Inflectional Syncretism and Corpora

D.Brown, C. Tiberius, G.G. Corbett
Surrey Morphology Group
University of Surrey
Guildford, Surrey, UK
GU2 7XH
{d.brown,c.tiberius,g.corbett}@surrey.ac.uk

Abstract

This paper describes a novel undertaking: comparing the relationship between grammatical ambiguity (syncretism) in nouns, as represented in a default inheritance hierarchy, with textual frequency distributions. In order to do this we consider a language with a reasonable number of grammatical distinctions and where syncretism occurs in different morphological classes. We investigated this relationship for Russian nouns. Our results suggest that there is an intricate relationship between textual frequency and inflectional syncretism.

1 Introduction

The treatment of syncretism, where a single form has more than one function, poses a particular challenge for theories of morphology. There are different ways of analyzing syncretism. One way is underspecification, where the form in question is treated as not realizing the syncretized functions. Another way is referrals (Zwicky 1985; Stump 2001: 212-41), where the form is associated with a basic function, and other cells in the paradigm refer to the cell with this basic function. Referrals are therefore asymmetrical in their nature, whereas underspecification is not. There is evidence that both types of analysis are required (Stump 2001: 212-41) and Evans, Brown & Corbett (2001: 216) argue that a kind of underspecified referral is required for analyzing syncretisms in Slovene and Dalabon. Therefore we cannot dispense with one at the expense of the other. It is therefore worthwhile examining whether the theoretical asymmetry of referrals can be observed in language use.

We consider the relationship between the frequency distributions of noun syncretism in texts and a formal implemented model of Russian inflection, created for other purposes. In particular we determine whether the asymmetry of referrals is reflected in the frequency distributions (i.e. does the referred-to cell in the paradigm typically occur more frequently than the cell which refers to it). Russian is an ideal candidate for testing this relationship, as it is a language with substantial paradigms and extensive syncretism of different types.

We used a manually compiled dataset (Corbett, Hippisley, Brown and Marriott 2001) of the most frequent nouns derived from the Uppsala corpus (Lönngren 1993; Maier 1994) containing 5440 noun lexemes, accounting for 243,000 word forms from the entire one million word corpus. On the basis of this dataset, we tested a set of hypotheses and established a model describing the relationship between paradigm structure and textual use in Russian nouns.

We cross-validate our results by testing the model on a pilot version of the Russian Standard Corpus (Sharoff 2004) which has been fully lemmatized and tagged.

The paper is structured as follows. Section 2 outlines the problem. We describe what we mean by syncretism and frequency and define a set of hypotheses. In Section 3, we test our hypotheses on the two corpora. Section 4 concludes the paper.

2 Morphological Theory and Frequency

2.1 Syncretism

Grammatical paradigms define the relationship between functions and forms. A reasonable assumption would be that functions and forms would match up one-to-one. Contrary to these expectations, however, languages often fail to meet this apparent ideal of one form for every function. In Russian nouns, for instance, there is a smaller number of forms.

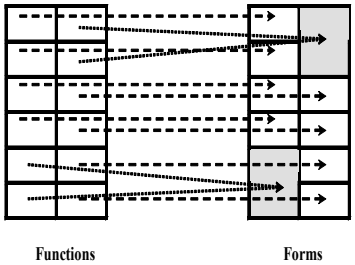

Figure 1: More functions than forms

Russian nouns have two number values (singular and plural) and six cases (nominative, accusative, genitive, dative, instrumental, and locative) which can be combined yielding 12 combinations of case and number. Despite the 12 combinations of case and number, a typical noun does not have more than 10 forms (Figure 1) in Russian[1], but the class in question will determine where the distinctions between functions are collapsed. For example, the form *комнате* (*komnate*) 'room' is ambiguous between dative singular and locative singular, the form *кости* (*kosti*) 'bone' is three-way ambiguous; it can function as a dative singular, a locative singular and a genitive singular.

Similarly, in English we expect that a verb will have forms for the present tense, the past tense and the passive participle. Given an ambiguous form, such as *hit* (as in 'Mary *hit* the nail with a hammer' as opposed to 'Mary was *hit* by a meteorite') we assume that there are two or more separate functions associated with it. This is because other verbs have different forms, such as the verb *eat* (as in 'John often *ate* chocolate' as opposed to 'The chocolate was *eaten* by John').

