Minimally Supervised Induction of Grammatical Gender

Silviu Cucerzan and David Yarowsky
Department of Computer Science and
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218, USA
{silviu,yarowsky}@cs.jhu.edu

Abstract

This paper investigates the problem of determining grammatical gender for the nouns of a language starting with minimal resources: a very small list of seed nouns for which gender is known or via translingual projection of natural gender. We show that through a bootstrapping process that uses contextual clues from an unannotated corpus and morphological clues modeled with suffix tries, accurate gender predictions can be induced for five diverse test languages.

1 Introduction

Grammatical gender is a common but not universal property of languages. English, the most studied language from a computational point of view, does not have grammatical gender. When grammatical gender is utilized in a language it can take various forms, as we show briefly in this introduction.

Merriam-Webster dictionary defines gender as being "a subclass within a grammatical class (as noun, pronoun, adjective, or verb) of a language that is partly arbitrary but also partly based on distinguishable characteristics (as shape, social rank, manner of existence, or sex) and that determines agreement with and selection of other words or grammatical forms". Although many grammatical classes in a language can have the gender subclass, grammatical gender is an intrinsic property of nouns, and the words in other classes use it to show agreement with the modified noun.

Because the grammatical gender (henceforth "gender") of a noun affects the possible inflections of the noun and the immediate context in which the noun occurs, gender information can be used as an important component in NLP tasks, such as text generation and machine translation, or as a secondary evidence source for part-of-speech (POS) tagging (to disambiguate the main POS tag of a word in a given context) and parsing (to determine the syntactic dependencies).

The overwhelming majority of nouns have only one grammatical gender; therefore, once gender is known for an exhaustive list of nouns in the language, very little work has to be done in terms of gender discovery in that language. Unfortunately, there is a shortage of gender information for many languages. We propose a general system that can serve as a basis for building more refined, language-dependent gender predictors, and we evaluate it on several languages for which gender information in electronic format was obtainable for large lists of nouns in those languages.

The current state-of-the-art POS taggers for inflected languages (such as Hajič and Hladká (1998)) discover grammatical gender at the time of annotating the text. Cucerzan and Yarowsky (2002) propose a model in which they postpone the gender resolution, keeping ambiguous gender sub-tags which are later disambiguated based on global agreement statistics in the tagged text.

We show in this paper that gender can be determined reliably independent of any other task, using minimal supervision. While the system obtains good results on a variety of Indo-European languages, it is difficult to claim full language independence because the gender system is culture specific and does not transfer well from one language to another (Foundalis, 2002).

Most Indo-European languages have a gender system with two or three classes. The typical genders for Indo-European languages are feminine, masculine and neuter. Latin has all three, but many of its descendants, such as French, Spanish, and Portuguese only kept two of these, the neuter gender being lost. In most Northern Germanic languages, feminine and masculine have merged into a common gender (such as in Swedish and Danish). In contrast, other languages such as Czech make a further distinction between masculine nouns by dividing them into

animate and inanimate. African languages such as the Bantu languages have a higher number of noun classes that can be interpreted as gender when the term is used in a more general sense. Even if a language has no concept of grammatical gender (for example, English and Hungarian), personal pronouns often have different forms that reflect the sex of the referenced noun.

In languages with masculine and feminine gender classes, the grammatical gender of nouns describing animate beings is related many times to their natural gender (sex). For example, in Romanian, *fiu* (meaning *son*) is masculine and *fiică* (*daughter*) is feminine. Even in the case of nouns having a natural gender, several exceptions exist: for example, in German, the noun *Mädchen* (girl) is neuter and not feminine as one might expect. Instead of being consistent with the natural gender, this noun's grammatical gender is consistent with a morphological rule, that states that words ending in *-lein* or *-chen* are neuter.

Many studies looking at the way people learn and remember gender have been carried out over the years, both for native speakers (L1) or second language (L2) learners, especially with English backgrounds, e.g. (Tucker et al., 1977), (Andersen, 1984), (Carroll, 1989), (Pérez-Pereira, 1991), (Surridge, 1993), (Taraban and Kempe, 1999). All these studies show that learning grammatical gender is seen as a complex process by the human subjects. There have been several studies in the connectionist literature on the fully supervised learning of gender, focusing on suffixes in French, e.g. (Sokolik and Smith, 1992) and (Rodrigues and Boivin, 2000).

Several computational problems can be formulated, from discovering the existence of gender in a language to building complex systems that can accurately predict gender for any noun in a given language. We show that it is possible to build an accurate gender-prediction system starting only from a raw corpus and a small gender-annotated noun list.

