Identification and handling of dialectal variation with a single grammar

António Branco and Francisco Costa

Universidade de Lisboa

Abstract

We present a study on approaches to handle variation in a deep natural language processing formalism. It allows a grammar to be parameterized as to what language variants it accepts, but also to detect such variants. In this respect, we compare it to standard language identification methods, employed here to detect variation in the same language.

1 Introduction

Variation in the same language is often regarded as a problem to categorical approaches of language, and as evidence for its probabilistic dimension (Abney 1996).

In this paper we focus on the problem of handling regional variation within a deep (categorical) natural language processing system, and present a simple way to model variation in a computational grammar using HPSG (Pollard and Sag 1994).

Support and control over variation is obviously important in these systems if they are to have practical application. On the one hand, it is desirable that such systems can cope with the analysis of as many language varieties as possible, since it is less economical to write a different grammar for each language variety. On the other hand, when computational grammars are used for natural language generation, users should be able to put bounds on what is generated variation-wise. Section 2 presents an HPSG design to handle variation in a symbolic model.

A related issue is: if a system can be fine-tuned to a particular regional variety, what is the best way to detect whether some text that is to be processed by that system is in that variety? We present two approaches to this question.

The first approach is to use independent components that can detect the language variety being used. We hypothesize that methods developed for language identification can be used to detect varieties. Section 3 presents an overview and develops on two of them. The second approach is to have the computational grammar prepared for multiple language varieties, with no preprocessing necessary.

We compare the two solutions. To this end we use Portuguese, and we focus on the differences between European Portuguese (henceforth EP) and Brazilian Portuguese (BP). The methods presented are applicable to other languages.

The HPSG setup described to handle variation and the experiments were carried out with a computational HPSG currently being implemented for Portuguese. It is being developed in the LKB (Copestake 2002) and it uses MRS semantics (Copestake, Flickinger, Pollard and Sag 2001). It is part of the DELPH-IN Consortium.[1] The grammar was modest at the time of the experiments (1;6 years of

[1] http://www.delph-in.net

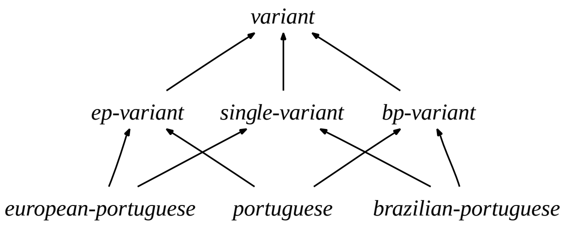

Figure 1: Type hierarchy under *variant*.

development).

2 HPSG Implementation of Variation

In a framework like HPSG, variation can be accounted for in the feature structures manipulated by the grammar.

It is important that the grammar can work with both EP and BP because of coverage, but accepting the two will necessary increase ambiguity. The ability to control variation is important in that it is a way to control the ambiguity generated from accepting both varieties.

Control on what is generated is also desirable. In general one wants to be able to parse as much as possible (e.g. EP and BP), but at the same time be selective in generation (i.e. generate in a specific variety), so that output is tailored to the expected audience.

If a grammar accepts both EP and BP ambiguity will rise because ambiguity inevitably goes up when coverage increases. But ambiguity can be put in check by restricting the grammar to reject analyses that involve marked constructions in more than one variety. More precisely, if an input string contains an element that can only be found in variety v_1 and that same string is ambiguous in a different place but only in varieties other than v_1, this ambiguity will not give rise to multiple analyses if the grammar can be constrained to accept strings with marked elements of at most one variety.

A feature VARIANT is employed to model variation, which encodes the variety of Portuguese being used. It is appropriate for all signs and declared to be of type *variant*. Its possible values are presented in Figure 1.

This attribute is constrained to take the appropriate value in lexical items and constructions specific to one of the two main Portuguese varieties. For example, a hypothetical lexical entry for the lexical item *autocarro* (*bus*, exclusive to EP) would constrain that attribute VARIANT to have the value *ep-variant* and the corresponding BP entry for *ônibus* would constrain the same feature to bear the value *bp-variant*. The only two types that are used to mark signs are *ep-variant* and *bp-variant*. The remaining types presented in Figure 1 are used to perform computations or to constrain grammar behavior, as explained below.

