What's in a name? In some languages, grammatical gender

Vivi Nastase
EML Research gGmbH
Heidelberg, Germany
nastase@eml-research.de

Marius Popescu
Department of Mathematics and Computer Science
University of Bucharest, Bucharest, Romania
mpopescu@phobos.cs.unibuc.ro

Abstract

This paper presents an investigation of the relation between words and their gender in two gendered languages: German and Romanian. Gender is an issue that has long preoccupied linguists and baffled language learners. We verify the hypothesis that gender is dictated by the general sound patterns of a language, and that it goes beyond suffixes or word endings. Experimental results on German and Romanian nouns show strong support for this hypothesis, as gender prediction can be done with high accuracy based on the form of the words.

1 Introduction

For speakers of a language whose nouns have no gender (such as modern English), making the leap to a language that does (such as German), does not come easy. With no or few rules or heuristics to guide him, the language learner will try to draw on the "obvious" parallel between grammatical and natural gender, and will be immediately baffled to learn that *girl* – *Mädchen* – is neuter in German. Furthermore, one may refer to the same object using words with different gender: *car* can be called *(das) Auto* (neuter) or *(der) Wagen* (masculine). Imagine that after hard work, the speaker has mastered gender in German, and now wishes to proceed with a Romance language, for example Italian or Spanish. He is now confronted with the task of relearning to assign gender in these new languages, made more complex by the fact that gender does not match across languages: e.g. *sun* – feminine in German (*die Sonne*), but masculine in Spanish (*el sol*), Italian (*il sole*) and French (*le soleil*); *moon* – masculine in German (*der Mond*), but feminine in Spanish (*la luna*), Italian (*la luna*) and French (*la lune*). Gender doesn't even match within a single language family: *travel* – masculine in Spanish (*el viage*) and Italian (*il viaggio*), but feminine in Portuguese (*a viagem*).

Grammatical gender groups nouns in a language into distinct classes. There are languages whose nouns are grouped into more or less than three classes. English for example has none, and makes no distinction based on gender, although Old English did have three genders and some traces remain (e.g. *blonde, blond*).

Linguists assume several sources for gender: (i) a first set of nouns which have natural gender and which have associated matching grammatical gender; (ii) nouns that resemble (somehow) the nouns in the first set, and acquire their grammatical gender through this resemblance. Italian and Romanian, for example, have strong and reliable phonological correlates (Vigliocco et al., 2004b, for Italian). (Doca, 2000, for Romanian). In Romanian the majority of feminine nouns end in *ă* or *e*. Some rules exists for German as well (Schumann, 2006), for example nouns ending in *-tät, -ung, -e, -enz, -ur, -keit, -in* tend to be feminine. Also, when specific morphological processes apply, there are rules that dictate the gender of the newly formed word. This process explains why *Frau* (woman) is feminine in German, while *Fräulein* (little woman, miss) is neuter – *Fräulein = Frau + lein*. The existing rules have exceptions, and there are numerous nouns in the language which are not derived, and such suffixes do not apply.

Words are names used to refer to concepts. The fact that the same concept can be referred to using names that have different gender – as is the case for *car* in German – indicates that at least in some cases, grammatical gender is in the name and not the concept. We test this hypothesis – that the gender of a noun is in its word form, and that this goes beyond word endings – using noun gender data for German and Romanian. Both Romanian and German have 3 genders: masculine, feminine and

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1368–1377,
Singapore, 6-7 August 2009.

neuter. The models built using machine learning algorithms classify test nouns into gender classes based on their form with high accuracy. These results support the hypothesis that in gendered languages, the word form is a strong clue for gender. This supplements the situation when some concepts have natural gender that matches their grammatical gender: it allows for an explanation where there is no such match, either directly perceived, or induced through literary devices.

The present research has both theoretical and practical benefits. From a theoretical point of view, it contributes to research on phonology and gender, in particular in going a step further in understating the link between the two. From a practical perspective, such a connection between gender and sounds could be exploited in advertising, in particular in product naming, to build names that fit a product, and which are appealing to the desired customers. Studies have shown that especially in the absence of meaning, the form of a word can be used to generate specific associations and stimulate the imagination of prospective customers (Sells and Gonzales, 2003; Bedgley, 2002; Botton et al., 2002).

2 Gender

What is the origin of grammatical gender and how does it relate to natural gender? Opinions are split. Historically, there were two main, opposite, views: (i) there is a semantic explanation, and natural gender motivated the category (ii) the relationship between natural and grammatical gender is arbitrary.