2 System Description

As observed in Section 1, grammatical gender is a rather complex property of nouns, based on several semantic and morphological phenomena and any attempt to build a unified gender model for many languages should have great flexibility. We formulate the problem of determining the grammatical gender as a classification problem and develop a unified language-independent learning technique. While it is very hard to account directly for the semantic properties of nouns when developing tools for a new language for which no annotated texts or lexicons are available, one can use context extracted
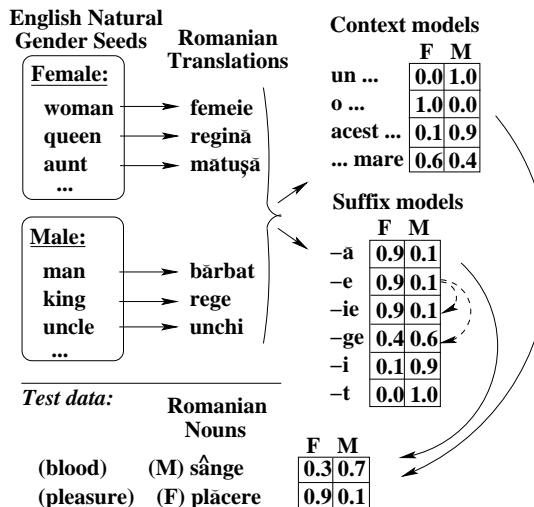

Figure 1: An illustration of the translingual projection of natural gender seeds via contextual and suffix models

from large unannotated corpora to compensate for the lack of this type of knowledge, by exploiting the gender agreement of nouns and immediate context.

We propose a two-component bootstrapping technique that can be summarized as follows:

Starting with a list of seed nouns, reliable contexts for every gender class are determined and more nouns that are seen with those contexts are extracted from the corpus. The morphological properties of the seeds and the nouns with relevant context are then used to hypothesize grammatical gender for all the nouns in the language.

This method assumes the existence of a list of nouns for the target language. This is not a major restriction; Cucerzan and Yarowsky (2002) show how nouns can be hypothesized reliably using a foreign language to English dictionary.

An in-depth analysis of the algorithm, problems encountered, and results, with examples in Romanian, is presented in the following subsections of the paper. Results for several other languages are summarized in Section 3.

2.1 Seeding

One of the most important issues in any bootstrapping process is the quantity and quality of the seeds. In a minimal-effort framework, one would like to use as few seeds as possible and have them be easily extracted from widely available resources.

A first approach uses the translingual projection of natural gender, starting with the translations of several English nouns that have a natural gender (Figure 1). We considered a total of 30 feminine and 30 masculine nouns in terms of natural gender[1], as

---

[1]Nouns such as *soldier* and *witch* are assumed to cor-

shown in Table 1.

When a bilingual dictionary with English is available electronically (as was the case for Romanian, French, and Spanish), the projection of natural gender labels is done automatically. When such a dictionary is not available in an electronic format, an English-to-target-language hard-copy dictionary can be used (this was done for Slovene, with 35 minutes necessary to extract all the translations of the English seeds). This first approach is limited to the languages in which grammatical gender is related to natural gender. Also, the results may vary according to the usage of the selected natural gender nouns in the corpus (for example, words like *niece* or *rooster* may not be very frequent in newswire corpora).

A second, more general approach is to extract from the corpus a list of nouns selected on the basis of frequency, number of contexts with which they co-occur, and suffix patterns, and to label them with gender information by dictionary look-up. This approach guarantees that the seeds used are representative for the given corpus, so that the bootstrapping has a better chance to succeed. On the other hand, it presumes the existence of a gender annotated mono- or bi-lingual dictionary (which might be more difficult to find than an unannotated bilingual dictionary). Our goal was to select a small set (50 or fewer[2]) of such words in order to minimize the human supervision required.

The system was initially developed for Romanian. A list of 17,053 nouns were extracted from the MULTEXT-East Romanian lexicon (Erjavec et al., 1997). Unfortunately, all neuter nouns are listed as masculine in this lexicon, probably because morphologically, neuters and masculine are alike in the singular – for example, *copac* (*tree*) is masculine while *conac* (*mansion*) is neuter. Syntactically, neuter singular nouns occur in the same context as masculine singular nouns, so the gender agreement cannot be used to choose between the two class labels.

Only 2,132 of the 17,053 known nouns appear in the corpus and we focus our attention on these nouns, as a relevant sample of nouns in the language.