It is not only lexical items that can have marked values in the VARIANT feature. Lexical and syntax rules can have them, too. Such constraints model marked constructions.

Feature VARIANT is structure-shared among all signs that comprise a full parse tree. This is achieved by having all lexical or syntactic rules unify their VARIANT feature with the VARIANT feature of all of their daughters.

If two signs (lexical items, syntax rules) in the same parse tree have different values for feature VARIANT (one has *ep-variant* and the other *bp-variant*), they will unify to *portuguese*, as can be seen in Figure 1. This type means that lexical items or constructions specific to two different varieties are used together. Furthermore, since this feature is shared among all signs, it will be visible everywhere, for instance in the root node.

It is possible to constrain feature VARIANT in the root condition of the grammar. If this feature is constrained to be of type *single-variant* (in root nodes), the grammar will accept EP and BP, but the sentences with properties of both may be blocked. As explained in the previous paragraph, feature VARIANT will have the value *portuguese* in this case, and there is no unifier for *portuguese* and *single-variant*. If this feature is constrained to be of type *european-portuguese* in the root node, the grammar will not accept any sentence with features of BP, since they will be marked to have a VARIANT of type *bp-variant*, which is incompatible with *european-portuguese*.

It is also possible to have the grammar reject EP (using type *brazilian-portuguese*) or to ignore variation completely by not constraining this feature in the start symbol.

With this mechanism, it is possible to use the grammar to detect to which variety input text belongs to. This is done by parsing that text placing no constraint on feature VARIANT of root nodes, and then reading the value of attribute VARIANT from the resulting feature structure: values *ep-variant* and *bp-variant* result from parsing text with features specific to EP or BP respectively; value *variant* indicates that no marked elements were detected and the text could be from both.

This mechanism achieves two goals:

1. Variation can be controlled. A grammar can be parameterized for language variants. Generation can be very specific by choosing values for feature VARIANT low in the type hierarchy, but good coverage variation-wise can be attained in parsing by using a more general type for the same feature. Furthermore, trade-offs between ambiguity and coverage can be explicitly controlled via intermediate types, like *single-variant*.

2. Language variants can be detected in the input.

If the input can be known to be specifically EP or BP before it is parsed, the constraints on feature VARIANT can be changed to improve efficiency. When parsing text known to be EP, there is no need to explore analyses that are markedly BP, for instance.[2]

[2] Currently, it is not possible to prune the parser's search space in such circumstances with the LKB,

It is thus interesting to know what other methods can do to detect varieties, and how they compare to the one just introduced, using real world data. In Section 3, some language identification models that can be used for this purpose are presented.

Because the two aspects of controlling variation and detecting variants are related by a single design, we assume that evaluating one indirectly evaluates the other. Therefore by investigating how good a grammar with such a mechanism can be in detecting language varieties, we can also have an idea of how well the same mechanism is used for the purpose of controlling ambiguity or specificity.

3 Language Detection Methods

Over the last years methods have been developed to detect the language a given text is written in. They have also been used to discriminate varieties of the same language, although less often. They can be based on words in text. Lins and Gonçalves (2004) look up words in dictionaries to discriminate among languages, and Oakes (2003) runs statistical tests on word frequencies, like the chi-square test, in order to differentiate between British and American English.

Many methods are based on frequency of byte sequences (byte n-grams) in text, because they can simultaneously detect language and character encoding (Li and Momoi 2001), and can reliably classify short portions of text, since they look at such short sequences. They have been applied in web browsers (to identify character encodings as in Li and Momoi (2001)) and information retrieval systems.

We are going to focus on methods based on character n-grams. Because all information used for classification is taken from character n-grams, and they can be found in text in much larger quantities than words or phrases, sparse data problems are attenuated. Therefore, high levels of n or very small training corpora can be used. Training data can also be found in large amounts because training corpora do not need to be annotated (it is only necessary to know the language they belong to).

More importantly, methods based on character n-grams can reliably classify small portions of text. The literature on automatic language identification mentions training corpora as small as 2K producing classifiers that perform with almost perfect accuracy for test strings as little as 500 Bytes (Dunning 1994) and considering several languages. With more training data (20K-50K of text), similar quality can be achieved for smaller test strings (Prager 1999).

