number values assigned. These differences, nevertheless, do not appear to substantially affect the overall picture of number in the languages examined. The distribution of lemmas by the ratio of singular and plural forms challenges the view commonly presented in grammars that most nouns occur in both numbers and that singular-only and plural-only nouns are rare. However, a closer analysis of nouns assumed to have defective number indicates that answers to more nuanced questions vary depending on the annotation tool used.

Keywords: number, comparable corpora, morphosyntactic annotation, English, Czech, Greek

1. Introduction

The comparison of morphological categories across languages has long been central to contrastive and typological research. Information on morphological features has been collected for many languages in typological databases. However, recent advances in cross-linguistic research on word order have shown that the verification of distinctions recorded in such databases against corpus data leads to a more realistic picture (e.g. Choi et al. 2021; Levshina et al. 2023; Jing et al. 2023). In this respect, cross-linguistic research on morphology has lagged behind syntactic research in using corpus data, possibly due to the limited availability of suitable resources.

The present paper examines the morphological category of number in nouns in three languages with different morphological profiles, namely English, Czech, and (Modern) Greek. The languages were selected based on the availability of the required resources and tools, and on the availability of native speakers' judgments, ensuring full control over both the data and their interpretation. In order to avoid the problems that monolingual corpora (varied size, time period covered, etc.) and parallel corpora (translationese, etc.) pose for cross-linguistic comparisons, we compiled corpora of comparable size and composition for the three languages from the Leipzig Corpora Collection. However, the aim is not only to compare the category of number, but also to assess the potential of automatically annotated data for such research. To this end, the corpora for each language are annotated using two open-access tools Stanza and UDPipe.

The paper is structured as follows. Section 2 outlines basic facts about the category of nominal number, first as a cross-linguistically attested category and then with specific reference to the three languages examined. Section 3 describes the construction and morphosyntactic annotation of the comparable corpora and discusses in particular the differences between the paired datasets for each language. Section 4 presents the quantitative and qualitative analysis of the data aimed at identifying both language-internal and cross-linguistic patterns in the use of number. The results of the study are summarized in Section 5.

2. Nominal Category of Number

2.1. Cross-linguistic attestation

Grammatical number is a fundamental morphological category across the world's languages. However, as Corbett (2000, p. 2) notes, although it is often regarded as a simply structured category with singular and plural values and overt marking on nouns, one or more of these assumptions may not hold universally. Typological databases provide a broad-coverage view of some aspects: Of the 291 languages for which the World Atlas of Language

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 30–40
11 May 2026.

Structures (WALS; Dryer and Haspelmath 2013) reports on plural marking, 28 lack nominal plural, and several dozen mark plurality only to a limited extent, for example only on human nouns (Feature 34A). The Grambank database (Skirgård et al., 2023) identifies plural marking on nouns in 1,282 of 2,389 languages (Feature GB044); the other languages may express plurality, for example, by means of a free-standing marker, noun reduplication, or not at all (cf. Feature 33A in WALS).

2.2. Number in grammars of English, Czech, and Greek

English, Czech, and Greek are Indo-European languages in which, despite their differing morphological profiles, the category of number is structured in a similar way. In the three languages, number in nouns is primarily realized as a binary opposition between singular and plural. English marks number most commonly by the suffix *-s* (as in *leg – legs*), with limited inflectional variation and a modest set of irregular forms (e.g. *foot – feet*; cf. Quirk et al. 1985; Huddleston and Pullum 2002; Bauer et al. 2013, among others).

Czech exhibits a morphologically rich inflectional system with numerous noun inflectional classes employing different formal markers to express singular and plural, while also retaining residual traces of the historical dual, preserved in a small group of nouns, especially those referring to parts of the body (e.g. Komárek et al. 1986; Havránek and Jedlička 2002). Number is realized jointly with morphological case in a single inflectional (portmanteau) ending; cf. selected forms of the noun *noha* 'foot': *noh-a* foot-NOM.SG, *noh-ám* foot-DAT.PL, *noh-ama* foot-INSTR.DUAL.

Greek is likewise fusional, with nominal number and case expressed cumulatively in portmanteau endings. Although it distinguishes singular and plural across several declensional classes, the inventory of formal number markers is comparatively smaller and more regular than in Czech (cf. Τριανταφυλλίδης 1979; Τζεβελέκου et al. 2007; Holton et al. 2012 (Triantaphillídis 1979, Tzevelékou 2007), or Χατζησαββίδης and Χατζησαββίδου 2014 (Khatzisavvídis and Khatzisavvídou 2014), or Κλαίρης and Μπαμπινιώτης 2010 (Klaíris and Bambiniótis 2010). Cf. selected forms of the noun *πόδι* (pód-i) 'foot', with portmanteau markers delimited: *πόδ-ι* (pód-i) foot-NOM/ACC.SG, *ποδ-ιού* (pod-ioú) foot-GEN.SG, *πόδ-ια* (pód-ia) foot-NOM/ACC.PL, *ποδ-ιών* (pod-ión) foot-GEN.PL.

2.3. Singularia and pluralia tantum

Although the grammars of individual languages follow different traditions, the assumption that "[m]ost nouns have both singular and plural" (Huddleston and Pullum 2002, p. 340), recurs, to varying degrees of explicitness, across them (cf. Komárek et al. 1986, p. 45 or Holton et al. 2012, Chapter 2). Nouns occurring exclusively in the singular or in the plural are typically treated as peripheral cases. These nouns are called singularia tantum, singular-only nouns, or singular invariable nouns for the first group, and pluralia tantum, plural-only nouns, or plural invariable nouns for the second group, and their precise delimitation may vary not only between languages, but also between individual works on a single language (Corbett, 2019; Acquaviva and Gardelle, 2023).