Starting from the English natural-gender words, seed nouns were obtained by projection using a scanned and OCRed bilingual Romanian-English dictionary. One major class of seed candidates were automatically and systematically eliminated, namely the multi-gender case where a male and female pair share the same translation (e.g. the masculine noun

---

respond to a certain gender for historical reasons, and this assumption is validated empirically.

[2]For some corpora, the system may not be able to extract 50 seeds because of the frequency constraints imposed.

Feminine	Freq	R/F/E/S	Masculine	Freq	R/F/E/S
woman	322	+/+/+/+	man	1396	+/+/+/+
girl	234	±̲/+/±̲/+	boy	261	±̲/+/±̲/+
sister	56	+/+/+/+	brother	106	+/+/+/+
mother	268	+/+/+/+	father	246	+/+/+/+
wife	302	+/+/+/+	husband	184	+/+/+/+
daughter	93	±̲/+/+/+	son	191	±̲/+/+/+
daughter-in-law	1	+/+/+/*	son-in-law	5	+/+/+/*
stepdaugther	1	?/?/+/+	stepson	3	?/?/+/+
grandmother	14	?/+/+/+	grandfather	17	?/+/+/*
granddaughter	3	+/+/+/+	grandson	7	+/+/+/+
aunt	11	+/+/?/+	uncle	26	+/+/+/+
niece	9	+/+/+/+	nephew	11	+/+/+/+
bride	39	?/+/+/+	groom	5	?/?/+/+
girlfriend	5	?/?/±̲/?	boyfriend	1	+/?/±̲/?
lady	62	+/?/+/+	gentleman	26	+/?/+/+
mistress	8	?/+/+/+	mister	5	?/?/?/+
queen	26	+/+/±/+	king	42	+/+/+/+
princess	7	?/+/+/+	prince	6	+/+/+/+
governess	4	+/?/+/*	governor	84	?/+/±/+
duchess	1	?/+/+/*	duke	6	+/+/+/+
empress	0	?/+/+/+	emperor	11	+/+/?/+
baroness	2	?/+/+/+	baron	3	?/+/+/+
witch	10	?/+/+/*	soldier	43	+/+/+/+
actress	17	+/+/±̲/+	actor	43	+/+/±̲/+
waitress	4	+/+/±̲/+	waiter	11	+/+/±̲/+
mare	15	+/?/+/+	stallion	7	+/?/+/*
cow	30	+/+/+/+	bull	29	+/+/±/+
bitch	8	+/+/+/*	dog	85	+/+/+/+
hen	23	+/±/?/+	rooster	5	?/+/+/?
doe	1	?/?/+/*	stag	9	+/?/+/+
	1575			2874	

Table 1: Natural gender seed nouns in English with their frequencies in the WSJ and Brown corpora combined and their projection onto 4 languages (R=Romanian, F=French, E=Spanish, S=Slovene): + means one or more translations, all with correct gender, ± means both gender-correct and incorrect translations, ? no translation, = male and female have the same unique translation, ±̲ both gender-correct and equal translations, ±̳ gender-correct and equal translations, the latter having a wrong gender (e.g. *ligue* and *pareja* in Spanish), * truth was not available (all three Slovene library dictionaries available to us do not include gender information and the MULTEXT lexicon does not contain these nouns)

copil is a translation of both *girl* and *boy*, as well as *daughter* and *son*). All other seeds were retained (including poor ones) in order to avoid use of language-specific knowledge. For example, one of the Romanian seeds, *fiu* (*son*) is also a very frequent form of the verb *a fi* (*to be*), and therefore many verbal contexts are considered by the system from the beginning. One possible approach to correct this problem is to verify that the seeds extracted from the bilingual dictionary have only nominal meanings by checking them individually in a library dictionary. We did not take this approach, trying to keep the amount of supervision minimal.

2.2 Evidence Modeling

Absolute rather than relative (normalized) frequencies are used to capture both the distributional infor-

mation and the reliability of that information (based on sample size). Uncertainty is explicitly represented as a class called "questionable". In this manner, one captures not only the evidence but also the lack of evidence. For example, if a noun w appears in 3 feminine contexts, 1 masculine context, and 6 other contexts with no available gender information, the gender distribution of w is modeled as Frequency(6,3,1) rather than Probability(0.75,0.25).