Formally syncretism can be treated as underspecification, where the multiple functions of the form have equal importance, or it can be seen in terms of referral (Zwicky 1985; Stump 2001: 212-41), where the form is associated with a basic function, and other cells in the paradigm refer to the cell with this basic function. Underspecification implies equal status for the functions in question, whereas the referral-based approach attributes greater importance to one function.

One measure of the relative importance of grammatical functions is textual frequency. We can make two different predictions about how the formal distinction between underspecification and referrals is reflected in textual frequency: Prediction 1) both functions are of equal importance (underspecification) and therefore the frequency of use is distributed equally between functions; or Prediction 2) one function is more important than another (referral) and therefore the frequency of use is not distributed equally between functions.

In this paper we test these predictions for Russian nouns. As the basis for our predictions, we start from a formal theoretical model of Russian morphology developed within the Network Morphology framework.

[1] We exclude from consideration here the question of the second locative and the second genitive.

2.2 Inheritance Hierarchy

In this model four noun classes are distinguished for Russian as is shown in Table 1.[2]

	I	**IV**	**II**	**III**
	завод	дело	комната	кость
	zavod	delo	komnata	kost´
Sing	'factory'	'thing'	'room'	'bone'
Nom	завод	дело	комната	кость
	zavod	delo	komnata	kost´
Acc	завод	дело	комнату	кость
	zavod	delo	komnatu	kost´
Gen	завода	дела	комнаты	кости
	zavoda	dela	komnaty	kosti
Dat	заводу	делу	комнате	кости
	zavodu	delu	komnate	kosti
Instr	заводом	делом	комнатой	костью
	zavodom	delom	komnatoj	kost´ju
Loc	заводе	деле	комнате	кости
	zavode	dele	komnate	kosti
Plur				
Nom	заводы	дела	комнаты	кости
	zavody	dela	komnaty	kosti
Acc	заводы	дела	комнаты	кости
	zavody	dela	komnaty	kosti
Gen	заводов	дел	комнат	костей
	zavodov	del	komnat	kostej
Dat	заводам	делам	комнатам	костям
	zavodam	delam	komnatam	kostjam
Instr	заводами	делами	комнатами	костями
	zavodami	delami	komnatami	kostjami
Loc	заводах	делах	комнатах	костях
	zavodax	delax	komnatax	kostjax

Table 1: Forms for major noun classes in Russian

We see that Class III nouns such as *кость* 'bone' have the highest ambiguity, three forms are used for six functions, one for the nominative and accusative, one for the genitive, the dative and the locative, and one for the instrumental. Class II nouns such as *комната* have syncretism in the dative and locative singular. Russian also has syncretism related to animacy. In the singular in Russian, masculine animate nouns, which belong to Class I, form their accusative on the basis of the genitive form. For classes IV and III, which are associated with neuter and feminine genders respectively, there is always nominative-accusative syncretism. Class II, of course, has a separate form for the accusative. In the plural, the situation is more straightforward: any animate noun forms its accusative on the basis of the genitive, and any inanimate noun forms its accusative on the basis of the nominative. In the above table only examples of inanimate nouns are given.

Counts based on both corpora show that, about 38%[3] of token occurrences in the singular in

[2] We use the following abbreviations: NOM – nominative, ACC – accusative, GEN – genitive, DAT – dative, INSTR – instrumental, LOC – locative. Forms are given in Cyrillic and in transliteration.

[3] This figure is obtained by summing the counts for

Russian are morphologically ambiguous (syncretic). Hence, in an apparently morphologically well-equipped language such as Russian, a great deal still depends on syntactic context.

Our default inheritance analysis of Russian treats syncretisms within paradigms as asymmetrical, in that a particular form is considered to have one function as basic. If this paradigmatic asymmetry is reflected in frequency distributions, then Prediction 2 would be the correct one. As the syncretisms within paradigms differ from class to class, we also have to take into account the hierarchical structure of morphological classes. Figure 2 represents the hierarchical structure of our model of Russian noun morphology using the Network Morphology framework (Corbett and Fraser 1993), with example lexemes at the bottom of the hierarchy.