For English, Quirk et al. (1985, pp. 297–318), for example, assesses the number of nouns primarily on the basis of syntactic behavior, meaning that nouns ending in *-s* are also classified as singularia tantum, such as the names of disciplines (*acoustics, linguistics*) or diseases (*measles, rickets*) that are used with a verb in the singular form. In contrast, Bauer et al. (2013, p. 124) favor morphological criteria and classify the names of diseases as pluralia tantum due to the presence of the ending *-s* and the lack of forms without this ending.

Grammars of Czech, as well as those of Greek, proceed relatively consistently within each linguistic tradition, starting from semantic criteria but checking for the presence of plural markers in the noun forms and the absence of singular forms, thus arriving at similar, though relatively modest, lists of singularia and pluralia tantum. Singularia tantum cover several semantic categories: abstract nouns (cf. Cz. *spravedlnost*, Gr. *δικαιοσύνη* (dikaiosíni), both 'justice'), mass nouns (Cz. *popel*, Gr. *στάχτη* (stákhti), both 'ash'), collective nouns (Cz. *nábytek* 'furniture', Gr. *κλήρος* (klíros) 'clergy'). Pluralia tantum include inherently paired objects (Cz. *brýle*, Gr. *γυαλιά* (yialiá), both 'glasses'), plural mass concepts (Cz. *splašky*, Gr. *λύματα* (límata), both 'dregs'), or names of events (Cz. *narozeniny*, Gr. *γενέθλια* (yenéthlia), both 'birthday').

2.4. A view from Universal Dependencies

A comment is due on the representation of the category of number in the Universal Dependencies collection, as the treebanks from this collection, in the respective versions specified below, were used to train the Stanza and UDPipe models, which we employ to annotate the comparable corpora for the present study.

Within Universal Dependencies treebanks, which are constructed on the basis of a unified annotation scheme (de Marneffe et al., 2021), number is encoded as a morphosyntactic feature

	Raw comparable corpora		Processed datasets		
Language	**Sentences**	**Tokens**	**Tool**	**Sentences**	**Tokens**
English	5,000,000	104,430,900	Stanza	5,095,753	115,948,006
			UDPipe	5,038,278	117,091,745
Czech	5,000,000	81,762,710	Stanza	5,043,844	87,456,006
			UDPipe	5,067,301	87,546,410
Greek	5,000,000	106,908,786	Stanza	5,046,190	112,260,066
			UDPipe	6,051,461	109,330,298

Table 1: Size of the raw comparable corpora and of the datasets processed by the two annotation tools.

of individual noun forms. This `Number` feature takes the values `Sing` (singular) and `Plur` (plural) in the treebanks for English, Czech, and Greek. While Greek is limited to these two values, the English data additionally include the value `Ptan` (plurale tantum) with nouns that appear only in the plural. In the Universal Dependencies treebanks of Czech, besides singular and plural, the value `Dual` is attested for forms referring to two entities, as well as the value `Coll` (collective), used for nouns that employ grammatical singular to denote sets of objects.

Linguistically, these values are clearly heterogeneous: At the level of the grammatical opposition between singular and plural, we find the class of pluralia tantum, defined precisely by the absence of one of the two values, alongside the lexical category of collective nouns. The inclusion of these values—together with additional ones attested in Universal Dependencies treebanks of languages other than the three analyzed here—likely reflects annotation decisions inherited from the original datasets prior to their harmonization within the Universal Dependencies framework.[1]

2.5. Expectations arising from research on paradigmatic defectiveness

A final line of research relevant to our study concerns morphological defectiveness. While not focusing specifically on number, it examines inflectional paradigms more broadly, challenging the assumption of paradigmatic completeness and regularity. Baerman et al. (2010) show that large portions of the lexicon exhibit systematic restrictions, with grammatically predictable forms often unattested. Based on corpus evidence, Janda and Tyers (2021) demonstrate that many Russian nouns systematically avoid certain case–number combinations, with defectiveness varying across inflectional classes. Nikolaev and Bermel (2022) report similar patterns in Czech, arguing that paradigmatic gaps extend across entire semantic domains rather than being isolated lexical anomalies.

[1] Cf. the Universal Dependencies documentation on English, Czech, and Greek, and on the feature itself.

3. Data and Methods

3.1. Construction of the comparable corpora

For the purposes of the present study, we constructed three comparable monolingual corpora of 5 million sentences each for Czech, English, and Greek, drawn from the Leipzig Corpora Collection (Goldhahn et al., 2012). The Leipzig Corpora Collection was chosen for its consistent preprocessing methodology, comparable text types, and sentence-level organization across languages. Each corpus combines news and Wikipedia texts in a 4:1 ratio (80% news, 20% Wikipedia), with news components spanning 2019–2024 and Wikipedia snapshots from 2016–2021. Despite identical sentence counts, the corpora differ in total tokens (cf. the left-hand side of Table 1), corresponding to average sentence lengths of 20.9 tokens for English, 16.4 tokens for Czech, and 21.4 tokens for Greek.

3.2. Morphosyntactic annotation

We annotated each corpus using two tools trained on treebanks from the Universal Dependencies collection: the latest version of Stanza, Stanza 1.11.0 (Qi et al., 2020) with models trained on Universal Dependencies 2.15, and the latest version of UDPipe 2 (Straka, 2018) with models trained on Universal Dependencies 2.17 (`english-gum`, `czech-pdtc`, `greek-gud`).[2] Both tools were ap-

[2] For the UDPipe models, high accuracy is reported across all tasks relevant to our study (`https://ufal.mff.cuni.cz/udpipe/2/models`); cf. the results for sentence segmentation, word-level tokenization, part-of-speech tagging, morphological feature prediction, and lemmatization: `english-gum`: 95.77, 99.74, 98.12, 98.04, 98.83; `czech-pdtc`: 94.82, 99.96, 99.24, 98.88, 99.54; `greek-gud`: 95.24, 99.94, 98.08, 94.36, 95.93.