Many n-gram based methods have been used besides the ones we present. Sibun and Reynar (1996) and Hughes, Baldwin, Bird, Nicholson and MacKinlay (2006) present good surveys. Many can achieve perfect or nearly perfect classification with small training corpora on small texts, so we just focus on two that use approaches very well understood in language processing and information retrieval.

because it is only possible to constrain the root node without changing and reloading the grammar. Therefore, incompatible analyses will only be discarded when that node is built, but not before that. Efficiency could be gained if it were possible to specify constraints that all nodes in a syntactic tree must obey. The limitation is system dependent, so, in theory, efficiency can be improved in such a way.

3.1 Markov Models

If one wants to know which language $L_i \in L$ generated string s, one can use Bayesian methods to calculate the probabilities $P(s|L_i)$ of string s appearing in language L_i for all $L_i \in L$, the considered language set, and decide for the language with the highest score (Dunning 1994). That is, in order to compute $P(L_i|s)$, we only compute $P(s|L_i)$. The Bayes rule allows us to cast the problem in terms of $\frac{P(s|L_i)P(L_i)}{P(s)}$, but, as is standard practice, we drop the denominator, since we are only interested in getting the highest probability score among several scores, not its exact value. The prior $P(L_i)$ is also ignored, assuming all languages are equally probable.

The way $P(s|L_i)$ is calculated is also the standard way to do it, namely assuming independence and just multiplying the probabilities of character c_i given the preceding $n-1$ characters (using n-grams), for all characters in the input string (which are estimated from n-gram counts in the training texts).

We implemented the algorithm as described in Dunning (1994) for the experiments presented in the following sections, which uses other common strategies, like prepending $n-1$ special characters to the input string to harmonize calculations, summing logs of probabilities instead of multiplying them to avoid underflow errors, and using Laplace smoothing to reserve probability mass to events not seen in training.

3.2 Vector Space Models

The second method using n-grams we employ in the following experiments is inspired in the vector space model of information retrieval to compare document similarity, but it uses n-gram counts instead of term frequency. It has been used for the purpose of language identification in Prager (1999).

Each language is represented by a vector built during training. Each possible n-gram corresponds to a component of that vector (e.g. if bigrams are used, the first component might represent the bigram *aa*), namely a number based on the frequency of occurrence of that that n-gram in the training corpus for that language.[3] Classification consists of creating a vector representing the input text in a similar way and choosing its nearest neighbor from the set of vectors that represent languages. The cosine of the angle between the two vectors is used as a measure of similarity. A number of well-known improvements can be used, like normalizing vectors in the training phase (make them of length = 1), so that calculating cosines amounts to calculating dot products during classification (after normalizing the vector representative of the test item).

In the literature it is also common to reduce dimensions by keeping the most frequent n-grams and discarding the rest, but we did not do this since we hypothe-

[3] In information retrieval tf-idf is often used (term frequency times inverse document frequency). Here we use n-gram frequency in that language divided by the frequency of that n-gram in all languages. Both numbers are estimated from the training corpora. Note that this is a literal interpretation of inverse document frequency: it is common practice to use a value based on that instead, like its log; but Prager (1999) reports that the literal version performs better for language identification.

size that the most frequent n-grams of EP and BP will largely overlap. It has been reported that the 300 most frequent n-grams are good predictors of language, and the others are representative of the textual topic (Cavnar and Trenkle 1994).

4 Data and Calibration

Some preliminary studies were conducted in order to investigate the performance of the language identification methods presented above at discriminating among languages (Section 4.1), and to find out the impact of training corpora size when they are employed to detect language variants and what values of n are reasonable (Section 4.3 and Section 4.4). The data used in all experiments concerning variety identification are presented in Section 4.2.

4.1 Language Identification Methods at Identifying Languages

We want to check that the language identification methods we are using are in fact reliable at identifying different languages. Although the literature reports good results, we wanted to test the exact implementation we will be using in distinguishing between EP and BP.

We ran the two classifiers on three languages showing strikingly different characters and character sequences. This is a deliberately easy test to get an upper bound on what these methods can do.