Grimm (1890) considered that grammatical gender is an extension of natural gender brought on by imagination. Each gender is associated with particular adjectives or other attributes, and in some cases (such as for *sun* and *moon*) the assignment of gender is based on personification. Brugmann (1889) and Bloomfield (1933) took the position that the mapping of nouns into genders is arbitrary, and other phenomena – such as derivations, personification – are secondary to the established agreement. Support for this second view comes also from language acquisition: children who learn a gendered language do not have a natural gender attribute that they try to match onto the newly acquired words, but learn these in a separate process. Any match or mapping between natural and grammatical gender is done after the natural gender "feature" is acquired itself. Kilarski (2007) presents a more detailed overview of currents and ideas about the origin of gender. Unterbeck (1999) contains a collection of papers that investigate grammatical gender in several languages, aspects of gender acquisition and its relation with grammatical number and agreement.

There may be several reasons for the polemic between these two sides. One may come from the categorization process, the other from the relation between word form and its meaning. Let us take them each in turn, and see how they influenced gender.

Grammatical gender separates the nouns in a language into disjoint classes. As such, it is a categorization process. The traditional – classical – theory of categorization and concepts viewed categories and concepts as defined in terms of a set of common properties that all its members should share. Recent theories of concepts have changed, and view concepts (and categories) not necessarily as "monolithic" and defined through rules, but rather as clusters of members that may resemble each other along different dimensions (Margolis and Laurence, 1999).

In most linguistic circles, the principle of arbitrariness of the association between form and meaning, formalized by de Saussure (1916) has been largely taken for granted. It seems however, that it is hard to accept such an arbitrary relation, as there have always been contestants of this principle, some more categorical than others (Jakobson, 1937; Jespersen, 1922; Firth, 1951). It is possible that the correlation we perceive between the word form and the meaning is something that has arisen after the word was coined in a language, being the result of what Firth called "phonetic habit" through "an attunement of the nervous system", and that we have come to prefer, or select, certain word forms as more appropriate to the concept they name – "There is no denying that there are words which we feel instinctively to be adequate to express the ideas they stand for. ... Sound symbolism, we may say, makes some words more fit to survive" (Jespersen, 1922).

These two principles relate to the discussion on gender in the following manner: First of all, the categories determined by grammatical gender need not be homogeneous, and their members need not all respect the same membership criterion. This frees us from imposing a matching between natural and grammatical gender where no such relation is obvious or possible through literary devices (personification, metaphor, metonymy). Nouns belonging to the same gender category may resemble each other because of semantic considerations, lexical derivations, internal structure, perceived associations and so on. Second, the fact that we allow for the possibility that the surface form of a word may encode certain word characteristics or attributes, allows us to hypothesize that there is a surface, phonological, similarity between words grouped within the same gender category, that can supplement other resemblance criteria in the gender category (Zubin and Köpcke, 1986).

Zubin and Köpcke (1981), Zubin and Köpcke (1986) have studied the relation between semantic characteristics and word form with respect to gender for German nouns. Their study was motivated by two observations: Zipf (1935) showed that word length is inversely correlated with frequency of usage, and Brown (1958) proposed that in choosing a name for a given object we are more likely to use a term corresponding to a "basic" level concept. For example, *chair, dog, apple* would correspond to the basic level, while *furniture, animal, fruit* and *recliner, collie, braeburn apple* correspond to a more general or a more specific level, respectively. Their study of gender relative to these levels have shown that basic level terms have masculine, feminine, and rarely neuter genders, while the more undifferentiated categories at the superordinate level are almost exclusively neuter.

In psycholinguistic research, Friederici and Jacobsen (1999) adopt the position that a lexical entry consists of two levels: form and semantic and grammatical properties to study the influence of gender priming – both from a form and semantic perspective – on language comprehension. Vigliocco et al. (2004a) study gender priming for German word production. While this research studies the influence of the word form on the production of nouns with the same or different grammatical gender, there is no study of the relation between word forms and their corresponding gender.

In recent studies we have found on the relation between word form and its associated gender, the only phonological component of a word that is considered indicative is the ending. Spalek et al. (2008) experiment on French nouns, and test whether a noun's ending is a strong clue for gender for native speakers of French. Vigliocco et al. (2004b) test cognitive aspects of grammatical gender of Italian nouns referring to animals.

Cucerzan and Yarowsky (2003) present a bootstrapping process to predict gender for nouns in context. They show that context gives accurate clues to gender (in particular through determiners, quantifiers, adjectives), but when the context is not useful, the algorithm can fall back successfully on the word form. Cucerzan and Yarowsky model the word form for predicting gender using suffix trie models. When a new word is encountered, the word is mapped onto the trie starting from the last letter, and it is assigned the gender that has the highest probability based on the path it matches in the trie. In context nouns appear with various inflections – for number and case in particular. Such morphological derivations are gender specific, and as such are strong indicators for gender.