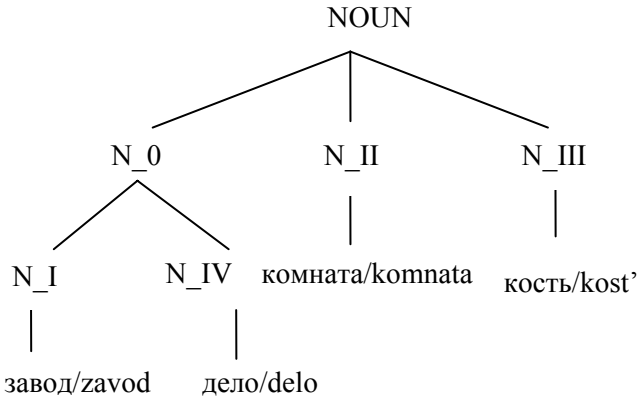

Figure 2: Inheritance structure for Russian nouns

Two points should be stressed. First, the original analysis was carried out with the goal of contributing to morphological theory, a goal which was achieved (see comments in Stump 2001:275-6). Second, in order to demonstrate that the analysis was valid, a substantial fragment of Russian, sufficient to include all instances of irregularity was implemented in the lexical knowledge representation language DATR (Evans and Gazdar 1996) and is available at the DATR archive from the DATR webpages (http://www.datr.org). The morphological model is structured as a hierarchy in which information is pushed as far up the hierarchy as it can go capturing as many generalisations as possible. Thus at the top of the hierarchy in Figure 2 we find a information associated with all nouns (such as the inflections for the dative, instrumental and locative plural) and that information is propagated to others by inheritance, and at the bottom we find information which is unique to particular instances. Node N_I (representing Class I) and node N_IV (representing Class IV) both inherit from node N_0 (representing Class O), sharing the inflections for the genitive, dative, instrumental and locative singular. DATR is a default inheritance formalism, which means that information specified under a particular class node takes precedence over that what is inherited, overriding the inherited information. This means that, for example, the value for locative singular which is *stem+e* in three of the four inflectional classes can be stated under the noun node and its value only needs to be overridden for Class III by *stem+i*.

We will use this model and link frequency information to the different components of the noun hierarchy, i.e nouns at the top, the different noun classes and the lexemes belonging to these noun classes at the bottom.

2.3 Research Questions

To check our predictions, we defined a set of hypotheses which we tested on two sets of corpus data.

The hierarchical organisation of paradigms.

Our inheritance model of Russian morphology involves a hierarchical organisation above the level of traditional word classes. A major question is whether this hierarchical organisation of paradigms is reflected by differences in frequency of use. Thus we will investigate the frequency of use of grammatical functions for each point on the hierarchy in Figure 2, and in particular we shall look at the frequency of use of those functions which are syncretic (grammatically ambiguous) at various points on that hierarchy. This research question will be studied by testing the following null hypothesis:

> H_0-*1*: Nodes at the same point in the hierarchy show no difference in frequency

paradigm cells which share the same form and dividing this by the overall number of singular tokens. In other words we use the sum of the token occurrences of the following ambiguous forms: i) all dative and locative singular occurrences of Class II nouns; ii) all nominative and accusative singular occurrences of Class IV nouns, Class III nouns and Class I nouns which are inanimate; iii) all accusative and genitive singular occurrences of Class I nouns which are animate; iv) all genitive, dative and locative singular nouns from Class III. It is likely that one can disambiguate most morphological syncretisms in Russian readily from the syntactic context, but our purpose is to demonstrate how often a morphologically well endowed language such as Russian still leaves much work to syntax.

distribution for any elements of their paradigms.

However, related research by Corbett et al (2001) who investigate the relationship between irregularity types and frequency suggests that we will find significant differences between nodes in the hierarchy. We therefore predict that the null hypothesis is false. There should be differences in frequency distributions for paradigms between nodes at the same point on the hierarchy. For example, for nominals there should be differences between nouns and adjectives. Within nouns, there should be differences between N_O (class N_I and class N_IV combined), class N_II and class N_III. This leads us on to a more specific hypothesis that specifies the nature of these differences.

> H_1-*1*: If two or more functions $f_1 \ldots f_z$ share the same form in one class, but not in another, then the combined relative frequency of those functions is higher for the classes where they are differentiated.

Syncretism and asymmetry of use

Do syncretisms occur where one function is more important than the other(s), but not where equal in status (as defined in terms of frequency of use)? Here we formulate the following null hypothesis:

> H_0-*2*: Where two or more functions share the same form, there is no difference in the relative frequency of those functions.

Again, we expect the null hypothesis to be false.

The context of syncretism

Which category is the most important factor for syncretism? According to Haspelmath (2001: 241) the frequency of a category determines the prevalence of syncretism within that category. Frequent grammatical categories, such as singular number, are more differentiated and thus show less syncretism than less frequent categories. This research question will be tested by the following null hypothesis.