Lang	Tool	Noun tokens	Sing	Plur	Dual	Ptan	No value
English	Stanza	22,079,865	16,098,396 (72.9%)	5,918,052 (26.8%)	—	63,417 (0.3%)	— —
	UDPipe	21,604,055	15,651,183 (72.4%)	5,867,912 (27.2%)	—	71,894 (0.3%)	13,066 (0.1%)
Czech	Stanza	20,730,449	14,595,098 (70.4%)	5,537,569 (26.7%)	4,603 (0.02%)	—	593,179 (2.9%)
	UDPipe	20,629,395	14,358,900 (69.6%)	5,636,277 (27.3%)	5,315 (0.03%)	—	628,903 (3.0%)
Greek	Stanza	21,671,723	14,642,436 (67.6%)	6,054,245 (27.9%)	—	—	975,042 (4.5%)
	UDPipe	21,526,746	14,656,751 (68.1%)	6,120,727 (28.4%)	—	—	749,268 (3.5%)

Table 2: Distribution of the values of the `Number` feature across noun tokens. Percentages calculated relative to total noun token counts in the individual datasets (rows). A dash (—) indicates the absence of the specific value in the dataset.

plied with default settings including sentence segmentation and word-level tokenization, ensuring easy replicability of the data compilation process and direct comparability with other Universal Dependencies research. By using two tools, we can validate quality through their agreement.

All processing was conducted on a high-performance computing cluster at the Institute of Formal and Applied Linguistics of Charles University using parallel processing strategies: input sentences were divided into 5,000-sentence bundles distributed across multiple cluster partitions. Output files follow CoNLL-U format (Nivre et al., 2016). The right-hand side of Table 1 lists sentence and token counts for the six resulting datasets.

The datasets have been made available in the LINDAT/CLARIAH-CZ repository at `http://hdl.handle.net/11234/1-6120`. Additional materials, including analysis scripts, manually annotated files for the evaluation of part-of-speech tagging and lemmatization quality (see Section 3.3), as well as grammar-derived lists of singularia and pluralia tantum (used in Section 4.3), are available on GitHub.

3.3. Annotation quality validation

Automated annotation tools may alter pre-existing segmentation—even when processing pre-segmented input (Demrozi et al., 2023; Bindi, 2025). This effect was observed in our processing, prompting an evaluation of consistency between Stanza and UDPipe. We therefore assessed the preservation of sentence and token segmentation relative to the original corpora and inter-tool differences in part-of-speech distributions.

Sentences and tokens in the CoNLL-U outputs were matched against the original plain-text corpora. Preservation refers to sentences or tokens remaining unchanged relative to the raw corpus; inter-tool agreement corresponds to the proportion preserved intact by both tools.

At the **sentence** level, Czech and English show high preservation rates and strong agreement between the tools. In Czech, Stanza preserved 97.62% and UDPipe 96.80% of sentences, with 96.07% jointly preserved. In English, Stanza preserved 96.70% and UDPipe 95.94%, with 94.20% jointly preserved. Greek shows lower agreement: Stanza preserved 93.35% of sentences, whereas UDPipe preserved 79.35% (77.35% jointly).

At the **token** level, preservation rates are comparable across tools. In Czech, Stanza preserved 85.33% and UDPipe 85.13% of tokens (85.08% jointly). In English, Stanza preserved 86.97% and UDPipe 88.26% (86.55% jointly). In Greek, Stanza preserved 89.22% and UDPipe 91.09%, with 88.20% jointly preserved.

Since the raw corpora are not tagged for grammatical categories, **part-of-speech** distributions were compared only between Stanza- and UDPipe-annotated datasets, focusing on nouns as the category of interest. As shown in Table 2, Stanza identified more noun forms than UDPipe in all three languages. The difference is largest in English (nearly 476 thousand tokens), smaller in Greek (146 thousand), and smallest in Czech (101 thousand tokens).

To assess the quality of part-of-speech tagging and lemmatization, 500 tokens per tool (1,000 per language, and 3,000 instances in total) were manually inspected. For Czech, POS tagging accuracy reached 95.6% for Stanza (22 incorrect assignments) and 97.6% for UDPipe (12 errors), while lemmatization accuracy was 95.0% (25 errors) and 98.4% (8 errors), respectively. For English, POS tagging accuracy was 94.6% (27 errors) for Stanza and 94.8% (25 errors) for UDPipe, with lemmatization accuracy at 92.6% (35 errors) and

93.6% (30 errors), respectively. For Greek, POS tagging accuracy was 88.4% for Stanza (58 errors) and 84.2% for UDPipe (79 errors), while lemmatization accuracy reached 84.4% (78 errors) and 77.2% (114 errors), respectively.

3.4. Extraction of number values and calculation of the plural ratio

From each annotated corpus, we extracted all tokens tagged as nouns[3] and, for each such token, retrieved and stored the value of the `Number` feature from the CoNLL-U `FEATS` column. As shown in Table 2, singular and plural forms were consistently attested in all six datasets. For English, forms annotated with the `Number` value `Ptan` (plurale tantum) were identified by both tools. For Czech, `Dual` forms were identified in both datasets, but no forms were assigned the value `Coll` (unlike in the Universal Dependencies treebanks of Czech; cf. Section 2.4).