For this test we used the Universal Declaration of Human Rights texts.[4] The languages used were Finnish, Portuguese and Welsh. Human inspection of texts in these languages immediately reveals highly idiosyncratic character sequences.[5]

The Preamble and Articles 1–19 were used for training (8.1K of Finnish, 6.9K of Portuguese, and 6.1K of Welsh), and Articles 20–30 for testing (4.6K of Finnish, 4.7K of Portuguese, and 4.0K of Welsh). Because these methods perform better if the text they are classifying is large, several tests were conducted, splitting the test data in chunks of text 1, 5, 10 and 20 lines long.

The Bayesian method obtained perfect accuracy on all test conditions (all chunk sizes), for all values of n between 1 and 7 (inclusively). For $n = 8$ and $n = 9$ there were errors only when classifying 1 line long test items. The vector space model obtained perfect accuracy on all test conditions, for all values of n between 2 and 8 (inclusively). For $n = 1$ and $n = 9$ there were errors once again only when classifying 1 line long test items.

[4] Available at http://www.unhchr.ch/udhr/navigate/alpha.htm.

[5] For the sake of illustration, examples (1), (2) and (3) present the first sentence of the first Article in Finnish, Portuguese and Welsh, respectively. (4) is the English version.

(1) Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan.

(2) Todos os seres humanos nascem livres e iguais em dignidade e em direitos.

(3) Genir pawb yn rhydd ac yn gydradd â'i gilydd mewn urddas a hawliau.

(4) All human beings are born free and equal in dignity and rights.

The average line length for the test corpora was 138 for Finnish, 141 for Portuguese and 121 for Welsh (133 overall). In the corpora we will be using in the following experiments, average line length is much lower (around 40 characters per line). Input length is obviously important for these methods. To make the results more comparable, we also evaluated these classifiers of Finnish, Portuguese and Welsh with the same test corpora, but truncated each line beyond the first 50 characters, yielding test corpora with an average line length around 38 characters (since some were smaller than that).

The results are similar, just slightly worse. The Bayesian classifier performed with less than perfect accuracy also with $n = 7$ when classifying 1 line at a time. The vector based classifier performed worse only with $n = 2$ and 1 line long test items. In all these less than perfect cases, accuracy was in the 80–90% range.

These methods thus perform very well at discriminating languages with reasonable values of n and can classify short bits of text, even with incomplete words.

4.2 Data

For the experiments on variety detection, we used two corpora from Portuguese and Brazilian newspaper text. They are CETEMPublico and CETENFolha. CETEMPublico contains text from the Portuguese newspaper *O Público*, and CETENFolha from the Brazilian *Folha de São Paulo*.

These corpora are minimally annotated (paragraph and sentence boundaries, *inter alia*), but are very large (CETEMPublico has 204M words and 1.2GB of text, and CETENFolha has 32M words and 183.2 MB).

Some preprocessing was carried out: all XML-like tags, like the <s> and </s > tags that mark sentence boundaries, were removed. Some heuristics were also employed to remove lines that are parts of lists, like sports results tables or music charts, since they might not be representative of language: only lines ending in ., ! and ? were considered, and lines containing less than 6 words (defined as strings delimited by whitespace) were discarded. Other character sequences that were judged irrelevant for the purpose at hand were normalized: URLs were replaced by the sequence URL , e-mail addresses by MAIL , hours and dates by HORA and DATA , etc. Names at the beginning of lines indicating speaker were removed since they are frequent and the grammar that will be used cannot parse name plus sentence strings.

4.2.1 The 400K Line Corpus

We ordered the remaining lines by line length in terms of words and kept the smallest 200K lines from each of the two corpora. Small lines were preferred as they are more likely to receive an analysis by the grammar.

From these 200K lines of text from each corpus, we randomly chose 20K lines for testing and the remaining 180K for training. This produced a large data set, that allows one to check how good n-grams based methods can be in detecting varieties given enough data, and what values of n are necessary. Since language varieties are more similar to each other than languages, it is expected that more data or more context will be required for comparable results. In the tests below, we refer to this data set as the 400K line corpus.

We assume that the sentences from the Portuguese corpus contain text belonging to EP, and that the sentences in the Brazilian corpus represent BP text. This is a simplification, since they can contain transcriptions from speakers of the other variety. A classification is thus considered correct if the classifier can guess the newspaper the text was taken from.