> H_0-*3*: The frequency of a category has no influence on the prevalence of syncretism within that category.

Following Haspelmath (2001), we predict that the null hypothesis is false. Based on our findings in Corbett et al (2001) we would predict that syncretism behaves in a similar way to various types of irregularity. That is, examples of syncretisms are found in the singular and plural in Russian, and the most important factor influencing syncretism will be the relative frequency distribution of the number subparadigm in which the syncretism is found, rather than the relative frequency distribution of the syncretic cells in question. Note that this hypothesis interacts with hypothesis H_1-*1*. If there were no evidence for that hypothesis, but there were evidence that H_0-*3* is false, then this would suggest that the number subparadigm is the most important factor.

3 Corpus Study and Results

To test the above hypotheses we looked at two separate sets of corpus data so as to cross-validate our results.

3.1 Corpus data

The first dataset consists of nouns from the Uppsala corpus. The Uppsala corpus is a set of sub-corpora of various genres, containing approximately one million words (Lönngren 1993, Maier 1994). The dataset was manually constructed by Andrew Hippisley[4] to investigate the relationship between number availability and number use. It is in the form of a spreadsheet and contains frequency information, case and number features, as well as semantic information, i.e. animacy category for 5440 lexemes, accounting for around 243,000 word forms from the entire one million word Uppsala corpus. The lexemes recorded in the dataset are those represented by word forms occurring in total at least five times. Lexemes occurring less than five times were excluded to avoid large standard errors in the estimates which occur when observed numbers in each category are small (Corbett, Hippisley, Bown and Marriott 2001:208).

The lexemes in the original dataset were transliterated. We have converted them into Unicode Cyrillic using the simplest encoding of unicode in data files, the "ucs2" encoding. In addition, we have included class information according to the Network Morphology analysis (Figure 2).

The second dataset consists of the nouns of a pilot version of the Russian Standard Corpus which is fully lemmatized and tagged. Based on this information a spreadsheet was created similar to the one we have for the Uppsala corpus. For consistency, we only took those nouns into account

[4] The dataset is available at `http://www.surrey.ac.uk/LIS/SMG`, along with a readme file.

which occur more than five times. This results in 3350 lexemes, making up 126,598 word forms of the 500 000 word corpus.

3.2 Results

We tested the hypotheses given in Section 2.3. We repeat the hypotheses here with the cross-validated results.

Hierarchical organisation of paradigms

> H_0-*1*: Nodes at the same point in the hierarchy show no difference in frequency distribution for any elements of their paradigms.

As expected, hypothesis H_0-*1*is false. There are differences in the frequency distributions, both in relative and absolute terms, of grammatical functions for classes represented as nodes at the same point in the hierarchy in Figure 2. For instance, there are differences in frequency of the singular and plural between noun classes at the same level in the hierarchy (See Table 2 in the Appendix). Class III has the lowest number of plurals in absolute terms (3421 in corpus 1 and 954 in corpus 2), and also in relative[5] terms (15.6% in corpus 1 and 14.5% in corpus 2). This contrasts with Class O which has the highest number of plurals in absolute terms (42677 in corpus 1 and 12241 in corpus 2), and also in relative terms (30.3% in corpus 1 and 23.9% in corpus 2).

We now turn to the dependent hypothesis H_1-*1*

Hierarchical organisation of paradigms(cont.)

> H_1-*1*: If two or more functions f_1... f_z share the same form in one class, but not in another, then the combined relative frequency of those functions is higher for the classes where they are differentiated.

In order to test this, we looked at the frequency distributions of the genitive singular, dative singular and locative singular for the different inflectional classes and calculated the proportion of the singular paradigm which they account for (See Appendix). As Class III has syncretism of all three of these functions, Class II has syncretism of two of them (dative singular and locative singular), and Class O differentiates all three, if the hypothesis were correct, then Class O should have the highest combined relative frequency for these functions, thereby accounting for a greater proportion of its singular paradigm. However, there is no evidence to support hypothesis H_1-*1*, because the combined relative frequency differs for the classes across the two corpora. In corpus 1 Class O shows the highest relative frequency, Class II the next, and Class III the least high relative frequency (which would support the hypothesis). In contrast, in corpus 2 Class III shows the highest relative frequency for the three functions.