To analyze number distribution within and across languages, we aggregated token counts by lemma and calculated a plural ratio for each distinct lemma. Forms with the values `Ptan` and `Dual` were treated as non-singular and added to plural forms. The plural ratio, representing the proportion of non-singular forms among all number-marked tokens of a lemma, was calculated for each language as follows:

– **English:**

$$\text{plural ratio} = \frac{\text{count}_{\texttt{Plur}} + \text{count}_{\texttt{Ptan}}}{\text{count}_{\texttt{Sing}} + \text{count}_{\texttt{Plur}} + \text{count}_{\texttt{Ptan}}}$$

– **Czech:**

$$\text{plural ratio} = \frac{\text{count}_{\texttt{Plur}} + \text{count}_{\texttt{Dual}}}{\text{count}_{\texttt{Sing}} + \text{count}_{\texttt{Plur}} + \text{count}_{\texttt{Dual}}}$$

– **Greek:**

$$\text{plural ratio} = \frac{\text{count}_{\texttt{Plur}}}{\text{count}_{\texttt{Sing}} + \text{count}_{\texttt{Plur}}}$$

A plural ratio of 0 indicates the lemma occurs only in singular, while a ratio of 1 indicates the absence of singular forms among the attested tokens of the lemma. Ratios between 0 and 1 show the lemma occurs in both singular and plural in varying proportions (see Section 4.2 for details).

Apart from the values included in the plural ratio, Table 2 also reports counts of forms with no assigned value. For English, this affects only one dataset (<0.1% of forms), while for Czech it is about 3% and for Greek up to 4.5%. These forms merit further investigation, as they reflect the tools' performance and data used for their training. For Section 4, however, they are disregarded.

[3]Thus, tokens tagged `PROPN` (reserved for proper nouns in the Universal Dependencies scheme) are not included in the analysis, as proper nouns are expected to be biased toward singular-only or plural-only usage and could skew the distributional patterns.

3.5. Compilation of defective-number noun lists for validation in the data

In order to assess how the automatically annotated comparable corpora reflect phenomena described in grammatical accounts within and across languages, we used the reference grammars cited in Section 2.2 to compile lists of nouns considered as singularia tantum and pluralia tantum. For English, in cases of conflicting classifications of particular lexemes—which, as noted above, can be explained by prioritizing morphological over syntactic criteria, or vice versa, in delimiting these categories—we followed Quirk's classification.

The lists compiled in this way contain 54 singularia tantum and 87 pluralia tantum candidates for English, 27 and 54 respectively for Czech, and 22 and 59 for Greek. For each candidate, the corresponding lemma was identified in the annotated datasets, and its plural ratio was determined; see Section 4.3 for a discussion of the attestation and distribution of these items.

4. Results and Linguistic Analysis

4.1. Singular vs. plural at the token level within and across languages

All six datasets exhibit a strong singular bias (67–73% of noun tokens) when considering the proportion of singular and plural forms—without yet linking forms to lemmas. The ratios of singular and plural forms are consistent for each language across both datasets and are also similar across languages: 2.7:1 for English, 2.6:1 for Czech, and 2.4:1 for Greek. Table 2 shows the distribution of `Number` values across all noun tokens in each dataset.

Before merging tokens marked with the values `Ptan` and `Dual` with plural forms for the calculation of the plural ratio in the next section, we briefly examine these categories, as they may reveal potential inconsistencies in the automatic annotation.

The `Ptan`, which occurs only in the English datasets, is assigned by Stanza to 63 thousand forms belonging to 363 lemmas, whereas UDPipe assigns it to 71 thousand forms associated with 800 lemmas. In both datasets, most cases involve expressions containing digits (e.g. *1970s*) or canonical examples of pluralia tantum (e.g. *thanks, clothes*). However, while in the Stanza dataset the lemma *clothes* is instantiated exclusively by `Ptan` forms, the UDPipe dataset also con-

tains some Sing and Plur forms under the same lemma. A similar inconsistency appears with *remains* and *remain* (the latter of which should not exist): cf. *remains* by Stanza: 117 Plur, 1,682 Ptan, and by UDPipe: 57 Plur, 764 Ptan; *remain* by Stanza: 76 Sing, 44 Plur, 7 Ptan, and by UDPipe: 38 Sing, 762 Plur, 321 Ptan.

The Dual value, provided only for Czech, was assigned by Stanza to 4.6 thousand forms belonging to three body-part lemmas (*oko* 'eye', *ruka* 'arm', *noha* 'leg'), which is consistent with Universal Dependencies guidelines. In contrast, the UDPipe dataset contains 5.3 thousand such tokens distributed across 260 lemmas, including nouns referring to persons or objects (e.g. *holka* 'girl', *droga* 'druh'), where the forms likely represent colloquial instrumental plurals rather than genuine dual forms. Among these lemmas, we also find instrumental forms that were not recognized as inflected forms and are incorrectly considered to be lemmas (e.g. *kanálama*, which should be lemmatized as *kanál* 'canal'), as well as non-existent strings (e.g. *pracha* instead of *prachy* 'money').

4.2. Proportion of singular and plural forms per lemma

A simple way to examine noun behavior with respect to number is to rank noun lemmas by their plural ratio, i.e., the percentage of non-singular forms among all forms of a lemma, as introduced above. Figures 1–3 depict this ranking for each dataset. The x-axis shows the plural ratio, with the value 0 on the left (i.e. exclusively singular forms and no plurals) and the value 1 on the right (i.e. exclusively plural forms and no singulars). The y-axis, on a logarithmic scale, represents the number of lemmas corresponding to a given plural ratio. The absolute frequencies of individual lemmas are not taken into account; thus, both high- and low-frequency lemmas fall into the same bar if they have the same proportion of plural forms. However, only lemmas with at least ten occurrences were included in the plots.