4.2.2 The 30KB Corpus

The use of two corpora, one from EP and the other from BP, does not allow the training of n-grams based classifiers to detect sentences that are possible in both EP and BP, because only a two-way classification is present in the training data, but we want these classifiers to produce a three-way distinction. If a sentence is found in the EP corpus, one can be relatively certain that it is possible in EP, but one does not know if it is BP, too. The same is true of any sentences in the BP corpus — it can also be a sentence of EP.

To address this limitation, a native speaker of EP was asked to manually decide from sentences found in the BP corpus whether they are markedly BP or are also acceptable in EP. Conversely, a Brazilian informant detected markedly European sentences from the EP corpus.

Because this task requires manual annotation, and the methods we are employing reportedly perform well even with small training sets (when identifying languages), we used only a small portion of text taken from these corpora.

We randomly selected 90K lines of text from each corpus and checked which ones could be parsed by the grammar. 25K lines of parsable BP and 21K of parsable EP (46K lines out of 180K, or 26%) were obtained. From these parsed lines we drew around 1800 random lines of text from each corpus, and had them annotated for whether they are possible in the other variety. Thus a three-way classification is obtained.

Perhaps not surprisingly, most of the sentences were judged to be possible in both EP and BP. 16% of the sentences in the Portuguese corpus were considered impossible in BP, and 21% of the sentences in the BP corpus were judged exclusive to it. Overall, 81% of the text was common to both varieties.

A hypothetical explanation of the asymmetry is that one of the most pervasive differences between EP and BP, clitic placement, is attenuated in writing: Brazilian text often displays word order between clitic and verb similar to EP, and different from oral BP. Therefore, European text displaying European clitic order does not look markedly European. In fact, we looked at the European sentences with clitic placement characteristic of EP that were judged possible in BP. If they were included in the markedly European sentences, 23% of the European text would be unacceptable BP, a number closer to the 21% sentences judged to be exclusively Brazilian in the Brazilian corpus.

Such information can be used to estimate prior probabilities for the Bayesian

Length of Test Item		1 line	5 lines	10 lines	20 lines
$n = 2$	Bayesian	0.84	0.99	**1**	**1**
	Vector based	0.62	0.59	0.56	0.52
$n = 3$	Bayesian	**0.96**	0.99	**1**	**1**
	Vector based	0.63	0.59	0.61	0.65
$n = 4$	Bayesian	**0.96**	**1**	**1**	**1**
	Vector based	0.63	0.73	0.79	0.87
$n = 5$	Bayesian	0.94	**1**	**1**	**1**
	Vector based	0.65	0.81	0.89	0.97
$n = 6$	Bayesian	0.92	0.99	**1**	**1**
	Vector based	0.67	0.86	0.94	0.98

Table 1: Precision with 360K lines of text for training, two-way classification.

method (which, as referred in Section 3.1, are ignored), creating a bias for classifying text as common to all varieties of Portuguese. This was not done, because like what happens for estimating the priors of any language in a set of languages in general, the difference between the priors of EP and BP is very difficult or even impossible to obtain.

The data were split into test and training data, but only a subset of what was judged common to both varieties was kept, since that data set was much larger than the other two. 10KB of text from each class were obtained. 5KB (of each class) were reserved for training and another 5KB for test. These values are close to the ones used for language discrimination in Section 4.1. There are approximately 140 lines for each class. For the test corpora, we kept exactly 140 lines for each: a multiple of 20 is convenient, because we want to create chunks of 1, 5, 10 and 20 lines for testing. In the following tests, this data set is referred to as the 30KB corpus.

4.3 Two-way Distinction with the 400K Line Corpus

Table 1 summarizes the results for the n-grams trained to distinguish between EP and BP with the 400K line corpus. The average line length of the test sentences is 43 characters. Several input lengths were tried out by dividing the test data into various sets with varying size.

The accuracy of the Bayesian classifier is surprisingly high, given that we can estimate the number of sentences that cannot be attributed to a single variety to be at least 80% (see Section 4.2.2). We hypothesize that this is a corpus sensitivity effect. For instance, German names are more frequent in Brazil. In fact, the absolute frequency $f_{Tr}(x)$ of n-gram x in the training data for n-grams *sch*/*Sch*, *ung*, *W* and *en*␣[6] is $f_{Tr}(\mathit{sch} \vee \mathit{Sch}) = 311$, $f_{Tr}(\mathit{ung}) = 194$, $f_{Tr}(\mathit{W}) = 1122$ and $f_{Tr}(\mathit{en}␣) = 529$ in Brazilian text and $f_{Tr}(\mathit{sch} \vee \mathit{Sch}) = 205$, $f_{Tr}(\mathit{ung}) = 98$,

[6] ␣ denotes a space.

