Patterns of Distribution of Noun Case Frequency

Arlee Pearlswig
George Mason University
English – Linguistics

Abstract

The distribution of case-marking across languages informs us of the linguistic processing of noun case. Seventy-five languages were parsed to test if case-marking supported a statistical model such as Zipf's law. By extracting case features from available UD annotations, the counted frequency of case-marked forms shows statistical models of the forms' distributions. Most languages showed a higher frequency for the first two to four case-markers and then a significant decrease in frequency for the remaining case-markers. Each language did not show the same curve, splitting results between the two statistical models, favoring either a Zipfian or exponential distribution. The trends were not uniform across all data samples. Data suggests a possible markedness of case-markers beyond a fourth semantic case.

Keywords: Noun case, Zipf's law, typology, statistics, linguistic processing

Introduction

Case is not a universal feature of language. In case-marking systems, nominal elements, modifiers, determiners, adpositions, and even verbs can show case. According to Zipf's law, languages should show a high frequency of usage of one or two forms in a given category with a significant decrease in frequency across forms (Manning & Schütze, 1999). Applications to linguistic forms have been cited numerous times (Li, 2002), the most notable example being by Zipf (1932) himself. Studies of grammatical case in other research show a variety of approaches to observing case in a typological fashion. These approaches either deal with the thematic roles of specific case-markers (Primus, 2011) or within the entire case-marking system, examples being morphological richness (Hawkins, 2002) and person-marking (Siewierska, 1998). Legendre et al. (1993) even observed case using Optimality Theory. Case-marking distribution across languages has not been cited in the reviewed literature. An approach to frequency distribution based on Zipf's law provides a new look into the phenomenon of noun case.

This study aims to identify a distributional pattern of case-marking to show a possible connection between linguistic processing and Case in human speech. I observe only the distribution of case-markers and do not make any claims about language universals. Instead, I draw conclusions to support or refute the presence of Zipf's law in case distribution. Analyzing case-marking from a statistical approach of distribution aims to accomplish three goals: 1) to observe and confirm whether case-marker frequency follows Zipf's law, 2) to determine whether an explanation for distributional patterns is supported by a common factor across these languages, and 3) to understand the linguistic processing of Case.

Data

The data in this study comes from the Universal Dependencies (UD) corpus.[1] Of the 96 languages listed there at the time of data collection, 75 languages were annotated with case-marking labels. These languages displayed the presence of 1-22 uniquely identified lexical case-bearing forms. Marked lexical items included nouns, adjectives, pronouns, determiners, adpositions, particles, adverbs, verbs, numerals, and subordinating conjunctions. Certain languages were labeled with specifically named case tags, while others, such as Japanese and Korean, were only marked as bearing a case particle. Those which were not labeled were annotated by hand.

```
# text = “Kalevala” kučču ativoih.
1       “       “       PUNCT   PUNCT   _       2       punct   _       SpaceAfter=No
2       Kalevala        Kalevala        PROPN   PROPN   Case=Nom|Number=Sing    4       nsubj   _       SpaceAfter=No|PropnType=Al
3       ”       ”       PUNCT   PUNCT   _       2       punct   _       _
4       kučču   kučču   VERB    VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 0       root    _       _
5       ativoih ativo   NOUN    NOUN    Case=Ill|Number=Plur    4       xcomp   _       SpaceAfter=No
6       .       .       PUNCT   PUNCT   _       4       punct   _       _
```

Figure 1.
Sample sentence of Karelian in UD with case labels for the nominative and illative.