The hypothesis tested here is that gender comes from the general sound of the language, and is distributed throughout the word. For this, the data used should not contain nouns with "tell tale" inflections. The data will therefore consist of nouns in the singular form, nominative case. Some nouns are derived from verbs, adverbs or adjectives, or other nouns through morphological derivations. These derivations are regular and are identifiable through a rather small number of regular suffixes. These suffixes (when they are indicative of gender) and word endings will be used as baselines to compare the accuracy of prediction on the full word with the ending fragment.

3 Data

We test our gender-language sounds connection through two languages from different language families. German will be the representative of the Germanic languages, and Romanian for the Romance ones. We first collect data in the two languages, and then represent them through various features – letters, pronunciation, phonetic features.

3.1 Noun collections

German data For German we collect nouns and their grammatical gender from a German-English dictionary, part of the BEOLINGUS multi-lingual dictionary[1]. In the first step we collected the German nouns and their gender from this dictionary. In step 2, we filter out compounds. The reason for this step is that a German noun compound will

[1] `http://dict.tu-chemnitz.de/`

have the gender of its head, regardless of its nominal modifiers. For the lack of a freely available tool to detect and split noun compounds, we resort to the following algorithm:

```
1. initialize the list of nouns $L_N$ to the empty list;
2. take each noun $n$ in the dictionary $D$, and
    (a) if $\exists n_i \in L_N$ such that $n$ is an end substring of $n_i$, then add $n$ to $L_N$ and remove $n_i$ from $L_N$;
    (b) if $\exists n_i \in L_N$ such that $n_i$ is a end substring of $n$, skip $n$;
```

Essentially, we remove from the data all nouns that include another noun as the end part (which is the head position in German noun compounds). This does not filter examples that have suffixes added to form the feminine version of a masculine noun, for example: *(der) Lehrer – (die) Lehrerin* (teacher). The suffixes are used in one of the baselines for comparison with our learning method.

We obtain noun pronunciation information from the Bavarian Archive for Speech Signals[2]. We filter again our list L_N to keep nouns for which we have pronunciation information. This allows us to compare the learning results when letter or pronunciation information is used.

After collecting the nouns and their pronunciation, we map the pronunciation onto lower level phonetic features, following the IPA encoding of sounds for the German language. The mapping between sounds and IPA features was manually encoded following IPA tables.

Romanian data We extract singular nominative forms of nouns from the Romanian lexical database (Barbu, 2008). The resource contains the proper word spelling, including diacritics and special characters. Because of this and the fact that there is a straightforward mapping between spelling and pronunciation in Romanian, we can use the entire data extracted from the dictionary in our experiments, without special pronunciation dictionaries. Following the example for the German language, we encode each sound through lower level phonological features using IPA guidelines.

As in Italian, in Romanian there are strong phonological cues for nouns, especially those having the feminine gender: they end in *ă* and *e*. To determine whether the connection between a word form and gender goes beyond this superficial rule, we generate a dataset in which the nouns are stripped of their final letter, and their representation is built based on this reduced form.

Table 1 shows the data collected and the distribution in the three classes.

	German		Romanian	
masc.	565	32.64%	7338	15.14%
fem.	665	38.42%	27187	56.08%
neut.	501	28.94%	13952	28.78%
total	1731		48477	

Table 1: Data statistics

Because for Romanian the dataset is rather large, we can afford to perform undersampling to balance our classes, and have a more straightforward evaluation. We generate a perfectly balanced dataset by undersampling the feminine and the neuter classes down to the level of the masculine class. We work then with a dataset of 22014 instances, equally distributed among the three genders.

3.2 Data representation

For each word in our collections we produce three types of representation: letters, phonemes and phonological features. Table 2 shows examples for each of these representations. The letter and phoneme representations are self-explanatory. We obtain the pronunciation corresponding to each word from a pronunciation dictionary, as mentioned in Section 3.1, which maps a word onto a sequence of phonemes (phones). For Romanian we have no such resource, but me make without since in most part the pronunciation matches the letter representation[3].

German		
letter	abend (m)	a b e n d
phoneme		a: b @ n d

Romanian		
letter	seară (f)	s e a r ă

Table 2: Data representation in terms of letters and phonemes for the German and Romanian forms of the word *evening*. For Romanian, the letter and phoneme representation is the same.