The level of confidence in the observed distribution is equal to $1 - P(quest)$; in the given example, the confidence is $1 - \frac{6}{6+3+1} = 0.4$ at an absolute frequency of 10.

questionable	masculine	feminine
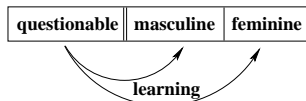
learning

In the bootstrapping process, learning becomes essentially shifting distributional mass from the questionable class to one of real gender classes.

2.3 Context Bootstrapping

The bootstrapping process starts by collecting statistics over the contexts in the corpus with which the seeds co-occur. These are filtered using a frequency threshold sensitive to both the size of the corpus and the seed list. Additional filtering is done by discarding contexts with high relative co-occurrence with words outside the noun list, based on the estimated coverage of this available noun list.

Once reliable contexts in terms of grammatical gender are determined, the distributions for all the words in the noun list are modified according to the number of reliable contexts with which they co-occur. Based on association statistics with these derived contexts, additional non-seed nouns with gender assignments above a threshold are added to the seed list and the bootstrapping algorithm iterates until no more such nouns or contexts are found.

The necessary statistics can be collected both in terms of type, by counting each type of context for each noun only once regardless how many times that context and the word form[3] co-occur in the corpus, and tokens, by counting each context as many times as it appears in the corpus with a given noun. The first approach has the advantage of being less sensitive to the particularities of a language, by not over-weighting any context – especially when seeds such as *fiu* in Romanian have multiple syntactic functions (see Section 2.1). Experiments in Romanian showed that results obtained by considering type statistics (99.79% accuracy at 44.37% coverage) are slightly

---

[3]Because the corpus in not considered POS-annotated, we use all instances of the word-forms in the noun list at the time of processing, even though some of these instances do not have nominal syntactic function.

better than those obtained using token statistics (99.69% accuracy at 45.78% coverage) when starting with natural gender seeds, although the difference is not statistically significant. We chose to use type statistics for all other languages because of their expected robustness to noise in the data.

We investigated 6 models of contextual information: left, right, and bilateral context based on whole words and word suffixes. For Romanian, we discovered that only the left whole-word context is a reliable indicator for gender, the right context and the contextual suffix information containing a lot of noise. For example, some very frequent adjectives (that typically constitute right context) such as *mare* (*big*), *tare* (*hard*), *iute* (*fast* or *spicy*), etc. do not distinguish between masculine and feminine in the indefinite form and they can occur by chance with only one type of noun seeds (especially when considering a small corpus). Also, Romanian verbs (that typically follow the nouns) do not change their form according to gender in most tenses, so they represent mostly noise.

As shown in Table 3 (and also in the tables from Section 3), the context bootstrapping yields very high precision, but the coverage is relatively low. Therefore, once grammatical gender is determined with high precision for a set of nouns, a morphological model to predict gender for the nouns that could not be disambiguated from contexts is used.

2.4 Morphological Analysis

In general, we are interested in predicting the properties of nouns that appear in a corpus. Many of the nouns in the list do not appear in the considered corpus at all and many of the nouns that appear are not in relevant contexts. Therefore, an approach based on morphology (in terms of word endings) has to be considered for these nouns. Table 2 shows the gender distribution over 10 frequent 1-letter suffixes in five languages, as obtained from the noun lists available for these languages.

2.4.1 Variable-length Suffix Interpolation

Tucker et al. (1977, pp 57-62) observed that noun endings are correlated with gender in "a systematic and predictable manner" in French. They found that their French subjects were determining the gender of a noun by making analogies with words that share a long suffix with the target word, using therefore a variable-length rather than a fixed bigram or trigram suffix model. For example, they assign feminine to a noun like *maison* based on the feminine ending *-aison* (*liaison, raison* etc.) rather than on the shorter ending *-on* observed with masculine nouns such as *son, blason, avion* etc.

	Romanian		French		Spanish		Slovene		Swedish	
	f	m/n	f	m	f	m	f	m/n	c	n
-a	130	11	0	0	5414	386	2029	10	2993	119
-ă	3987	38	0	0	0	0	0	0	0	0
-d	1	175	0	0	693	16	20	105	1670	670
-e	4943	148	404	192	167	1395	0	0	2538	2070
-i	14	149	2	20	14	75	0	0	516	214
-n	2	775	256	36	9	74	14	181	2213	596
-o	0	0	0	0	81	6095	0	0	172	65
-r	9	1456	16	62	33	795	5	272	1651	806
-s	2	249	3	40	76	123	4	93	1318	463
-t	2	1908	4	192	0	0	465	263	3939	944

Table 2: Type-based statistics showing the correlation between one-letter suffixes and grammatical gender in 5 languages

Following these observations, we considered a variable-length suffix model, in which nouns with no reliable contexts borrow more gender information from nouns with which they share longer suffixes and less from those with which they share shorter suffixes. This is easily implemented using suffix trie models and taking advantage of the class conditional distribution model by using the questionable mass as a measure of uncertainty at different suffix positions.