UD provided accessible data with detailed syntactic and morphological annotations. This data includes disproportionately more Indo-European samples than samples from other language families. These families were observed: Austronesian, Baltic, Celtic, Dravidian, Germanic, Indo-Aryan, Indo-European, Indo-Iranian, Japonic, Koreanic, Mongolic, Niger-Congo, Northwest Caucasian, Pama-Nyugan, Romance, Semitic, Sino-Tibetan, Slavic, Turkic, and Uralic. There is one language isolate present in the data, Basque. Both nominative-accusative and ergative-absolutive alignments were represented. Inflectional, lexical, and mixed systems are represented

[1] https://universaldependencies.org/

in the data. Word order and level of synthesis in these languages varies. Data samples for each language range from 10 items (Abaza) to over a million (Czech, German). The differences in these sample sizes were considered when analyzing the data. Each language's data resides in at least one treebank from UD; language-internal data across multiple treebanks were analyzed as a whole.

Case sorting labels were drawn directly from the UD annotations and not from the descriptions found of languages' case inventories. For Japanese and Korean, annotations were not present for case markers, instead placing them all under a broad case tag. As distribution of each case was needed for the analysis, these were counted based on the case analysis provided by Suzuki and Toutanova (2006). The goal was to identify the most frequently used cases and separate them from the least frequently used cases. Thus, identifying forms serving the same grammatical purpose allowed for a clearer analysis of distribution. For this study, the semantic function of cases was not relevant. Cases were labeled numerically during the analysis based on frequency to remove assumptions drawn from looking at semantic category.[2]

Methods

Case-marker counts were extracted from UD annotations and recorded in a spreadsheet. Data from inflections on lexical items were counted and organized by category (noun, proper noun, adjective, etc.). A count was taken for all case-marked items as well as individually for each lexical category. For certain languages which included co-occurrences of more than one

[2] As cases often do not match up semantically cross-linguistically, this study avoids a semantic analysis for determining a common factor among the data. A genitive in German may serve a different purpose than a genitive in Japanese, though both receive the same name when studied. This difference in semantic scope could skew data, making a broad claim where a closer look at semantic role is necessary.

lexical item being marked at a time, counts were taken for co-occurrences. Distributions by language were initially overlaid on a large graph to show an overall statistical curve reflecting the model based on Zipf's law.

Several individual graphs show the distribution of frequencies in each separate case-bearing language and observe trends between lexical items. The individual items' data was fit to different regression models to compare data across languages. The two models used for comparison were power-law distribution (Zipf's model) and exponential distribution. Both models reflect a non-linear, inversely proportional relationship between the rank of a given case and frequency of a case being marked on a form and the frequency of forms inflected for that case among all case-marked forms. In a Zipfian distribution, the frequency of a case is predicted as the case's rank is raised to a certain constant value. Alternatively, in an exponential distribution, the case rank is the exponent of some constant value, meaning that the rate of decay in frequency is directly proportional to the rank of the case.

Results were sorted according to which regression model showed a closer fit. Separate comparisons reflected all case-marked forms and exclusively nominal forms for each language. Each language's distributions were compared to the regression models and sorted according to best fit. Languages were placed into three categories: 1) Zipfian, 2) exponential, and 3) inconclusive. This categorization was then analyzed to determine distribution patterns. To look for a connection to linguistic processing, categories were analyzed to determine a pattern for model fit. Languages were observed according to language family, data inventory size, number of attested cases, speaker population, language status, alignment, and word order. They were sorted by each of these variables to find a common factor driving the split between the observed distributional models.

Results

Initial results showed that languages displayed between 1 and 4 cases that occurred with a higher frequency than other cases, displaying a significant gap between the most frequent and least frequent cases. Results for the first extraction showed errors which needed to be addressed to continue. A first error analysis was performed at this point to clean up the results.

Some files, when parsed, extracted items which were not case-marking tags, but rather features of animacy, verbal mood, and other unrelated categories.[3] For languages which extracted these elements in the initial parse, the extraction code was adjusted, which revealed case-markings which had not been previously counted in the distribution. Recounts were made, removing inapplicable features and adding in missing data to the final numbers.