[2] http://www.phonetik.uni-muenchen.de/Bas/

[3] The exceptions are the diphthongs and a few groups of letters: ce, ci, che, chi, oa, and the letter x.

Phonemes, the building blocks of the phonetic representation, can be further described in terms of phonological features – “configurations” of the vocal tract (e.g tongue and lips position), and acoustic characteristics (e.g. manner of air flow). We use IPA standards for mapping phones in German and Romanian onto these phonological features. We manually construct a map between phones and features, and then automatically binarize this representation and use it to generate a representation for each phone in each word in the data. For the word *abend (de) / seara (ro) (evening)* in Figure 2, the phonological feature representation for German is:

0000100000001000010000000001
0001000100000000000010000000
0000100000000100000000010001
1000000100000010000000000000
1000000100000000000010000000,

with the feature base:

$<$ *alveolar, approximant, back, bilabial, central, close, closemid, consonant, fricative, front, glottal, labiodental, long, mid, nasal, nearclose, nearopen, open, openmid, palatal, plosive, postalveolar, rounded, short, unrounded, uvular, velar, vowel* $>$.

For Romanian, the phonological feature base is:

$<$ *accented, affricate, approximant, back, bilabial, central, close, consonant, dental, fricative, front, glottal, labiodental, mid, nasal, open, plosive, postalveolar, rounded, trill, unrounded, velar, voiced, voiceless, vowel* $>$,

and the phonological feature representation of the word changes accordingly.

4 Kernel Methods and String Kernels

Our hypothesis that the gender is in the name is equivalent to proposing that there are sequences of letters/sounds/phonological features that are more common among nouns that share the same gender or that can distinguish between nouns under different genders. To determine whether that is the case, we use a string kernel, which for a given string (sequence) generates a representation that consists of all its substrings of length less than a parameter l. The words are represented as strings with boundaries marked with a special character (’#’). The high dimensional representation generated by the string kernel is used to find a hyperplane that separates instances of different classes. In this section we present in detail the kernel we use.

Kernel-based learning algorithms work by embedding the data into a feature space (a Hilbert space), and searching for linear relations in that space. The embedding is performed implicitly, that is by specifying the inner product between each pair of points rather than by giving their coordinates explicitly.

Given an input set $\mathcal{X}$ (the space of examples), and an embedding vector space $\mathcal{F}$ (feature space), let $\phi : \mathcal{X} \rightarrow \mathcal{F}$ be an embedding map called feature map.

A *kernel* is a function k, such that for all $x, z \in \mathcal{X}$, $k(x,z) =< \phi(x), \phi(z) >$, where $< .,. >$ denotes the inner product in $\mathcal{F}$.

In the case of binary classification problems, kernel-based learning algorithms look for a discriminant function, a function that assigns $+1$ to examples belonging to one class and -1 to examples belonging to the other class. This function will be a linear function in the space $\mathcal{F}$, that means it will have the form:

$$f(x) = \text{sign}(< w, \phi(x) > +b),$$

for some weight vector w. The kernel can be exploited whenever the weight vector can be expressed as a linear combination of the training points, $\sum\limits_{i=1}^{n} \alpha_i \phi(x_i)$, implying that f can be expressed as follows:

$$f(x) = \text{sign}(\sum_{i=1}^{n} \alpha_i k(x_i, x) + b)$$

.

Various kernel methods differ in the way in which they find the vector w (or equivalently the vector α). Support Vector Machines (SVM) try to find the vector w that define the hyperplane that maximum separate the images in $\mathcal{F}$ of the training examples belonging to the two classes. Mathematically SVMs choose the w and b that satisfy the following optimization criterion:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} [1 - y_i(< w, \phi(x_i) > +b)]_+ + \nu ||w||^2$$

where y_i is the label ($+1/-1$) of the training example x_i, ν a regularization parameter and $[x]_+ = \max(x, 0)$.

Kernel Ridge Regression (KRR) selects the vector w that simultaneously has small empirical error and small norm in Reproducing Kernel Hilbert Space generated by kernel k. The resulting minimization problem is:

$$\min_w \frac{1}{n} \sum_{i=1}^{n} (y_i - < w, \phi(x_i) >)^2 + \lambda ||w||^2$$

where again y_i is the label ($+1/-1$) of the training example x_i, and λ a regularization parameter. Details about SVM and KRR can be found in (Taylor and Cristianini, 2004). What is important is that above optimization problems are solved in such a way that the coordinates of the embedded points are not needed, only their pairwise inner products which in turn are given by the kernel function k.