All figures show the same trend: A high number of singular-only noun lemmas creates a peak on the left in all bar plots, while plural-only nouns form a smaller peak on the right. While this pattern deviates from grammatical expectations, it is consistent with our previous results based on different datasets annotated with different tools (namely, on the British National Corpus for English and the SYN2020 corpus for Czech; Ševčíková and Diamantopoulos 2025).

Absolute counts and percentages for the extremes, for (arbitrarily chosen) adjacent ranges (plural forms up to 10 or above 90%) and for the intermediate range are reported on the right-hand side of Table 3, which, like Figures 1–3, includes only lemmas with at least ten occurrences. Lemmas with ten or more forms but no plural among them account for roughly 30% of nouns in each dataset—far exceeding their peripheral status in grammars. Relaxing the ten-token threshold (left-hand side of Table 3) increases the proportion of singular-only nouns to 60–70%. Plural-only nouns remain lower, between 2–12% with the threshold and 12–25% without it.

The difference between the two sides of the table shows that low-frequency lemmas with fewer than 10 occurrences are primarily found in the extreme groups with a plural ratio of 0 and 1. Applying a minimum threshold of 10 forms reduces these extreme groups dramatically; at the same time, however, it leads to a relative increase—given the newly established totals—in the groups adjacent to these extremes.[4]

4.3. Singular-only and plural-only nouns: data vs. grammars

The correspondence between the analyzed datasets and the picture of paradigmatic defectiveness depicted in the grammars is illustrated by colored dots in Figures 1–3. Nouns listed in grammars as singularia tantum (red dots) and pluralia tantum (blue dots) are plotted on bar plots showing the distribution of words according to the plural ratio; the candidate dots are placed at arbitrary vertical positions with randomly varied sizes for visual distinction.

For English, candidate words in both datasets are only partially located in the expected extremes, with a substantial proportion scattered across the middle range among nouns attested in both singular and plural. Examples of words whose annotation matches grammatical expectations include singularia tantum such as *honesty* (plural ratio 0.00 in both datasets) and *physics* (Stanza 0.000, UDPipe 0.001; the same order applies throughout this section), as well as pluralia tantum like *people* (0.98, 1.00). However, the tools differ in annotating other predicted singularia tantum (*economics*: 0.35, 0.01; *politics*: 0.94, 0.16) and pluralia tantum (*police*: 0.60, 0.53; *vermin*: 0.69, 0.00).

For Czech and Greek, the data show greater agreement with the grammars: In the UDPipe-annotated dataset for Czech and the Stanza dataset for Greek, most candidates cluster in the respective extremes or close to them. In the Stanza dataset for Czech and the UDPipe dataset for Greek, singularia tantum remain on

[4] The absolute counts of lemmas in the (0, 0.1] and [0.9, 1) intervals remain unchanged, as only lemmas with at least ten occurrences fall into these intervals, even without applying a minimum frequency threshold.

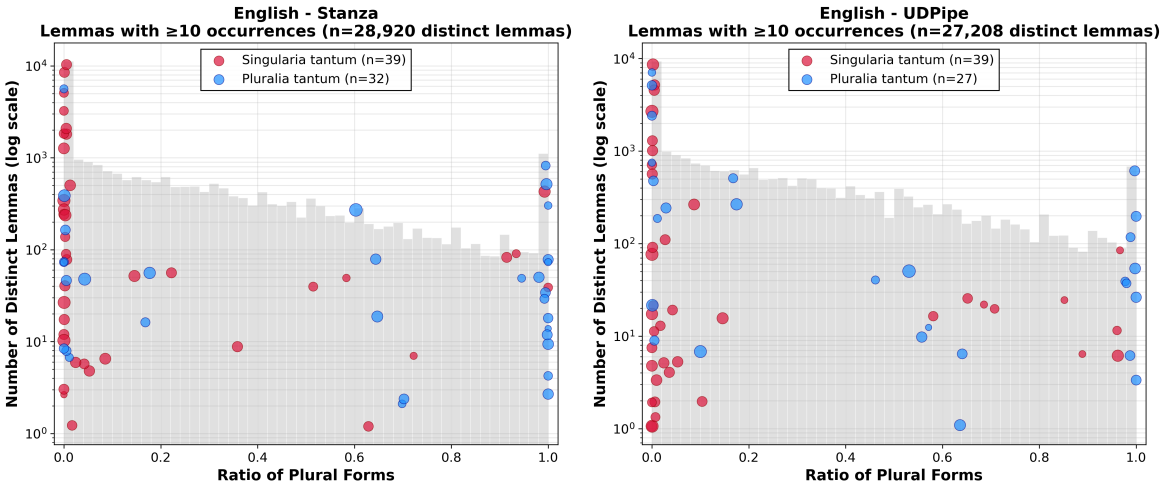

Figure 1: Distribution of English noun lemmas (with ≥10 occurrences) by plural ratio. Singularia tantum (red, 54 candidates) and pluralia tantum (blue, 87 candidates) overlaid. Left: Stanza; right: UDPipe.