Two case names which were not listed in the UD feature guidelines appeared in the data sets, the caritive and approximative cases in Komi Zyrian, neither of which were attested in other sampled languages nor were marked under a different tag in the UD feature guidelines. Hamari (2009) highlights that caritive is an alternative marking for the abessive case in Uralic languages. Aikhenvald (2008) attests for another instance of the approximative case in Kham. As these cases were confirmed to be in the literature, they were not removed from the data set. The corpus also included some mismatched labels for the cases temporal and considerative.[4] These data were unified for clearer counts.

A second set of results showed minor differences, but otherwise maintained a similar pattern to the first. Data across all of the languages still showed that for every case-bearing

[3] One category was "Yes." The "Yes" label on the UD corpus can refer to a few different features, which are binary features. It marks the presence or absence of a certain feature and was therefore unrelated to case-marking.
[4] Some languages marked "Tem" versus "Temp" for temporal; some languages marked "Cns" versus "Con" for considerative.

language, there are between 1-4 cases occurring with very high frequency when compared to the rest of the cases. For some languages, the slope between the most highly frequent case(s) and the following cases was steep. Others showed a steadier slope with a more even distribution as the frequencies approached the final case marker.

Individual results showed a varied image of what is happening with case distribution by language. The number of cases appeared to play a role in the shape of the statistical curve when divided into groups of ≤4 cases and ≥5 cases. Three main trends appeared in the individual data: 1) exponential, 2) Zipfian, and 3) linear (languages with 2 cases).

A second error analysis revealed data miscounts stemming from results that took both a first and second line of UD data instead of just the first. These languages were all parsed again for correct counts using a more specific line of code. During this error analysis, data from more languages and different lexical categories was included to improve results, bringing the number of languages represented from 53 to 75.

This third set of results included older forms of languages (French, Russian) and languages which only bear pronominal case (English, French, Spanish), and languages which are no longer spoken (Ancient Greek). The inclusion of these languages continued to reflect the previous sets of results statistically, showing that languages with a pronominal form of case-marking reflected similar patterns to the first two sets of results.

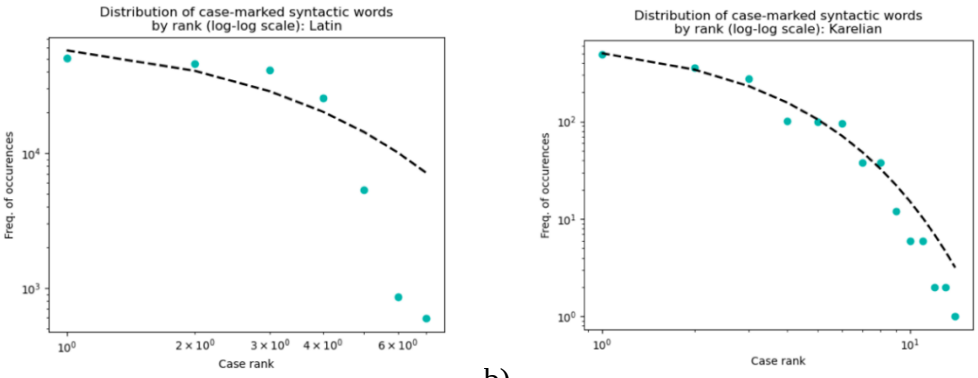
Regression models served as a basis to compare the languages. This took the data from looking at three general patterns graphically to two main theoretic models for comparison. Here it was observed that the data from each language aligned more closely with either a Zipfian or an exponential model, showing a variation between these two distributions. More languages fit best to the exponential model than to the Zipfian model.

From the regression models created from the third analysis, languages were again compared and organized based on shared features, such as language family, data inventory size, number of attested cases, speaker population, language status, alignment, and word order. The results from this comparison were inconclusive, showing no apparent pattern in the data which suggested the split between the two separate statistical models.

Discussion

Results across all observed languages show that between 1 and 4 cases occur at a higher frequency than the subsequent cases in a language. Some languages see a gap in the frequencies between these first 1-4 cases while others continue to decline without significant deviation. The languages are split on which regression models they best fit, and this split shows no evident pattern from observing other factors.