2.4.2 Trie Models

Initially, all nouns are introduced in the trie and their class conditional distributions are added to each ramification or terminal node on the path, as shown in the tutorial-example in Figure 2. Again, there are two choices, whether to use type-based statistics (each word has a mass contribution equal to 1 on its path) or token-based statistics (each word has a contribution equal to the number of occurrences in the corpus). Also, a compromise solution would be to weight the contextual class distribution not by the number of occurrences in the corpus, but by the number of relevant contexts with which it has been seen (for example, a word that occurs 6 times from which 3 times in masculine contexts is given more weight than a word occurring 4 times from which twice in masculine contexts).

Once the trie is built, the probability of a gender for a given word w composed of letters $l_1 l_2 ... l_n$ is computed along the path corresponding to w by using a recursive smoothing formula which naturally takes advantage of the distributional representation of uncertainty. The conditional class probabilities given suffix $l_n l_{n-1} ... l_i$ (i.e. the path in the trie is $root - l_n - l_{n-1} - ... - l_i$) are estimated as:

$$\widehat{P}(gen_j | l_n l_{n-1} ... l_i) = P_{node(l_n l_{n-1} ... l_i)}(gen_j) + P_{node(l_n l_{n-1} ... l_i)}(quest) \cdot \widehat{P}(gen_j | l_n l_{n-1} ... l_{i+1}) \quad (1)$$

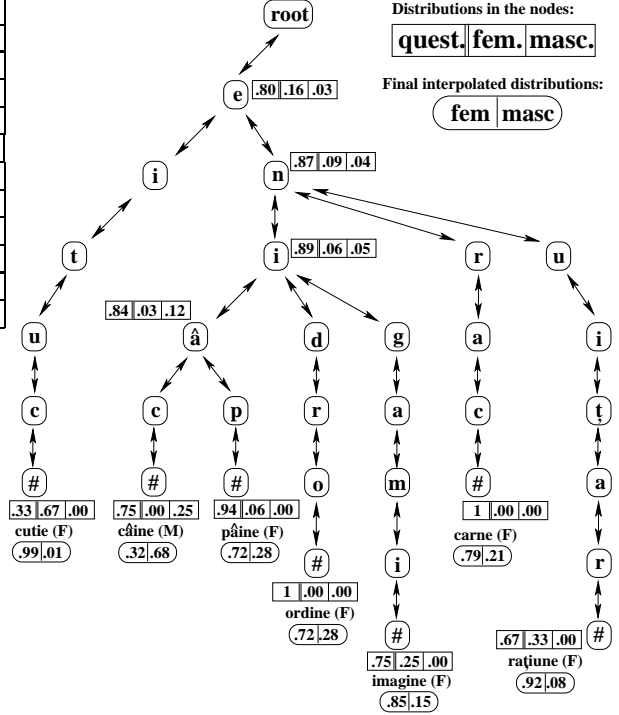The recursion can stop either at root level, where an

Figure 2: Abbreviated example showing the final interpolated feminine/masculine distributions in the leaf nodes. For 5 of the words (*cutie*, *câine*, *pâine*, *imagine*, and *rațiune*) context-based distributions are available, while for *ordine* and *carne* final distributions are based only on morphological trie smoothing. The distributions in rectangular boxes show initial propagation of context-based probabilities from the leaves.

a priori distribution over genders is stored or at any other suffix-length level. We chose to stop at length-1 level. This has the advantage that only information from words that have a suffix of at least one letter in common with the current word (for which a the class-conditional distribution is computed) is used. Because of that, rare nouns that end in letters not seen in any of the nouns for which their gender can be disambiguated through context bootstrapping, are not covered by the trie interpolation algorithm and some other method has to be employed (see 2.5).