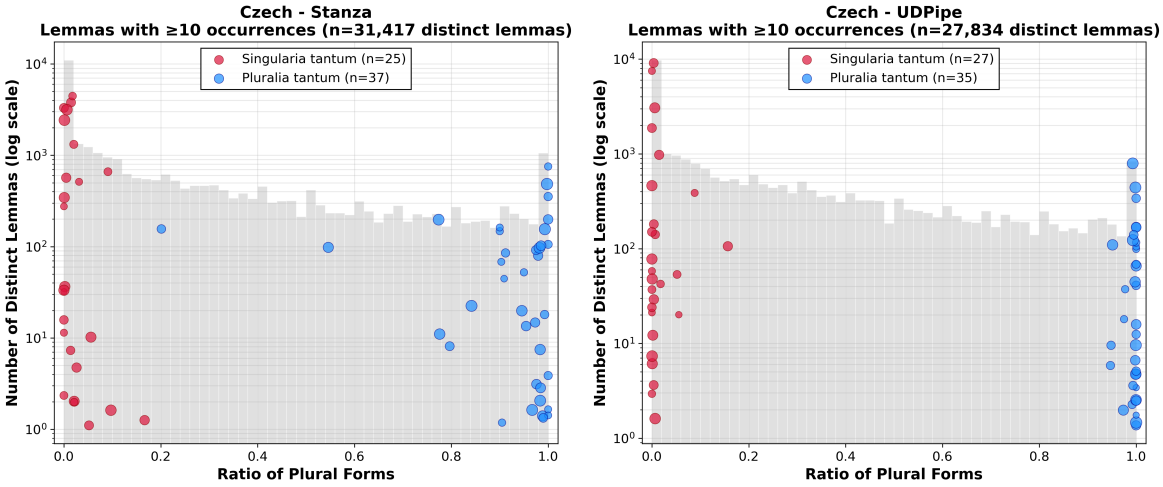

Figure 2: Distribution of Czech noun lemmas (with ≥10 occurrences) by plural ratio. Singularia tantum (red, 27 candidates) and pluralia tantum (blue, 54 candidates) overlaid. Left: Stanza; right: UDPipe.

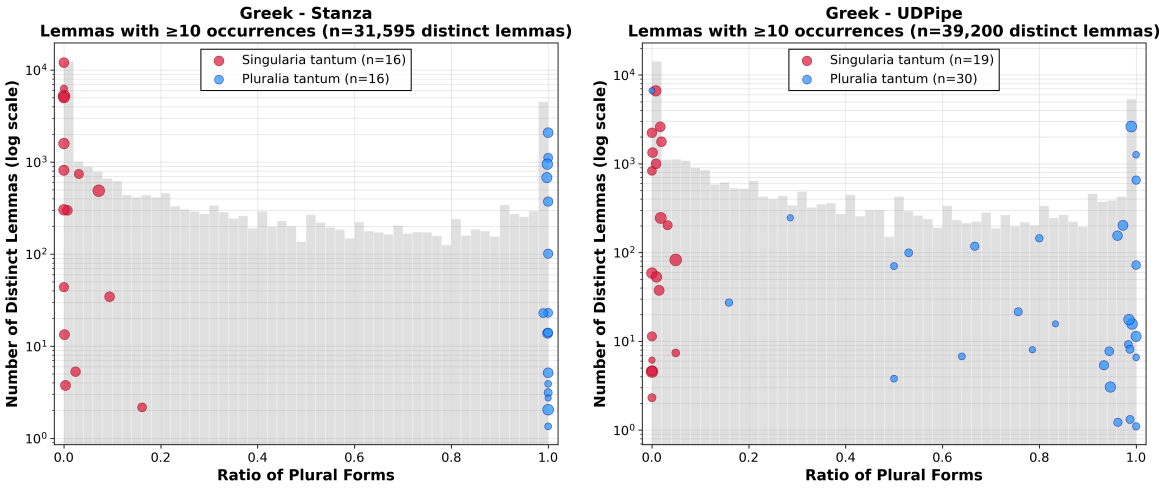

Figure 3: Distribution of Greek noun lemmas (with ≥10 occurrences) by plural ratio. Singularia tantum (red, 22 candidates) and pluralia tantum (blue, 59 candidates) overlaid. Left: Stanza; right: UDPipe.

All noun lemmas						Noun lemmas with ≥10 forms					
Total nouns	Plural ratio = 0	(0,0.1]	(0.1,0.9)	[0.9,1)	1	Total nouns	Plural ratio = 0	(0,0.1]	(0.1,0.9)	[0.9,1)	1
English Stanza											
146,892	102,554	4,897	20,827	522	18,092	28,920	9,930	4,897	12,549	522	1,022
	(69.8%)	(3.3%)	(14.2%)	(0.4%)	(12.3%)		(34.3%)	(16.9%)	(43.4%)	(1.8%)	(3.5%)
English UDPipe											
147,966	100,204	4,990	22,451	528	19,793	27,208	8,024	4,990	13,064	528	602
	(67.7%)	(3.4%)	(15.2%)	(0.4%)	(13.4%)		(29.5%)	(18.3%)	(48.0%)	(1.9%)	(2.2%)
Czech Stanza											
145,635	91,321	6,921	25,504	993	20,896	31,417	8,922	6,921	13,633	993	948
	(62.7%)	(4.8%)	(17.5%)	(0.7%)	(14.3%)		(28.4%)	(22.0%)	(43.4%)	(3.2%)	(3.0%)
Czech UDPipe											
124,745	75,155	5,597	22,303	851	20,839	27,834	7,947	5,597	12,727	851	712
	(60.2%)	(4.5%)	(17.9%)	(0.7%)	(16.7%)		(28.6%)	(20.1%)	(45.7%)	(3.1%)	(2.6%)
Greek Stanza											
194,559	118,945	5,076	21,219	1,622	47,697	31,595	10,943	5,076	9,877	1,622	4,077
	(61.1%)	(2.6%)	(10.9%)	(0.8%)	(24.5%)		(34.6%)	(16.1%)	(31.3%)	(5.1%)	(12.9%)
Greek UDPipe											
272,126	163,821	6,195	30,235	2,200	69,675	39,200	12,603	6,195	13,393	2,200	4,809
	(60.2%)	(2.3%)	(11.1%)	(0.8%)	(25.6%)		(32.2%)	(15.8%)	(34.2%)	(5.6%)	(12.3%)

Table 3: Lemma-level distribution of `Number` values for all nouns (left) vs. nouns with ≥10 forms (right). Lemma counts by plural ratio: 0 (only singular); (0,0.1] (mostly singular); (0.1,0.9) (both singular and plural); [0.9,1) (mostly plural); 1 (only plural). Percentages sum to 100% per total count of lemmas.