Length of Test Item		1 line	5 lines	10 lines	20 lines
$n = 2$	Bayesian	**0.86**	**0.98**	**0.96**	**1**
	Vector based	0.61	0.75	0.86	0.85
$n = 3$	Bayesian	0.82	0.73	0.64	0.5
	Vector based	0.64	0.61	0.71	0.79
$n = 4$	Bayesian	0.68	0.55	0.5	0.5
	Vector based	0.64	0.71	0.79	0.93

Table 2: Precision with 10K lines of text for training, two-way classification.

$f_{Tr}(W) = 680$ and $f_{Tr}(\mathit{en}\textvisiblespace) = 305$ in Portuguese text. This might also explain the lower performance of the vector space model, where infrequent n-grams have a lower impact on the result since the individual values derived from the n-grams are summed together rather than multiplied.

The amount of training data is very large because these methods look at characters. There are 15.5M of them in the training sets. The fact that relatively high values of n (4 and 5 for 5 lines of input) are necessary to achieve perfect accuracy on small inputs (and perfection is never found with 1 line long test items) suggests that variety discrimination is much harder than language identification.

4.4 Two-way Distinction with the 30KB Corpus

The same experiment was conducted, using only the EP and BP data (not the sentences judged to be common to both) of the 30KB corpus (only 20KB of it).

Although the size of training data is much smaller than in the test reported in Section 4.3, the two classes are expected to be farther apart since sentences judged to be common to the two varieties were not included.

The results are in Table 2. The Bayesian classifier is very good with bigrams, but because of the small training data, it is heavily biased at classifying everything as EP. The vector space model cannot achieve as good performance with bigrams, but is less affected by sparseness of training data.

4.5 Differences between EP and BP

We proceeded to an analysis of the training data resulting from the manual classification described in Section 4.2.2 (the 30KB corpus). A brief typology of the markedly Brazilian elements found in the BP training corpus is presented. We also present the relative frequency of these phenomena based on the same data.

1. Mere orthographic differences (24%)
   e.g. *ação* vs. *acção* (*action*)

2. Phonetic variants reflected in orthography (9.3%)
   e.g. *irônico* vs. *irónico* (*ironic*)

3. Lexical differences (26.9% of differences)
    (a) Different form, same meaning (22.5%)
    e.g. *time* vs. *equipa* (*team*)
    (b) Same form, different meaning (4.4%)
    e.g. *policial* (*policeman/criminal novel*)
4. Syntactic differences (39.7%)
    (a) Possessives without articles (12.2%)
    (b) In subcategorization frames (9.8%)
    (c) Clitic placement (6.4%)
    (d) Singular bare NPs (5.4%)
    (e) In subcat and word sense (1.9%)
    (f) Universal *todo* occurring with article (0.9%)
    (g) Contractions of preposition and article (0.9%)
    (h) Questions without subject-verb inversion (0.9%)
    (i) Postverbal negation (0.5%)
    (j) other (0.5%)

One third of the differences found would be avoided if the orthographies were unified (items (2) and (1)).

Some differences cannot be detected by n-gram based methods or the grammar. This is the case of item (3b), which would require word sense disambiguation. When word sense differences are accompanied by different syntax, they can be detected by the grammar (item (4e)) in limited circumstances (in that example, only if the complement is expressed).

Differences that are reflected in spelling can be modeled by the grammar via multiple lexical entries, with constraints on feature VARIANT reflecting the variety in which the lexical with that spelling item is used.[7]

Interestingly, 40% of the differences are syntactic. These cases are expected to be difficult to detect with n-gram based approaches, but not by a grammar.

