The Romanian Neuter Examined Through A Two-Gender N-Gram Classification System

Liviu P. Dinu[1, 3]
ldinu@fmi.unibuc.ro

Vlad Niculae[1, 3]
vlad@vene.ro

Octavia-Maria Şulea[1, 2, 3]
mary.octavia@gmail.com

[1] Faculty of Mathematics and Computer Science
[2] Faculty of Foreign Languages and Literatures
[3] Center for Computational Linguistics
University of Bucharest

I. How many genders are there in Romanian?

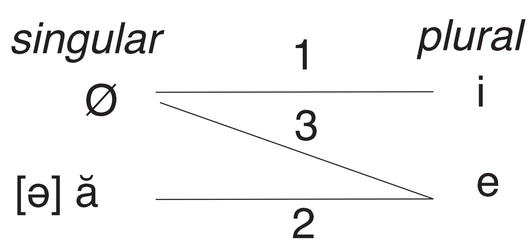
The data shows **three genders** (feminine, masculine, neuter) but only **two agreement patterns** (feminine and masculine)

The "neuter" systematically follows the rule:
singular neuter ~ masculine agreement
plural neuter ~ feminine agreement

	Singular		Plural	
Masculine	un	copac	doi	copaci
	one.**M**	tree.**M**	two.**M**	trees.**M**
Neuter	un	cub	două	cuburi
	one.**M**	cube.**N**	two.**F**	packages.**N**
Feminine	o	amânare	două	amânari
	one.**F**	delays.**F**	two.**F**	delays.**F**

Traditional (3-gender) system: the three genders are marked in the lexicon and different rules for how gender is assigned are posited.

Different gender assignment rules are given by Graur, Corbett, Farkas. For example, from Corbett:



Modern (2-gender) analysis: two gender classes (lexically unspecified) in the singular (**m/f**), and two different classes in the plural (also **m/f**).

Gender assignment in the singular and in the plural is done separately, based on semantic cues (natural gender) and phonology. The neuter corresponds to different assignment in singular and plural

II. Gender classifiers

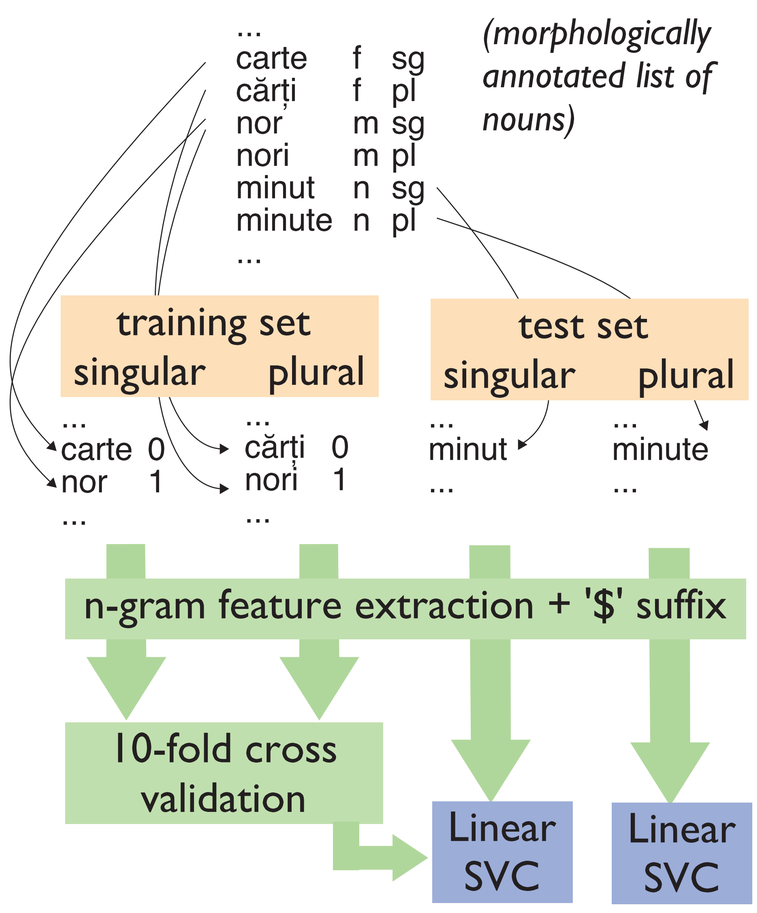
Previous gender classification systems using machine learning are based on the traditional model, *i.e.* three-way classification. We pose the problem as **two** separate binary classifica-tion problems (singular / plural).

Our system's input contains both the singular and the plural form.

We have better accuracy because the neuter is almost indistinguishable from the masculine in the singular form.

To test the two-gender approach, we checked whether the neuter forms classify as masculine in the singular and as feminine in the plural.