What is most notable about the data is the 1-4 cases which occur most frequently. Each distribution has at least one case that occurs significantly more frequently the other cases, however, there appears to be a trend in a second, third, or even fourth case also having a rather high frequency. This, in certain languages, sometimes creates a plateau in the frequency before the more significant decline in frequency occurs. Languages with a more subtle slope in the trend do not show as significant of a curve. Examples are displayed in Figure 2.



a) b)

Figure 2.

Samples of the data aligning with exponential regression showing a) a plateau dropping off from the 4 highly frequent case-markers in Latin for all case-marked items and b) a subtle slope in frequencies in case-markers in Karelian for all case-marked items.

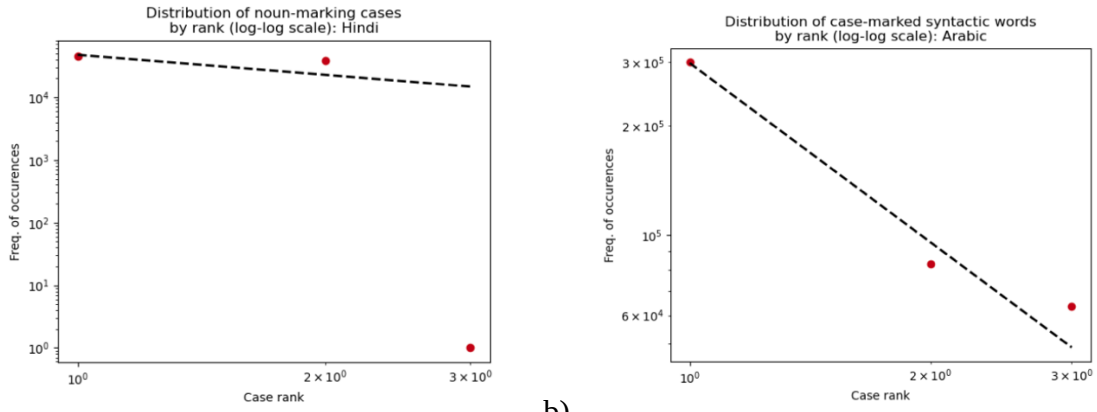
Certain case-markers occurring more frequently than others shows that for every language, there is an instance of 1-4 cases which occur most frequently in each language. Linking this back to linguistic processing, this demonstrates that language shows a tendency towards favoring between 1 and 4 semantic categories of case-marking. This would suggest that marking ≤4 cases is linguistically processed as less marked than marking ≥5 cases.

Languages were fit either into the Zipfian or exponential regression models. Results show more languages fitting to the exponential model than to the Zipfian model. This difference suggests that case cannot be looked at universally from a single trend. This could be largely dependent on the type of system using case; Korean case particles marking the noun can co-occur[5] (Bak, 2004) where other systems will not allow a compounded marking. Other systems only morphologically mark pronouns, while many mark several different word categories.

[5] In Jaehee Bak's dissertation, it is noted that there are instances where two case particles mark the same noun. The abstract calls attention to accusative and genitive (possessor) particles marking the same noun, which is contrary to the traditional view of case in which one case maps to one noun.

However, differences between systems showed no clear pattern. Language family, case inventory, sample size, syntactic word order, and speaker population[6] all showed no bearing on the grouping of languages into these two models. The connecting factor that unites the languages that fit to the Zipfian model is that all of these languages have synthetic case-marking morphology. Languages which were more analytical fit better to the exponential model, though numerous synthetic languages also align with this category. This could suggest that non-synthetic languages will continue to fit to the exponential distribution, though more data is needed to support this.

Patterns in trends differ between languages with ≤4 cases and languages with ≥5 cases. The smaller inventory of cases shows two significant patterns (displayed in Figure 3) in those languages: 1) two cases are highly frequent and one is low frequency (Bhojpuri, Hindi, and Urdu) and 2) one case is highly frequent and two are low frequency (Arabic, Romanian, and Swedish).