A parametrized form of the natural trie-smoothing formula 1 is employed:

$$\widehat{P}(gen_j | l_n l_{n-1} ... l_i) = \lambda_{node,\alpha,\beta} P_{node(l_n l_{n-1} ... l_i)}(gen_j) + \beta P_{node(l_n l_{n-1} ... l_i)}(quest)^\alpha \cdot \widehat{P}(gen_j | l_n l_{n-1} ... l_{i+1}) \quad (2)$$

where $\alpha \in (0, \infty)$, $\beta \in (0, 1)$, and $\lambda_{node,\alpha,\beta} = \frac{1 - \beta P_{node(l_n l_{n-1} ... l_i)}(quest)^\alpha}{1 - P_{node(l_n l_{n-1} ... l_i)}(quest)}$. For example, by choosing $\beta = 0.5$, we enforce that, at each node level on a suffix path, at most half of the probability mass used in the computation of the current node is borrowed from the preceding node on the path.

Romanian	Natural gender seeds		(19 fem., 26 masc.)	
	by type		by token	
2132 nouns	context	+morph.	context	+morph.
coverage	44.37	100	66.25	100
accuracy	99.79	99.25	99.89	97.90

Romanian	System extracted seeds		(17 fem., 23 masc.)	
	by type		by token	
2132 nouns	context	+morph.	context	+morph.
coverage	48.26	100	70.10	100
accuracy	**99.81**	**99.44**	99.90	98.25

Table 3: Results for Romanian

	Romanian	French	Spanish	Slovene	Swedish
unk	0.19%	0.08%	0.03%	0.46%	0.09%
M.L.	0	0	100	10.00	41.18
M.V.	100	100	100	90.00	41.18

Table 4: Percentage of nouns for which predictions cannot be made and the accuracy obtained for these nouns by predicting the most likely class (M.L.) and the class with most endings (M.V.) in the language

2.5 Treatment of "Unknown" Words

We expect that there will still exist some nouns that are rare and unusual (in terms of morphology) for which grammatical gender cannot be predicted by using both the contextual and morphological clues. In the Romanian test set, there were four such nouns: the single letters *A*, *B*, *C*, and the word *şah*, with 35, 15, 5, and 11 occurrences in the corpus respectively. One solution is to predict for these nouns the most common gender in the language. This is not necessarily optimal, because the words corresponding to the most common gender might have very regular morphology. A better solution, which we employ, may be to assign for the uncommon words in terms of morphology the gender with greater suffix variability (masculine in Romanian), as in Table 4.

2.6 Evaluation Methods

We compute the coverage and accuracy of the system considering two estimation methods: type-based (all nouns are considered equally important and get a weight of one) and token-based (each prediction is weighted by the number of occurrences of the target noun in the corpus). For most languages, we had at least a main-POS annotated corpus, so that we could consider only the nominal instances of the words when computing the token coverage and accuracy. When such information is not available, one can approximate the token coverage and accuracy by considering all instances of the word forms from the noun list in the corpus, regardless their true POS.

Noun	Gender	Freq	o ...	un ...
autobuz	m/n	3,930	1	865
cascadă	f	46	15	0
concept	m/n	13,200,000	3,030	185,000
lichid	m/n	10,600	1	607
prostie	f	3,560	954	0
sticlă	f	469	83	0
umilinţă	f	7	0	0
univers	m/n	1,740,000	230	186,000
verificare	f	520,000	3,920	7
viitor	m/n	50,700	1	5,420

Table 5: Statistics for 10 randomly chosen Romanian nouns obtained using Google: the number of hits for the word itself and the word preceded by feminine and masculine contexts (indefinite articles)

This estimation may provide inaccurate results. For example, if true POS were ignored when computing token accuracy in Romanian, the accuracy would be 96.75 instead of 98.25 at 100% coverage, mainly because of two very frequent words, *bine* (meaning *good*) and *cât* (meaning *quotient, how*, and *as*), never used with their nominal senses in the corpus.

2.7 Further Considerations

The main problem encountered is the initial coverage given by the context bootstrapping phase. The coverage can be increased by lowering the thresholds for context reliability, but this has the negative effect of lowering accuracy. The problem occurs primarily because of the size of the corpora used. Thus we ran the following experiment using Google-based searches of the largest corpus publicly available: the Web. Keller et al. (2002) show that the Web can be used as a reliable general source of lexical and syntactical information using search engines.

When using natural gender seeds, the system finds two reliable left contexts, *un* and *o*, which are the indefinite articles in Romanian. When using system extracted seeds, six left contexts are found: *un, nici un, acest, unui* for masculine, *o*, and *nici o* for feminine. We randomly selected 50 nouns from Romanian imposing only the restriction that they are longer than 5 characters (to reduce the chance of getting results from documents in other languages that also have those lexical items) and formed queries of type "masculine-context noun" and "feminine-context noun" using the reliable contexts found by the system. By hypothesizing the gender as the one corresponding to the query with more hits, we obtained 100% accuracy at 94% coverage (the coverage problem is caused by the rarity of some words as well as the fact that most Internet texts in Romanian do not use diacritics in order to be font independent). The frequencies for 10 such nouns are shown in Table 5.