SVM and KRR produce binary classifiers and gender classification is a multi-class classification problem. There are a lot of approaches for combining binary classifiers to solve multi-class problems. We used *one-vs-all* scheme. For arguments in favor of one-vs-all see (Rifkin and Klautau, 2004).

The kernel function offers to the kernel methods the power to naturally handle input data that are not in the form of numerical vectors, for example strings. The kernel function captures the intuitive notion of similarity between objects in a specific domain and can be any function defined on the respective domain that is symmetric and positive definite. For strings, a lot of such kernel functions exist with many applications in computational biology and computational linguistics (Taylor and Cristianini, 2004).

Perhaps one of the most natural ways to measure the similarity of two strings is to count how many substrings of length p the two strings have in common. This give rise to the p-spectrum kernel. Formally, for two strings over an alphabet Σ, $s, t \in \Sigma^*$, the p-spectrum kernel is defined as:

$$k_p(s,t) = \sum_{v \in \Sigma^p} \text{num}_v(s)\text{num}_v(t)$$

where $\text{num}_v(s)$ is the number of occurrences of string v as a substring in s [4] The feature map defined by this kernel associate to each string a vector of dimension $|\Sigma|^p$ containing the histogram of frequencies of all its substrings of length p. Taking into account all substrings of length less than p it will be obtained a kernel that is called the *blended spectrum kernel*:

$$k_1^p(s,t) = \sum_{q=1}^{p} k_q(s,t)$$

The blended spectrum kernel will be the kernel that we will use in conjunction with SVM and KRR. More precisely we will use a normalized version of the kernel to allow a fair comparison of strings of different length:

$$\hat{k}_1^p(s,t) = \frac{k_1^p(s,t)}{\sqrt{k_1^p(s,s)k_1^p(t,t)}}$$

5 Experiments and Results

We performed 10-fold cross-validation learning experiments with kernel ridge regression and the string kernel (KRR-SK) presented in Section 4. We used several baselines to compare the results of the experiments against:

BL-R Gender is assigned following the distribution of genders in the data.

BL-M Gender is assigned following the majority class (only for German, for Romanian we use balanced data).

BL-S Gender is assigned based on suffix-gender relation found in the literature. We use the following mappings:

- German (Schumann, 2006):

  feminine *-ade, -age, -anz, -e, -ei, -enz, -ette, -heit, -keit, -ik, -in, -ine, -ion, -itis, -ive, -schaft, -tät, -tur, -ung, -ur*;

  masculine *-ant, -er, -ich, -ismus, -ling*;

  neuter *-chen, -ist, -lein, -ment, -nis, -o, -tel, -um*.

  In our data set the most dominant gender is feminine, therefore we assign this gender to all nouns that do not match any of the previous suffixes. Table 4 shows a few suffixes for each gender, and an example noun.

- Romanian: in Romanian the word ending is a strong clue for gender, especially for feminine nouns: the vast majority end in either *-e* or *-ă* (Doca, 2000). We design a heuristic that assigns the gender “preferred” by the last letter – the

[4] Note that the notion of substring requires contiguity. See (Taylor and Cristianini, 2004) for discussion about the ambiguity between the terms ”substring” and ”subsequence” across different traditions: biology, computer science.

Method	Accuracy	masc. F-score	fem. F-score	neut. F-score
German				
BL-R	33.79			
BL-M	38.42			
BL-S	51.35	40.83	62.42	26.69
KRR-SK	72.36 ± 3	64.88 ± 5	84.34 ± 4	64.44 ± 7
KRR-SK$_{noWB}$	66.91	58.77	79.19	58.26
Romanian				
BL-R	33.3			
BL-S	74.38	60.65	97.96	63.93
KRR-SK	78.83 ± 0.8	68.74 ± 0.9	98.05 ± 0.2	69.38 ± 2
KRR-SK no last letter	65.73 ± 0.6	56.11 ± 1	85.00 ± 0.5	55.05 ± 1
KRR-SK$_{noWB}$	77.36	67.54	96.75	67.39

Table 3: 10-fold cross-validation results – accuracy and f-scores percentages (± variation over the 10 runs) – for gender learning using string kernels

German

gender	suffix	example	
fem.	-e	Ecke	(corner)
	-heit	Freiheit	(freedom)
	-ie	Komödie	(comedy)
masc.	-er	Fahrer	(driver)
	-ich	Rettich	(radish)
	-ling	Frühling	(spring - season)
neut.	-chen	Mädchen	(girl)
	-nis	Verständnis	(understanding)
	-o	Auto (car)	