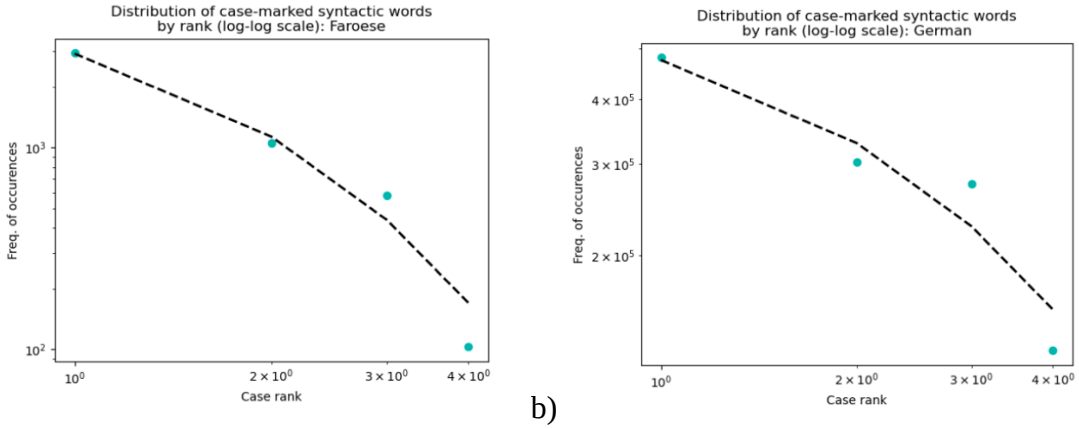


a) b)

Figure 3.
Samples of the data aligning with Zipfian regression showing a) a pair of case-markers in higher frequency compared to a third in Hindi nouns and b) one case-marker in higher frequency compared to a pair of lower frequency case-markers in Arabic for all case-marked elements.

[6] Population of speakers, both L1 and L2 speakers, as reported by Ethnologue.

Languages with four cases show a steady decreasing slope which can be fit to the distributional models; Faroese, Irish, Kurmanji, and Scottish Gaelic show a better fit with the Zipfian model while German is split between the two models based on analysis of overall versus nominal distributions.[7] Examples of the frequencies in languages with 4 cases can be seen in Figure 4 below.



a) b)

Figure 4.
Note the similar slope patterns between the case-marker frequencies between a) Faroese and b) German when comparing all case-marked forms with an exponential regression. Observing nouns alone in German fit better to the Zipfian model.

The languages with one or two cases were outliers in that they either did not show regression or had a steep, expected decline between the two cases.

Clearer patterns were present among the languages with more cases. Unlike with an inventory of fewer cases, the number of cases and the pattern shown in the trend do not appear to have a correlation. Languages which fit to the Zipfian model best had between 4 and 22 cases. Those which fit best to the exponential model had as few as 3 and as many as 17 cases. Sample size was expected to contribute to these differences in patterns, however, no such evidence is

[7] German nominal data fit best to the Zipfian model, but the total case-marking data fit better to the exponential model, showing that trends among parts of speech can differ within an individual language's data set.

present from the current data. A comparison of the sample sizes shows both large data sets and small data sets present in both categories of distributional models.

From these findings, it shows that the number of cases in high distribution is not affected by number of cases. There are examples of Zipfian and exponential distributions attested in both languages with few cases and languages with many. The only apparent finding is that languages with ≥5 cases can have up to 4 cases which occur in high frequency. This becomes more apparent as languages with larger case inventories are observed. The significance of the most frequent cases is clearer to see as the number of cases in the language increases. From this, case number appears to restrict a certain number of cases which occur in high frequency together, potentially those semantically linked to verbal arguments.