French	Natural gender seeds		(31 fem., 35 masc.)	
	by type		by token	
1317 nouns	context	+morph.	context	+morph.
coverage	77.15	100	86.00	100
accuracy	**97.51**	**95.44**	98.26	97.18

French	System extracted seeds		(19 fem., 29 masc.)	
	by type		by token	
1317 nouns	context	+morph.	context	+morph.
coverage	76.31	100	94.28	100
accuracy	**99.50**	**96.81**	99.73	98.81

Table 6: Results for French

Spanish	Natural gender seeds		(53 fem., 51 masc.)	
	by type		by token	
2993 nouns	context	+morph.	context	+morph.
coverage	54.06	100	72.71	100
accuracy	**98.70**	**95.59**	99.47	98.45

Spanish	System extracted seeds		(18 fem., 30 masc.)	
	by type		by token	
2993 nouns	context	+morph.	context	+morph.
coverage	50.84	100	77.33	100
accuracy	**98.69**	**95.49**	99.51	98.13

Table 7: Results for Spanish

3 Evaluation on Multiple Languages

We chose the following languages both because of the ground-truth resources available to us and because they represent three different language families (Romance, Slavic, and Germanic). Most languages have characteristic gender systems, so various adjustments of the general system described in this paper might be necessary. Our purpose has not been to do such adjustments: all results were obtained by the same system, with identical methodology of computing the parameters (considering corpus size, noun list size, number of occurrences of the known noun word forms in the corpus), as established for Romanian.

3.1 French

French is a Romance language that has two grammatical genders, masculine and feminine. The results presented were obtained using 100,000 sentences from the French side of the Hansards, containing 2,767,775 words. We only had a list of 1,488 gender-annotated French nouns. The left contexts extracted as reliable by the system when starting with system-extracted seeds are the following: masculine: *dernier, bon, ce, quel, nouveau, avant, excellent, Le, présent, grand, le, aucun, meilleur, seul, 1er, premier, certain, un*; feminine: *nouvelle, aucune, quelle, telle, toute, grande, bonne, sa,* and *cette*.

3.2 Spanish

Like French, Spanish has two genders, which normally respect the natural gender for animate beings, but are semantically arbitrary for inanimate nouns. The neuter was eliminated historically, usually in favor of masculine. Most nouns in Spanish (73.5% in our test data) end in what can be considered a gender marker: *-a* for feminine and *-o* for masculine (Table 2). However, there are exceptions to these conventions, and words with other endings than *-o* and *-a* can correspond to either gender.

The performance of the system in Spanish is presented in Table 7. We used a newswire corpus of 12,258 sentences accounting for a total of 369,136 words, and a list of 16,302 nouns extracted from an on-line Spanish-English dictionary.

We were surprised that system performance on Spanish was worse than for French and Romanian. The left contexts extracted by the system-seeded algorithm in the first step were quite reliable (*última, último, misma, buen, una, alguna, nuevo, esa, ese, Esta, Este, algún, citado, otra, otro, todo, cuyo, El, esta, este, primer, el, ningún, ninguna, un*) and would have been expected to yield higher accuracy. However, Spanish has the rather particular characteristic that masculine determiners are used with feminine nouns starting with stressed *a-* or *ha-*. 13 of the 20 errors by type were made for such nouns and they further propagated in the morphology step. This suggests again that there can be language-specific idiosyncrasies limiting performance.

3.3 Slovene

We chose Slovene as a representative for the Slavic languages (more precisely, South Slavonic) because of the availability of gender data and corpora from MULTEXT-East (where, as in Romanian, masculine and neuter are collapsed) and the IJS-ELAN Slovene Corpus (Erjavec, 2002).

Four main groups of stems account for three gender classes: *-i* and *-a* stems are largely feminine, *-o* stems are masculine when followed by a final consonant or neuter when correspond to nouns ending in *-o* or *-e*, consonant stems are mostly neuter.