Note that on average each sentence contained 1.46 marked elements. Spelling differences (items (2) and (1)), which account for 33.3% of all differences appear in 47.9% of them (in the BP training corpus). N-grams models can detect them.[8]

In the Portuguese grammar we use for the experiments, only clitic word order (item (4c)) and co-occurrence of prenominal possessives and determiners (item

[7] In some cases a different solution would be preferable. When the difference is systematic (e.g. the EP sequence *´on* always corresponds to a BP sequence *ˆon*, with an example in item (2)), it would be best to have a lexical rule that affects only spelling and the VARIANT feature producing one variant from the other. This is not implemented, because string manipulation is limited in the LKB.

[8] Even when these differences are not absolute, they are often strongly unbalanced. For instance, the bigram *ct* appears 22 times in the Portuguese training corpus and only once in the Brazilian one.

(4a)) are marked with respect to the VARIANT feature. The main limitation is grammar immaturity, in that several differences involving phenomena that are not implemented yet cannot be taken into account. These two phenomena do account for 18.6% of the differences found.

We expanded the grammar with many lexical items markedly EP or markedly BP. These were taken from the Portuguese Wiktionary,[9] where this information is available. We did not include all of the ones there, since some were judged infrequent and manual expansion of a lexicon for a deep grammar is time consuming. At the end, around 740 lexical items were added. Variety specific lexical items found in the training corpora (80 more) were also incorporated in the lexicon.

5 Results

We report on the evaluation of the n-gram based methods presented in Section 3 and the grammar-based mechanism to handle variation described in Section 2, tested with the 30KB corpus (Section 4.2.2).

When the grammar produced multiple analyses for a sentence, we only considered that sentence to be classified as EP if all the parses produced VARIANT with type *ep-variant*, and similarly for BP. In all other cases the sentence would be considered common to both.

The grammar can only look at one line at a time, but several input sizes are tested. In order to make the grammar results comparable, this is done also for the grammar. In this case, the result for chunks of more than one line is the unification of the values for each line. If the unification result is *portuguese* (see the hierarchy in Section 2), signaling inconsistency, the grammar does not decide, affecting recall but not precision. For this reason, precision, recall, and the F-measure can be different and are all reported. With the n-grams based models, they are always identical. The results for the three-way classification are in Table 3.

Error analysis shows that the BP sentences classified as EP contain clitics following the EP syntax, and misspellings conforming to the EP orthography.[10] Most of the sentences common to EP and BP that were classified as EP also present clitics with this syntax. A large proportion of the errors consisted in classifying as common to all varieties of Portuguese sentences that were in fact marked. Inspection of these sentences reveals many marked lexical items.[11] It is thus a problem of lexical coverage.

The Bayesian method works well with small values of n, but it tends to classify everything as EP, producing correct classifications for only one third of the test items. The vector space model is more affected by input length.

[9] http://pt.wiktionary.org

[10] In Brazil a diaeresis is used on *u* when it follows *q*, precedes *e* or *i* and is pronounced. In the Portuguese orthography it is no longer used. The errors were due to the spellings *aguentar* (*to bear*) and *tranquilo* (*calm*), instead of *agüentar* and *tranqüilo*.

[11] Note that, in order to increase grammar coverage, we used a POS tagger to get information about unknown words. Obviously, feature VARIANT was left underspecified in these items.

Length of Test Item		1 line	5 lines	10 lines	20 lines
	Precision	0.57	**0.78**	0.72	0.64
Grammar	Recall	0.57	**0.72**	0.62	0.43
	$F_{\alpha=1}$	0.57	**0.75**	0.67	0.51
$n = 2$	Bayesian	**0.59**	0.67	**0.76**	**0.76**
	Vector based	0.43	0.52	0.55	0.57
$n = 3$	Bayesian	0.55	0.52	0.45	0.33
	Vector based	0.47	0.48	0.67	**0.76**
$n = 4$	Bayesian	0.48	0.39	0.33	0.33
	Vector based	0.41	0.5	0.71	0.67

Table 3: Evaluation of variety identification, three-way classification. With the n-grams based method, precision, recall and the F-measure are identical under the same conditions.

6 Discussion and Conclusions

Before getting into the analysis of the quantitative results obtained above, some remarks on the two approaches, with the grammar and with the stochastic techniques, follow from the very nature of these methods.