Table 4: Gender assigning rules and examples for German

Romanian

gender	ending	example		Prec.
fem.	-ă	masă	(table)	98.04
	-e	pâine	(bread)	97.89
masc.	-g	sociolog	(sociologist)	72.77
	-r	nor	(cloud)	66.89
	-n	domn	(gentleman)	58.45
neut.	-m	algoritm	(algorithm)	90.95
	-s	vers	(verse)	66.97
	-t	eveniment	(event)	51.02

Table 5: Word-ending precision on classifying gender and examples for Romanian

majority gender of all nouns ending in the respective letter – based on analysis of our data. In Table 5 we include some of the letter endings with an example noun, and a percentage that shows the precision of the ending in classifying the noun in the gender indicated in the table.

The results of our experiments are presented in Table 3, in terms of overall accuracy, and f-score for each gender. The performance presented corresponds to the letter-based representation of words. It is interesting to note that this representation performed overall better than the phoneme or phonological feature-based ones. An explanation may be that in both the languages we considered, there is an (almost) one-to-one mapping between letters and their pronunciation, making thus the pronunciation-based representation unnecessary. As such, the letter level captures the interesting commonalities, without the need to go down to the phoneme-level.

We performed experiments for Romanian when the last letter of the word is removed. The reason for this batch of experiments is to further test the hypothesis that gender is more deeply encoded in a word form than just the word ending. For both languages we observe statistically significant higher performance than all baselines. For Romanian, the last letter heuristic gives a very high baseline, confirming that Romanian has strong phonological cues for gender in the ending. Had the word ending been the only clue to the word's gender,

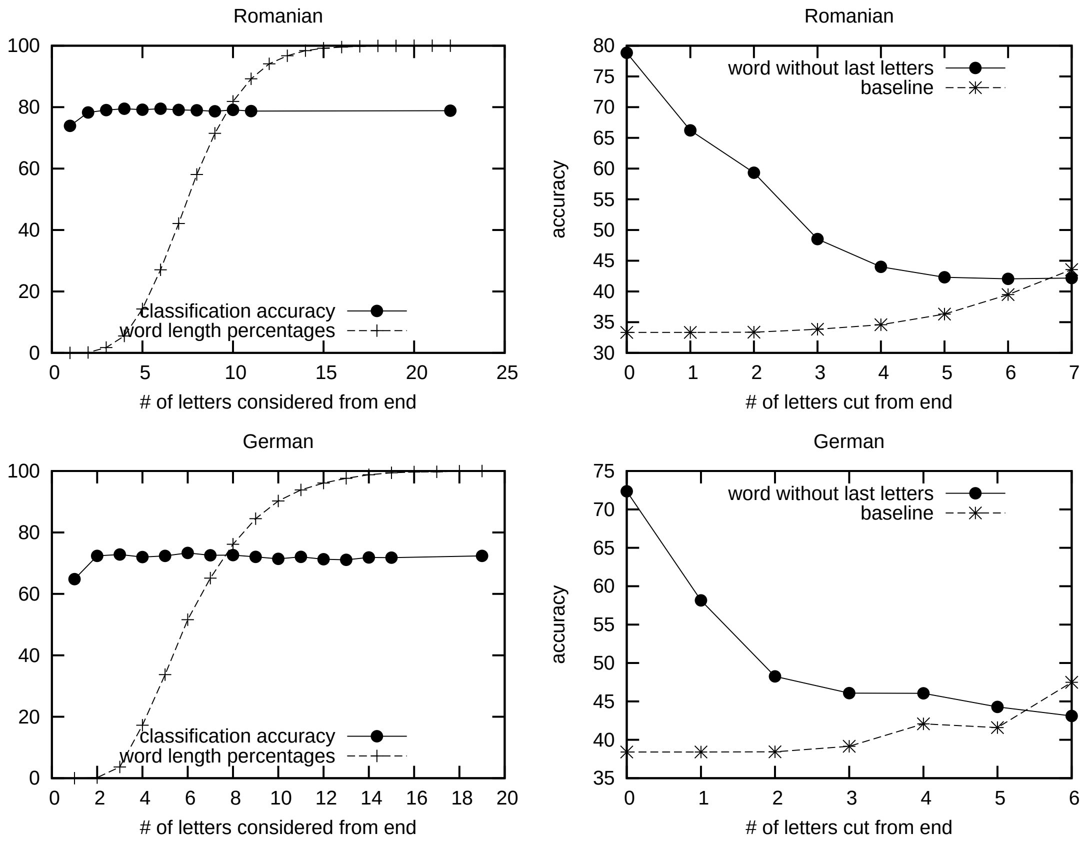

Figure 1: Gender prediction based on the last N letters, and based on the word minus the last N letters

once it is removed the performance on recognizing gender should be close to the random assignment. This is not the case, and the improvement over the random baseline is 32% points. It is interesting to notice that when cutting off the last letter the class for which the gender assignment heuristic was clearest – the feminine class with -ă and -e endings – the performance remains very high – 85% F-score.