The absence of definite and indefinite articles in Slovene poses an additional challenge for the bootstrapping algorithm, the language missing some of the most reliable grammatical gender clues. Therefore, we were not surprised to see that the coverage obtained when first running the system using natural gender nouns was very low, as shown in Table 8-a. Starting with system-extracted seeds, the initial coverage given by contextual clues is better and the performance is comparable to those for Romance

Slovene	Natural gender seeds		(44 fem., 40 masc.)	
	by type		by token	
2170 nouns	context	+morph.	context	+morph.
coverage	2.26	100	3.64	100
accuracy	**100**	**90.60**	100	78.32

Slovene	System extracted seeds (27 fem., 19 masc.)			
	by type		by token	
2170 nouns	context	+morph.	context	+morph.
coverage	18.99	100	64.86	100
accuracy	**99.51**	**95.62**	98.18	96.71

Table 8: Results for Slovene

Swedish	Natural gender seeds		(38 ~fem., 41 ~masc.)	
	by type		by token	
19877 nouns	context	+morph.	context	+morph.
coverage	0.30	100	1.81	100
accuracy	44.07	46.21	46.21	45.92

Swedish	System extracted seeds (27 comm., 23 neut.)			
	by type		by token	
19877 nouns	context	+morph.	context	+morph.
coverage	35.61	100	72.73	100
accuracy	**98.84**	**94.41**	99.62	96.50

Table 9: Results for Swedish

languages, as shown in Table 8-b.

3.4 Swedish

Swedish is part of the Northern Germanic (Scandinavian) group of languages. We extracted a gender-annotated noun list from the fine-grained-POS annotated (SUC-1). 150 of the nouns were marked both as neuter and common in the corpus, so they were eliminated from the list of nouns.

Because Swedish gender does not follow a standard masculine/feminine distinction, the projection of the English natural gender nouns (obtained from www.freedict.com) gives results close to random (Table 9-a). This is not necessarily a negative result. Rather, it shows that the system can not only predict gender for languages with masculine/feminine genders but can also identify languages in which the grammatical gender classes do not follow the natural gender distinction.

When starting with system-extracted seeds (labeled correctly as common or neuter), the results become good, especially for the context-bootstrapping step. Although Table 2 shows that a word's last letter alone is not a reliable gender indicator in Swedish, the variable-length suffix approach seems to handle the correlation between morphology and gender well.

4 Conclusion

This paper has presented a set of algorithms for the high-accuracy, minimally supervised induction of grammatical gender. One approach is based on the translingual projection of natural gender, using only the translations of 60 words for which natural gender is marked in English. The second approach uses only a 50 word seed-exemplar set entered from a library dictionary. Using no other language-specific knowledge, the presented bootstrapping algorithm robustly combines both morphological and contextual models to achieve full gender annotation of the nouns in 5 diverse test languages.

References

R. W. Andersen. 1984. What's gender good for anyway? In *Second Languages: A cross-linguistic perspective*, pages 77–100. Newbury House.

S. E. Carroll. 1989. Second-language acquisition and the computational paradigm. *Language Learning*, 39:535–594.

S. Cucerzan and D. Yarowsky. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of CoNLL-2002*, pages 132–138.

T. Erjavec, N. Ide, and D. Tufiş. 1997. Development of common lexical specifications for six Eastern European languages and their application to stochastic part of speech tagging. In *Proceedings of ACH/ALLC'97*.

T. Erjavec. 2002. The IJS-ELAN Slovene-English parallel corpus. *International Journal of Corpus Linguistics*, 7(2):in print.

H. Foundalis. 2002. Evolution of gender in Indo-European languages. CogSci2002. http://www.citeseer.nj.nec.com/545913.html.

J. Hajič and B. Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of COLING-ACL-98*, pages 483–490.

F. Keller, M. Lapata, and O. Ourioupina. 2002. Using the web to overcome data sparseness. In *Proceedings of EMNLP-2002*, pages 230–237.

M. Pérez-Pereira. 1991. The acquisition of gender: what Spanish children tell us. *Journal of Child Language*, 18:571–590.

P. Rodrigues and R. Boivin. 2000. Connectionism and gender assignment in French: Acquisition model or classification model? *Revue Quebecoise de Linguistique*, 28:29–50.

M. E. Sokolik and M. E. Smith. 1992. Assignment of gender to French nouns in primary and secondary language: a connectionist model. *Second Language Research*, 8(1):39–58.

M. E. Surridge. 1993. Gender assignment in French. *International Review of Applied Linguistics*, 31:77–95.

R. Taraban and V. Kempe. 1999. Gender processing in native and nonnative Russian speakers. *Applied Psycholinguistics*, 20:119–148.

G. Tucker, W. Lambert, and A. Rigault. 1977. *The French speaker's skill with grammatical gender: an example of rule-governed behavior*. The Hague: Mouton.