		English		Czech		Greek	
		Sg tantum	Pl tantum	Sg tantum	Pl tantum	Sg tantum	Pl tantum
Tool	**Lemmas**	Candidates / Attested / Confirmed					
Stanza	All	54 / 40 / 10	87 / 37 / 12	27 / 25 / 7	54 / 39 / 8	22 / 20 / 10	59 / 29 / 21
	≥10	54 / 39 / 9	87 / 32 / 8	27 / 25 / 7	54 / 37 / 7	22 / 16 / 7	59 / 16 / 12
UDPipe	All	54 / 39 / 5	87 / 33 / 9	27 / 27 / 10	54 / 38 / 20	22 / 20 / 8	59 / 45 / 14
	≥10	54 / 39 / 5	87 / 27 / 3	27 / 27 / 10	54 / 35 / 18	22 / 19 / 7	59 / 30 / 6

Table 4: Attestation of grammar-derived singularia tantum and pluralia tantum candidates in the datasets. For each language, the table shows the number of candidates, how many of them are attested (in the full dataset vs. among lemmas with ≥10 occurrences), and how many of those attested are found exclusively in singular or plural in the datasets.

the left side of the plot, whereas pluralia tantum are dispersed across the scale, reaching the plural ratios of singularia tantum. In Czech, singularia tantum like *lidstvo* 'humankind' (0.00 in both datasets) and *kvítí* 'flowers' (0.00, 0.09), and pluralia tantum such as *kleště* 'pliers' (1.00) and *dveře* 'door' (0.99, 1.00) match grammatical expectations, though *vrátka* 'gate' shows large discrepancies (0.20, 0.99). In Greek, both tools largely agree on singularia tantum like *οξυγόνο* (oxigóno) 'oxygen' (0.00 in both datasets) and *ζάχαρη* (zákhari) 'sugar' (0.00, 0.01), and pluralia tantum such as *περίχωρα* (períkhora) 'suburbs' (1.00, 0.96) and *γενέθλια* (yenéthlia) 'birthday' (0.99, 0.95). However, expected pluralia tantum like *πρόθυρα* (próthira) 'threshold' are analyzed differently (1.00, 0.50). Other predicted pluralia tantum, such as *έξοδα* (éxoda) 'expenses', are absent in the Stanza dataset, while UDPipe treats them mostly as singular forms (plural ratio 0.29).

Table 4 reports the exact counts of candidate nouns, how many were found in the data, and how many were confirmed as restricted exclusively to singular or plural forms. For all languages, more singularia tantum than pluralia tantum candidates appeared on the lists. Many pluralia tantum were not found even when searching the entire dataset, with further reductions when

limiting to lemmas with ≥10 occurrences; this effect is smaller for singularia tantum. The absence of the Greek plurale tantum *άρματα* (ármata) 'chariots' likely reflects its archaic character, while the failure to find singularia tantum like *γυμναστική* (yimnastikí) 'gymnastics' or pluralia tantum like *μαθηματικά* (mathimatiká) 'mathematics', as with English pluralia tantum *outskirts* or *shorts*, points to lemmatization issues. Problematic lemmatization is also seen in the Czech pluralia tantum *Velikonoce* 'Easter', which was not found in the UDPipe dataset, as it appears there only in lowercase (*velikonoce*), along with two incorrect lemmas (*velikonoc* and *velikonce*).

Last but not least, the data also reveal that, particularly among nouns attested in the datasets exclusively in the singular (plural ratio 0.00), there are nouns that grammatical descriptions do not consider to be singularia tantum. In the languages analyzed, these include spatial or temporal expressions; cf. English *north* and *southwest*; Czech *jih* 'south' and *sever* 'north', or *minulost* 'past', *budoucno* 'future', *dnešek* 'this day', *leden* 'January', *úterý* 'Tuesday'; and Greek *μέλλον* (méllon) 'future', μεσημέρι (mesiméri) 'noon', *νύχτα* (níkhta) 'night', *Κυριακή* (Kiriakí) 'Sunday'.

5. Conclusions

In this study, we examined the morphological category of number in English, Czech, and Greek using comparable corpora annotated with Stanza and UDPipe, tools reported to achieve excellent performance. By systematically comparing paired datasets for each language, we aimed to assess whether such annotated corpora constitute a reliable basis for drawing linguistic conclusions.

Despite differences between the paired datasets in sentence and word segmentation, noun counts, and number values, these discrepancies did not appear to substantially affect the overall patterns. In all datasets for each language, singular forms were two to three times more frequent than plurals, and lemma-level analysis revealed a picture differing from grammar descriptions, showing a strong inclination of a substantial portion of nouns toward singular and, to a lesser extent, toward plural forms.

A more detailed inspection of a small set of nouns expected to exhibit defective number revealed inconsistencies, indicating that such annotated comparable corpora may yield divergent results on linguistically nuanced questions. The observed differences seem to stem from variations in lemmatization—Stanza producing more stable and linguistically conformant lemmas, while UDPipe sometimes generates nonexistent lemmas—and likely point to additional issues that should be addressed in dedicated follow-up experiments.

6. Acknowledgements

The research reported in the present paper has been supported by the Czech Science Foundation (Project No. GA26-21822S). It has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062).