The next hypothesis concerns the relationship between syncretism and the importance of the functions involved. If the hypothesis is false, then this is support for prediction 2 that syncretisms occur where one function is more frequent than another.

Assymmetry of use

H_0-*2*: Where two or more functions share the same form, there is no difference in the relative frequency of those functions.

Again, we looked at the frequency distributions of the genitive, dative and locative singular. In Table 1, we saw that the lexemes belonging to Class N_I and Class N_IV (which are grouped together under a Class N_O in Figure 2) have the same forms, but they unambiguously distinguish all three cases. Class N_II has ambiguity between dative singular and locative singular, and Class N_III completely fails to differentiate the three combinations. In the default inheritance representation of Russian nominal morphology, the default specification of locative singular at NOUN is *stem* + *e*, as this is used across three inflectional classes. In classes II and III, where the dative singular has the same ending as the locative singular, the dative singular will take over its ending from the locative singular, and in Class III where the locative singular has the same ending as the genitive singular, the locative singular will take over its ending from the genitive singular.

Hence, where there is syncretism, we can claim that a form is primarily associated with a particular function, but may be used by another function. In the case of Class III, composition of the referrals means that the dative singular takes over the locative singular, and this in turn takes over the genitive singular, giving the form *stem* + *i* for all three functions. According to prediction 2, we expect that this asymmetry, where one cell refers to another for its form, is reflected in the frequency of use of the particular functions in question. This is true for Class III, which has all three functions syncretic. But it is also true for nouns as a whole, including the Class O which has different forms for each function. In Figure 3 we see a decrease in frequency of use from genitive singular, through locative singular to dative singular.

[5] Relative frequency with respect to the class.

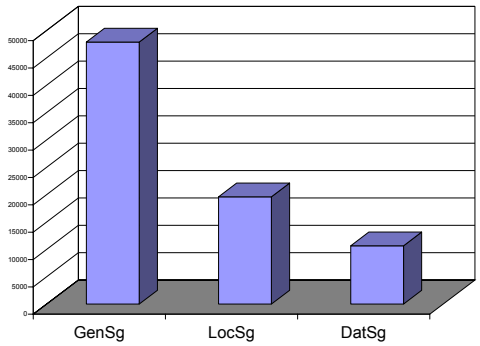

Figure 3: Frequency related to the NOUN node

Hence, if there is a relationship between syncretism and frequency, this asymmetry may be a necessary, but not a sufficient, condition.

The final hypothesis we tested concerned the frequency of the category which provides a context for syncretism.

> H_0-*3*: The frequency of a category has no influence on the prevalence of syncretism within that category.

Class III which has the greatest amount of ambiguity in the singular has the lowest absolute frequency of singulars. This means that the hypothesis H_0-*3* is false.

A note on animacy

Animacy is a distinction which is orthogonal to the individual morphological classes in Figure 2. We were not in a position to cross-validate the results for nominative-accusative syncretism in inanimate nouns, as only corpus 2 has disambiguated the nominative and accusative for inanimates. The results for corpus 2 suggest (Table 4), however, that there is no directional effect for inanimates. We were in a position to check the animates, for which the genitive-accusative syncretism had been disambiguated in both corpora.

In the DATR representation there is a referral of the accusative singular for masculines of Class I to the genitive singular, and there is a referral of the accusative plural to the genitive plural for all animate nouns. Our corpus study showed that the null hypothesis H_0-*2* is also false when we look at animates in the plural: for animates the accusative plural is less frequent than the genitive plural in both corpora (Table 5), which, given the referral of the accusative plural to genitive plural, is the same effect we have seen for genitive singular, locative singular and dative singular in Figure 3 (the referred-to element of the paradigm is more frequent). Recall that, in the singular, only masculine nouns have genitive-accusative syncretism for animates. Our results proved to be very interesting for the singular. For animate nouns as a whole (i.e. across all the morphological classes), the two corpora did not match: while genitive singular was more frequent in corpus 1, there was little difference between genitive singular and accusative singular for all animates in corpus 2 (Table 5). However, if the comparison is restricted to precisely the class which may have the genitive-accusative syncretism, namely Class I, the two corpora match in that they both show genitive singular as more frequent (Table 6).

4 Conclusion

We wished to see whether there is a relationship between syncretism and its representation in a model based on a default inheritance hierarchy. We investigated this relationship for Russian nouns. Our results showed that there is a difference in the frequency distributions for particular functions at the different nodes in the hierarchy. It is also true that there is a relationship with syncretism when one compares classes at the same point in the hierarchy. Class III compared with classes II and O had the lowest absolute frequency for both singulars and plurals, and the most syncretism in the singular. This could suggest that low absolute frequency is a necessary condition for syncretism.