This data is inconclusive on the typological distinction between the two models found to fit to the data set. From the data observed, no clear pattern arises to suggest why certain languages fit better to a Zipfian or exponential model. It can be concluded from this data set that, statistically speaking, a Zipfian distribution of case-markers is less frequent than an exponential one.

As stated above, the existence of 1-4 highly frequent cases in each language is significant to the second and third goals this study aimed to answer. No language showed a fifth case in high distribution, capping each language's highly frequent items at 4. Considering the linguistic concept of markedness, it appears that having >4 cases may be marked, thus causing a significant gap between a highly frequent fourth case and a considerably less frequent fifth case. This could explain the lack of drastic gaps in frequency in languages with ≤4 cases. In conjunction with this, this finding suggests that up to 4 cases are more cognitively accessible, while more marked cases are less so.

In terms of case categories, these 1-4 cases are used in high frequency regardless of label or semantic category. This phenomenon occurs independently of case semantics, and a look at the numerical data can show that each most frequent case is not the same semantic class. While there are a few repeating highly frequent cases from a semantics standpoint, it is more significant that regardless of semantic category, every language stops having a similar frequency at the fifth most frequent case.

Conclusion

Results show that across all tested languages, up to the first 4 cases have the highest frequency and the subsequent cases decrease in frequency. Further work will be needed to form conclusions about what determines the separation into two distribution patterns. The most significant finding is that there is a higher distribution shared between 1-4 cases in each language, which appears to have a restriction on languages with ≥9 cases that results in a plateau in the graphical data. This could indicate that there is an upper limit to how many cases may be in a similar high frequency with one another.

Further analysis shows that the exponential distribution fits better to more languages than the Zipfian model. The reasoning for this is unclear from these findings. Several factors were observed to locate a pattern that united the two separate statistical models, but the closest indication was related to morphological synthesis. As no analytic language had a better fit to the Zipfian model, it could suggest that there is a pattern present among analytic languages that prevents a Zipfian distribution, but more data is needed.

Moving forward, this raises the question of why the distributions of case did not fit to the expected Zipfian model. This study could be lacking in data, though without a pattern in data sets

affecting the results, it could also reflect that case-marking does not follow Zipf's law. As noted in the discussion, it appears that regardless of the differences in the distributions, humans are processing up to 4 cases more frequently than other cases. For each language, this could show which semantic roles are most communicatively useful.

With more data, results may show that languages may fit better to one model over the other. It was seen in languages like German that case observed on a single word class could reflect a different best fit than the data taken as a whole. This study presents a look at the entire case-marking system of a language and provides a look at distributions of only nominals for comparison. Observing those individual patterns may answer the question as to why no clear pattern exists from looking at the data in its entirety.

References

Aikhenvald, A. (2008). Versatile cases. *Journal of Linguistics*, *44*(3), 565–603.

Bak, J. (2004). *Optional case marking of the possessor in Korean: A study of double accusative marking of possessor/possessee in Korean based on corpus study and optimality computation*. (Master's thesis).

Hamari, A. (2009). The abessive case of the Uralic languages. In *Book of Abstracts*, 35.

Hawkins, J. (2002). Symmetries and asymmetries: Their grammar, typology and parsing. *Theoretical Linguistics*, *2*(28), 95–150.

Legendre, G., Raymond, W., and Smolensky, P. (1993). An optimality-theoretic typology of case and grammatical voice systems. In *Annual Meeting of The Berkeley Linguistics Society, 19*(1), 464-478.

Li, W. (2002). Zipf's law everywhere. *Glottometrics*, *5*, 14–21.

Manning, C. and Schütze, H. (1999). Foundations of statistical natural language processing. *Computational linguistics: Statistical methods, 1,* 23-25.

Primus, B. (2011). Case-marking typology. In *The Oxford Handbook of Linguistic Typology*, 1-35.

Siewierska, A. (1998). On nominal and verbal person marking. *Linguistic typology, 2*, 1-53.