7. Bibliographical References

Paolo Acquaviva and Laure Gardelle. 2023. Pluralia tantum and singularia tantum. *The Wiley Blackwell Companion to Morphology*, pages 1–28.

Matthew Baerman, Greville G Corbett, and Dunstan Brown. 2010. *Defective paradigms: Missing forms and what they tell us*. Liverpool University Press.

Laurie Bauer, Rochelle Lieber, and Ingo Plag. 2013. *The Oxford reference guide to English morphology*. Oxford University Press, Oxford.

Beatrice Bindi. 2025. Evaluating Stanza and UDPipe for Morphosyntactic Annotation of Old Russian: A Case Study on Maximus the Greek. *Scripta & e-Scripta*, pages 39–60.

Hee-Soo Choi, Bruno Guillaume, and Karën Fort. 2021. Corpus-based language universals analysis using Universal Dependencies. In *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pages 33–44, Sofia, Bulgaria. Association for Computational Linguistics.

Greville G. Corbett. 2000. *Number*. Cambridge University Press.

Greville G Corbett. 2019. Pluralia tantum nouns and the theory of features: A typology of nouns with non-canonical number properties. *Morphology*, 29(1):51–108.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Florenc Demrozi, Cristian Turetta, Fadi Al Machot, Graziano Pravadelli, and Philipp H. Kindt. 2023. A comprehensive review of automated data annotation techniques in human activity recognition.

Matthew S. Dryer and Martin Haspelmath. 2013. The World Atlas of Language Structures Online.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Bohuslav Havránek and Alois Jedlička. 2002. *Stručná mluvnice česká*. Fortuna, Praha.

David Holton, Peter Mackridge, Irene Philippaki-Warburton, and Vassilios Spyropoulos. 2012. *Greek: A comprehensive grammar of the modern language*. Routledge.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge.

A. Laura Janda and M. Francis Tyers. 2021. Less is more: why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory*, 17(1):109–141.

Yingqi Jing, Paul Widmer, and Balthasar Bickel. 2023. Word order evolves at similar rates in main and subordinate clauses. *Diachronica*, 40(4):532–556.

Miroslav Komárek, Jan Kořenský, Jan Petr, and Jarmila Veselková. 1986. *Mluvnice češtiny 2. Tvarosloví*. Academia, Praha.

Natalia Levshina, Savithry Namboodiripad, and Marc Allassonnière-Tang et al. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Alexandre Nikolaev and Neil Bermel. 2022. Explaining uncertainty and defectivity of inflectional paradigms. *Cognitive Linguistics*, 33(3):585–621.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Randolph Quirk, Sidney Greenbaum, Geoffrey N. Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Magda Ševčíková and Konstantinos Diamantopoulos. 2025. Word-formation of singular-only nouns: A pilot study in four languages. Presented at the Word-Formation Theories VII & Typology and Universals in Word-Formation VI conference, Košice, Slovakia.

Hedvig Skirgård, Hannah J. Haynie, and Damián E. Blasi et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Χρήστος Κλαίρης and Γεώργιος Μπαμπινιώτης. 2010. *Γραμματική της νέας ελληνικής: δομολειτουργική-επικοινωνιακή. Το όνομα της νέας Ελληνικής.* Ελληνικά Γράμματα.

Μαρία Τζεβελέκου, Βασιλική Κάντζου, and Σπυριδούλα Σταμούλη. 2007. *Βασική γραμματική της Ελληνικής*. Ινστιτούτο Επεξεργασίας του Λόγου - Ε.Κ. "Αθηνά".

Μανόλης Α. Τριανταφυλλίδης. 1979. *Νεοελληνική γραμματική: αναπροσαρμογή της Μικρής Νεοελληνικής Γραμματικής του Μανόλη Τριανταφυλλίδη*. Οργανισμός Εκδόσεων Διδακτικών Βιβλίων.

Σοφρώνης Χατζησαββίδης and Αθανασία Χατζησαββίδου. 2014. *Γραμματική της Νέας Ελληνικής Γλώσσας Α', Β', Γ' Γυμνασίου*. Οργανισμός Εκδοσης Διδακτικών Βιβλίων - Υπουργείο Παιδείας και Θρησκευμάτων.

8. Language Resource References

Leipzig Corpora Collection. 2019. *English news corpus based on material from 2019*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2020a. *Czech news corpus based on material from 2020*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2020b. *English news corpus based on material from 2020*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2021a. *Czech Wikipedia corpus based on material from 2021*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2021b. *English Wikipedia corpus based on material from 2021*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2021c. *Modern Greek news corpus based on material from 2021*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2021d. *Modern Greek Wikipedia corpus based on material from 2021*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2022a. *Czech news corpus based on material from 2022*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2022b. *Modern Greek news corpus based on material from 2022*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2023a. *Czech news corpus based on material from 2023*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2023b. *English news corpus based on material from 2023*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2023c. *Modern Greek news corpus based on material from 2023*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2024a. *Czech news corpus based on material from 2024*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2024b. *English news corpus based on material from 2024*. University of Leipzig. Accessed: 2026-03-04.

Leipzig Corpora Collection. 2024c. *Modern Greek news corpus based on material from 2024*. University of Leipzig. Accessed: 2026-03-04.

Diamantopoulos, Konstantinos. 2026. *Automatically Annotated Corpora with Stanza and UDPipe for Czech, English, and Greek*. LINDAT/CLARIAH-CZ digital library.

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell et al. 2024. *Universal Dependencies 2.15*. LINDAT/CLARIAH-CZ. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell et al. 2025. *Universal Dependencies 2.17*. LINDAT/CLARIAH-CZ. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).