We have also found evidence that the frequency of use of a particular function reflects directionality in the formal model, in that the function which is referred to for its form (the function with which a form is pimarily associated) has higher textual frequency. This may be a necessary condition, but it is not a sufficient one. Our results show that there is an intricate relationship between syncretism and frequency, one which should be investigated further.

5 Acknowledgements

The research reported here is supported by the Economic and Social Research Council (UK) under grant RES-000-23-0082 ‘Paradigms in Use’. Their support is gratefully acknowledged. We would like to thank Anja Belz, Kees van Deemter, Roger Evans, Gerald Gazdar and Adam Kilgarriff for useful discussion on aspects of this paper, and the anonymous referees for helpful suggestions for its improvement. All errors are, of course, our responsibility.

References

Greville G. Corbett and Norman M. Fraser. 1993. Network morphology: A DATR account of Russian nominal inflection. *Journal of Linguistics* 29: 113-42.

Greville G. Corbett, Andrew Hippisley, Dunstan Brown and Paul Marriott. 2001. Frequency, regularity and the paradigm: a perspective from Russian on a complex relation. In J. Bybee and P. Hopper, editors, *Frequency and the Emergence of Linguistic Structure*, pages 201-226. John Benjamins, Amsterdam.

Nicholas Evans, Dunstan Brown and Greville G. Corbett. 2001. Dalabon pronominal prefixes and the typology of syncretism: a Network Morphology analysis. In G. Booij and J. van Marle, editors, *Yearbook of Morphology 2000*, pages 187-231. Dordrecht: Kluwer.

Roger Evans and Gerald Gazdar. 1996. DATR: A Language for Lexical Knowledge Representation. *Computational Linguistics* 22: 167-216.

Norman M. Fraser and Greville G. Corbett. 1995. Gender, animacy and declensional class assignment: a unified account for Russian. In G. Booij and J. van Marle, editors, *Yearbook of Morphology 1994*, pages 123-150. Kluwer, Dordrecht.

Martin Haspelmath. 2001. *Understanding Morphology*. Edward Arnold, London.

Lennart Lönngren. 1993. *Častotnyj slovar´ sovremennogo russkogo jazyka*. Uppsala University, Uppsala.

Ingrid Maier. 1994. Review of Lönngren Častotnyj slovar´ sovremennogo russkogo jazyka. *Rusistika Segodnja* 1: 130-136.

Serge Sharoff. 2004. Methods and tools for development of the Russian Reference Corpus. In D. Archer, A. Wilson, P. Rayson, editors, *Corpus Linguistics Around the World*. Amsterdam: Rodopi.

Gregory T. Stump. 2001. *Inflectional Morphology*. Cambridge: Cambridge University Press.

Arnold Zwicky. 1985. How to describe inflection. In M. Niepokuj, M. Van Clay, V. Nikiforidou and D. Feder, editors, *Proceedings of the eleventh annual meeting of the Berkeley Linguistics Society*, pages 372-386. Berkeley Linguistics Society, Berkeley.

Appendix

	Number of Tokens in			
	Corpus 1		Corpus 2	
	SG	PL	SG	PL
Class O	98173	42677	38877	12241
Class II	48843	17688	21047	5700
Class III	18522	3421	5619	954

Table 2: Absolute and Relative Frequency

	Corpus 1	Corpus 2
Class O	48.1%	34.7%
Class II	46.4%	32.9%
Class III	42.0%	36.1%

Table 3: Combined Relative Frequency for Gen-Dat-Loc Animate/Inanimate Singular

	Number of Tokens in Corpus 2			
	NomSg	AccSg	NomPl	AccPl
Class O	6873	7804	2050	1987
Class II	3695	5223	1006	1174
ClassIII	1231	1105	202	230

Table 4: Frequency distributions for Inanimate Nouns in Corpus 2

	Number of Tokens in	
	Corpus 1	Corpus 2
GenSg	2966	1569
AccSg	1595	1606
GenPl	3484	1131
AccPl	1003	536

Table 5: Frequency distribtions for Animate Nouns

	Number of Tokens in	
	Corpus 1	Corpus 2
GenSg	2260	1033
AccSg	1107	981

Table 6: Frequency distributions for Class I Animate Nouns