To further test where the gender indicators are located, we performed two more sets of experiments: (i) classify words in their corresponding gender class using the word minus the last N letters; (ii) classify words based on the last N letters. The results of these experiments in terms of accuracy are presented in Figure 1. When considering only the last N letters the performance is high for both German and Romanian, as expected if the gender indicators are concentrated at the end of the word. It is interesting though to notice the results of classification based on the word without the last N letters. The prediction accuracy monotonically decreases, but remains above the baseline until more than 6 letters are cut. Because as letters are cut some words completely disappear, the baseline changes accordingly. 94.07% of the words have a length of at most 12 letters in the Romanian dataset, and 96.07% in the German one. Because gender prediction can be done with accuracy higher than the random baseline even after 6 letters are cut from the ending of the word indicate that for more than 94% of the words considered, gender clues are spread over more than the second half of the word. Again, we remind the reader that the word forms are in nominative case, with no case or number inflections (which are strong indicators of gender in both Romanian and German).

Except for lines $KRR - SK_{noWB}$, the results in Table 3 are obtained through experiments conducted on words containing word boundary markers, as indicated in Section 4. Because of these markers, word starting or word ending substrings are distinct from all the others, and information about their position in the original word is thus preserved. To further explore the idea that gender indicators are not located only in word endings, we ran classification experiments for German and Romanian when the word representation does not contain word boundary markers. This means that the substrings generated by the string kernel have no position information. The results of these experiments are presented in rows $KRR-SK_{noWB}$ in Table 3. The accuracy is slightly lower than the best results obtained when word boundaries are marked and the entire word form is used. However, they are well above all the baselines considered, without no information about word endings.

For both German and Romanian, the gender that was learned best was feminine. For German part of this effect is due to the fact that the feminine class is more numerous in the data. For Romanian the data was perfectly balanced, so there is no such bias. Neuter and masculine nouns have lower learning performance. For Romanian, a contribution to this effect is the fact that neuter nouns behave as masculine nouns in their singular form (take the same articles, inflections, derivations), but as feminine in the plural, and our data consists of nouns in singular form. It would seem that from an orthographic point of view, neuter and masculine nouns are closer to each other than to feminine nouns.

From the reviewed related work, the one that uses the word form to determine gender is Cucerzan and Yarowsky (2003) for Romanian. There are two important differences with respect to the approach presented here. First, they consider words in context, which are inflected for number and case. Number and case inflections are reflected in suffixes that are gender specific. The words considered here are in singular form, nominative case – as such, with no inflections. Second, Cucerzan and Yarowsky consider two classes: feminine vs. masculine and neuter. Masculine and neuter nouns are harder to distinguish, as in singular form neuter nouns behave like masculine nouns in Romanian. While the datasets and word forms used by Cucerzan and Yarowsky are different than the one used here, the reader may be curious how well the word form distinguishes between feminine and the other two classes in the experimental set-up used here. On the full[5] Romanian dataset described in Section 3, a two class classification gives 99.17% accuracy. When predicting gender for all words in their dataset, Cucerzan and Yarowsky obtain 98.25% accuracy.

6 Conclusion

When a speaker of a genderless language tries to learn a language with grammatical gender, it is very tempting to try to assign grammatical gender based on perceived or guessed natural gender types. This does not work out well, and it only serves to confuse the learner even more, when he finds out that nouns expressing concepts with clear feminine or masculine natural gender will have the opposite or a neutral grammatical gender, or that one concept can be referred to through names that have different grammatical genders. Going with the flow of the language seems to be a better idea, and allow the sound of a word to dictate the gender.

In this paper we have investigated the hypothesis that gender is encoded in the word form, and this encoding is more than just the word endings as it is commonly believed. The results obtained show that gender assignment based on word form analysis can be done with high accuracy – 72.36% for German, and 78.83% for Romanian. Existing gender assignment rules based on word endings have lower accuracy. We have further strengthened the point by conducting experiments on Romanian nouns without tell-tale word endings. The accuracy remains high, with remarkably high performance in terms of F-score for the feminine class (85%). This leads us to believe that gender information is somehow redundantly coded in a word. We plan to look closer at cases where we obtain different predictions based on the word ending and the full form of the word, and use boosting to learn weights for classifiers based on different parts of the word to see whether we can further improve the results.