Suzuki, H. and Toutanova, K. (2006). Learning to predict case markers in Japanese. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annynal Meeting of the Association for Computational Linguistics* (pp. 1049–1056).

Zipf, G. K. (1932). Selected studies of the principle of relative frequency in language.

311 **Appendix – Sample of Case Frequency Counts**

LANG	1	2	3	4	5	6	7	8	9	10	11	1	1	1	1
UAGE												2	3	4	5
ARA	274	757	596												
N/ADJ	356	93	99												
HYE	628	502	746	78	657	599									
N/DET	0	9													
EUS	110	367	272	226	170	770	64	62	48	47	47	7	7	2	1
N/DET	01	9	9	4	7		2	6	9	0	0	8	4	7	2
EUS	123	359	235	231	182	789	68	56	56	48	46	7	6	6	1
N/ADJ	93	3	0	2	2		5	2	0	3	4	1	8	4	6
BEL	172	681	678	638	342	96									
N/DET	5														
BEL	236	934	903	805	493	115									
N/ADJ	4														
BHO	122	103													
N/ADP	6	2													
BHO	119	711	6	1											
N/DET	9														
BHO	113	723	1												
N/ADJ	7														
HRV	178	147	123	656	282	129	22								
N/DET	78	77	89	2	7	3									

HRV	229	200	149	781	359	160	27								
N/ADJ	80	71	52	8	1	8									
CES	159	135	106	649	444	210	43								
N/DET	663	456	911	26	19	06	1								
CES	218	170	132	825	589	255	50								
N/ADJ	187	675	601	56	52	96	3								
EST	369	330	208	791	784	552	48	45	32	19	11	7	6	4	2
N/DET	50	45	62	2	0	7	53	23	89	26	70	0	9	0	8
												7	0	7	6
EST	481	404	242	876	869	630	50	48	48	19	13	8	7	4	4
N/ADJ	00	61	84	8	8	5	96	84	59	82	21	9	3	1	0
												5	5	7	4
FAO	161	687	309	45											
N/DET	6														
FAO	199	765	381	48											
N/ADJ	9														
FIN	173	102	100	287	284	258	21	14	12	85	55	1	8	6	
N/ADJ	15	20	04	3	9	3	26	36	28	3	4	9	8	1	
												3			
DEU	164	155	146	106											
N/DET	322	065	082	143											
DEU	758	441	390	372											
N/ADJ	58	82	96	48											

GOT	375	334	328	208	202								
N/DET	1	2	9	2									
GOT	441	411	406	214	224								
N/ADJ	8	5	8	2									
ELL	119	635	590	173	37								
N/DET	92	0	8										
ELL	870	452	441	186	40								
N/ADJ	3	2	6										
HIN	598	513	121	593	256	42	25	2	1	1			
N/ADP	12	92	5										
ISL	698	602	515	212									
N/ADJ	86	13	04	96									
GLE	437	236	36	28									
N/DET	3	9											
GLE	475	183	36	28									
N/ADJ	8	6											
KRL	294	231	230	77	76	73	36	14	12	5	1	1	1
N/ADJ													
LAT	672	585	614	340	757	856	72						
N/ADJ	85	41	52	08	2		1						
LAV	193	148	138	903	628	55							
N/DET	07	34	01	3	6								

LAV	200	178	147	938	670	58									
N/ADJ	61	97	33	0	1										
LIT	104	528	411	157	132	111	16	9							
N/DET	40	0	0	3	4	6									
LIT	116	629	465	184	137	121	17	9							
N/ADJ	08	7	5	8	8	9									
OLO	142	111	77	39	31	31	25	18	16	13	11	3	2	2	1
N/ADJ															
MAG	110	400	207	127	16	13									
N/DET	2														
MAG	108	391	200	127	16	13									
N/ADJ	6														
MAR	418	185	161	69	56	40	13	4	2	1					
N/ADJ															

312