III. Training the models



(morphologically annotated list of nouns)

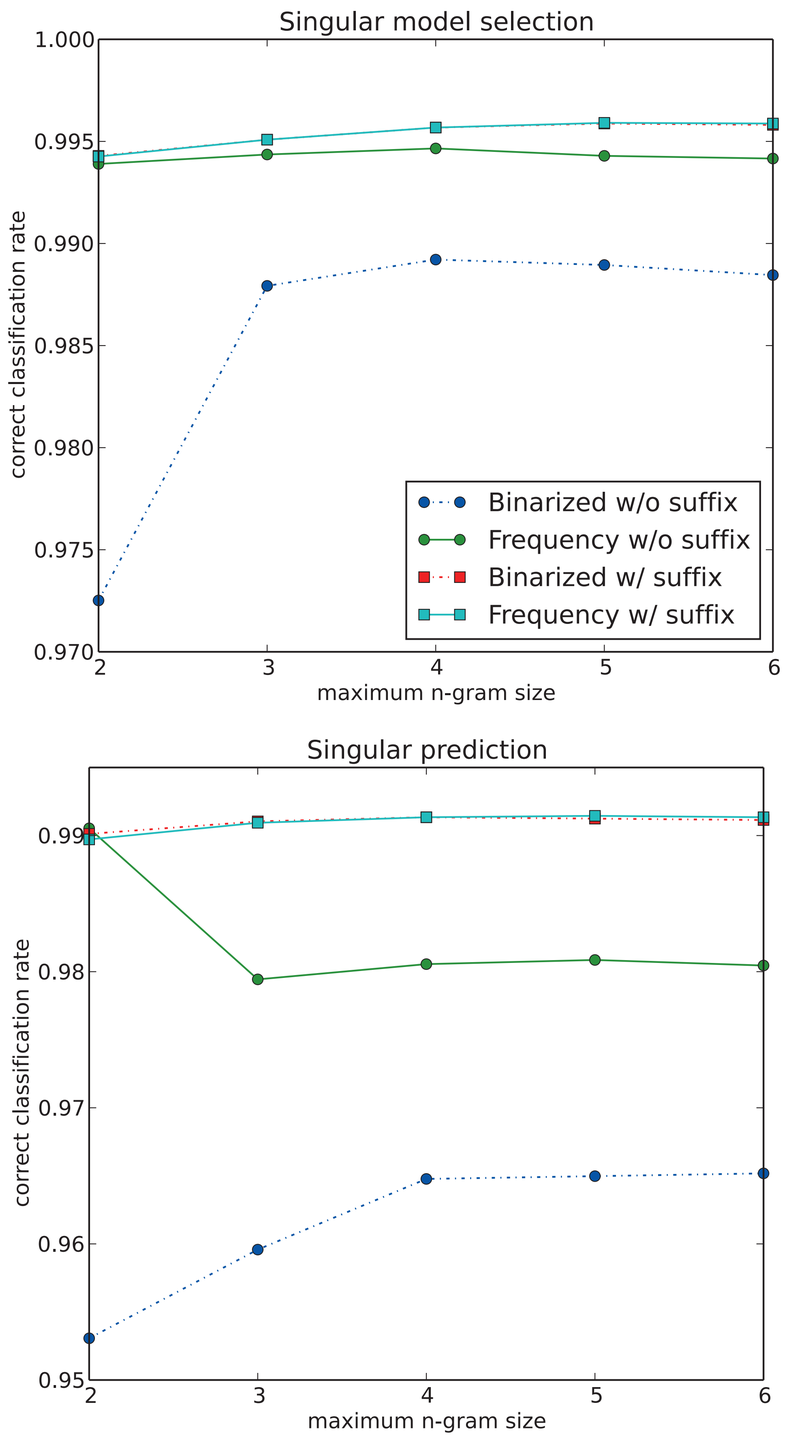
IV. Results & Conclusions

The chosen system parameters are:
5-grams, no binarization, append '$' suffix.

The scores estimated by cross-validation:
singular model: accuracy: 99.59%
precision: 99.63%, recall: 99.80% F_1: 99.71%
plural model: accuracy: 95.98%
precision: 97.32%, recall: 97.05% F_1: 97.18%

Evaluating the neuter nouns: 99.14% accuracy in the singular and 92.30% in the plural. The interesection accuracy is 91.60%.

The plural is less reliable due to hyphenated compounds and confusing endings: balaur/balaur**i** vs. bord/bord**uri**.



sg/pl	*f*	*m*
m	8997	741
f	69	15

Contingency table describing the results of the singular and the plural classifiers on pairs of neuter singular and plural forms.

The 15 complete misclassifications are interesting. 10 of them are french borrowings that kept near-original form: café/caféuri and there-fore, do not follow standard patterns. The rest either have abnormal labels in the dataset or look like mistakes to native speaker

Our system showed that the modern analysis performs better in gender assignment.