Bayesian and vector-similarity methods are expected to be easier to scale up with respect to the number of varieties considered given that the size of the type hierarchy under *variant* is exponential on the number of language varieties if all variety combinations are taken into account.[12]

In turn, provided the symbolic method is supported by a more matured grammar than the one we could use in the present experiments, with a large enough lexicon, stochastic methods are expected to show more dependency on the text domain they are applied to than the grammar, and it is likely that their performance tends to degrade more severely when applied over texts from domains which they were not trained with.

Focusing on the results obtained with the grammar, the fact that the best score results from setting the input with 5 lines/sentences is understandable at the light of the following considerations: on the one hand, taken individually, there is a certain chance that each sentence ends up not being specified with respect to any language variant at stake; on the other hand, when they are bundled together, there happens the incremental effect that the resolution obtained at one or several of them in each bunch unifies with the underspecified values of the remaining ones that did not get resolved; however, when they are bundled into a too large bunch ($\geq$ 10 lines/sentences) chances also increase that different sentences get different specifications, which induces incorrect or even non resolution for the whole bunch, thus canceling the beneficial effect of the sentences being bundled together.

By the same token, it is also worth noting that with larger bundles, the perfor-

[12] This may be necessary. For instance, *bu´e* (*very*, *much*) is a word in European and Angolan Portuguese, but not in Brazilian Portuguese; *moleque* (*boy*) is a word in Angolan and Brazilian Portuguese, but not in EP, etc.

mance of classifiers based on the grammar is thus expected to degrade more than the performance of classifiers based on n-grams.[13]

Note however that this may not be a shortcoming for the grammar-based methods in every application scenario. For instance, when the input text is a dialog, such input may have to be entered in small chunks (a chunk per turn) if one wants to contemplate conversations between speakers of different varieties.

Turning now to the evaluation results obtained above, both the grammar and the stochastic approaches displayed similar results. In both cases the best score is around $F = 0.75$.

For both approaches, our experiments were limited in several respects and there is plenty of room for improvement. The n-grams methods can be enhanced by using more training data, since only 15KB were used. With the grammar, lexical coverage can be augmented, and more marked constructions can be added — the syntactic differences considered cover half of the occurrences of all syntactic differences found in the BP training data (Section 4.5).

In spite of the limitations of these first experiments, results are encouraging. The design we presented to account for variation can be adapted to other feature-type formalisms, and the experimental setup used to compare performance in face of language varieties, which takes into account the fact that they largely overlap, is new and extensible to other languages as well.

References

Abney, S.(1996), Statistical methods and linguistics, *in* J. Klavans and P. Resnik (eds), *The Balancing Act*, The MIT Press, Cambridge, MA.

Cavnar, W. B. and Trenkle, J. M.(1994), N-gram-based text categorization, *Proceedings of the 1994 Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV USA.

Copestake, A.(2002), *Implementing typed feature structure grammars*, CSLI Publications, Stanford, California.

Copestake, A., Flickinger, D., Pollard, C. and Sag, I. A.(2001), Minimal Recursion Semantics: An introduction, *Language and Computation*.

Dunning, T.(1994), Statistical identification of language, *Technical Report MCCS-94-273*, Computing Research Lab (CRL), New Mexico State University.

Hughes, B., Baldwin, T., Bird, S., Nicholson, J. and MacKinlay, A.(2006), Reconsidering language identification for written language resources, *Proceedings of LREC2006*, Genoa, Italy.

Li, S. and Momoi, K.(2001), A composite approach to language/encoding detection, *Proceedings of the Nineteenth International Unicode Conference*.

Lins, R. D. and Gonçalves, P.(2004), Automatic language identification of written texts, *Proceedings of the 2004 ACM Symposium on Applied Computing*.

Oakes, M. P.(2003), Text categorization: Automatic discrimination between US

[13] Recall for common Portuguese is 0.89 in the 1 line test, and 0.14 in the 20 lines case. Overall, 68% of the test items were classified as common in the 1 line test, but only 5% in the 20 lines test.

and UK English using the chi-square test and high ratio pairs, *Research in Language*.

Pollard, C. and Sag, I.(1994), *Head-driven phrase structure grammar*, Chicago University Press and CSLI Publications.

Prager, J. M.(1999), Linguini: Language identification for multilingual documents, *Journal of Management Information Systems*.

Sibun, P. and Reynar, J. C.(1996), Language identification: Examining the issues, *5th Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, U.S.A.