As we have underlined before, word form similarity between words under the same gender is one criterion for gender assignment. It would be interesting to verify whether gender recognition can be boosted by using lexical resources that capture the semantics of the words, such as WordNets or knowledge extracted from Wikipedia, and verify whether similarities from a semantic point of view are also responsible for gender assignments in various languages.

References

Ana-Maria Barbu. 2008. Romanian lexical data bases: Inflected and syllabic forms dictionaries. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. http://www.lrec-conf.org/proceedings/lrec2008/.

Sharon Bedgley. 2002. Strawberry is no blackberry: Building brands us-

[5] By "full" we mean the dataset before balancing the classes 48,477 instances (see Table 1).

ing sound. http://online.wsj.com/article/0,,SB1030310730179474675.djm,00.html.

Leonard Bloomfield. 1933. *Language*. Holt, Reinhart & Winston, New York.

Marcel Botton, Jean-Jack Cegarra, and Beatrice Ferrari. 2002. *Il nome della marca: creazione e strategia di naming, 3rd edition*. Guerini e Associati.

Roger Brown. 1958. *Words and Things*. The Free Press, New York.

Karl Brugmann. 1889. Das Nominalgeschlecht in den indogermanischen Sprachen. In *Internationale Zeitschrift für allgemenine Sprachwissenschaft*, pages 100–109.

S. Cucerzan and D. Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proceedings of HLT-NAACL 2003*, pages 40–47.

Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Harrassowitz, Wiesbaden.

Gheorghe Doca Doca. 2000. *Romanian language. Vol. II: Morpho-Syntactic and Lexical Structures*. Ars Docendi, Bucharest, Romania.

John Rupert Firth. 1951. Modes and meaning. In *Papers in linguistics 1934-1951*. Oxford University Press, London.

Angela Friederici and Thomas Jacobsen. 1999. Processing grammatical gender during language comprehension. *Journal of Psychological Research*, 28(5):467–484.

Jacob Grimm. 1890. *Deutsche Grammatik*.

Roman Jakobson. 1937. *Lectures on Sound and Meaning*. MIT Press, Cambridge, MA.

Otto Jespersen. 1922. *Language - its Nature, Development and Origin*. George Allen & Unwim Ltd., London.

Marcin Kilarski. 2007. On grammatical gender as an arbitrary and redundant category. In Douglas Kilbee, editor, *History of Linguistics 2005: Selected papers from the 10th International Conference on the History of Language Sciences (ICHOLS X)*, pages 24–36. John Benjamins, Amsterdam.

Eric Margolis and Stephen Laurence, editors. 1999. *Concepts: Core Readings*. MIT Press.

Ryan Rifkin and Aldebaro Klautau. 2004. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5(January):101–141.

Johannes Schumann. 2006. *Mittelstufe Deutsch*. Max Hueber Verlag.

Peter Sells and Sierra Gonzales. 2003. The language of advertising. http://www.stanford.edu/class/linguist34/; in particular unit 8: ˜/Unit_08/blackberry.htm.

Katharina Spalek, Julie Franck, Herbert Schriefers, and Ulrich Frauenfelder. 2008. Phonological regularities and grammatical gender retrieval in spoken word recognition and word production. *Journal of Psycholinguistic Research*, 37(6):419–442.

John S. Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

Barbara Unterbeck, editor. 1999. *Gender in Grammar and Cognition. Approaches to Gender*. Trends in Linguistics. Studies and Monographs. 124. Mouton de Gruyter.

Gabriela Vigliocco, David Vinson, Peter Indefrey, Willem Levelt, and Frauke Hellwig. 2004a. Role of grammatical gender and semantics in german word production. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30(2):483–497.

Gabriela Vigliocco, David Vinson, and Federica Paganelli. 2004b. Grammatical gender and meaning. In *Proc. of the 26th Meeting of the Cognitive Science Society*.

George Zipf. 1935. *The Psychobiology of Language*. Addison-Wesley.

David Zubin and Klaus-Michael Köpcke. 1981. Gender: A less than arbitrary grammatical category. In R. Hendrick, C. Masek, and M. F. Miller, editors, *Papers from the seventh regional meeting*, pages 439–449. Chicago Linguistic Society, Chicago.

David Zubin and Klaus-Michael Köpcke. 1986. Gender and folk taxonomy: The indexical relation between grammatical and lexical categorization. In C. Craig, editor, *Noun classes and categorization*, pages 139–180. Benjamins, Philadelphia.