Accommodating Language Variation in Deep Processing

António Branco and Francisco Costa
University of Lisbon

Proceedings of the GEAF 2007 Workshop

Tracy Holloway King and Emily M. Bender (Editors)

CSLI Studies in Computational Linguistics ONLINE

Ann Copestake (Series Editor)

2007

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

We present an approach to handle variation in deep linguistic processing. It allows a grammar to be parameterized as to what language variants it accepts and also to detect the variant of the input. We also report on the evaluation of this approach by having the system detect the dialect of input sentences extracted from corpora of two different dialects.

1 Introduction

This paper proposes a design strategy for deep language processing grammars to appropriately handle language variants of a given language.

In the benefit of generalization and grammar writing economy, it is desirable that a grammar can handle language variants — that is variants which share most grammatical structures and lexicon — in order to avoid the multiplication of individual grammars, motivated by inessential differences.

The design presented here allows a grammar to be restricted as to what language variant it is tuned to, but also to detect the variant a given input pertains to. Evaluation of this design is also reported.

We assume the HPSG framework (Pollard and Sag, 1994) and a grammar that handles two close variants of the same language, European and Brazilian Portuguese. These assumptions are merely instrumental, and the results obtained can be easily extended to other languages and variants, and to other grammatical frameworks for deep linguistic processing.

The HPSG setup for handling variation and the experiments themselves were carried out with a computational HPSG for Portuguese. It is being developed in the LKB (Copestake, 2002) on top of the Grammar Matrix (Bender et al., 2002), and it uses MRS for semantic description (Copestake et al., 2001). This grammar is part of the DELPH-IN Consortium (http://www.delph-in.net).[1]

2 Handling variation

We propose an approach that allows flexibility with respect to variation in the same language and also permits a grammar to be tuned to a particular variant. It relies on the use of a feature VARIANT to model variation. This feature is appropriate for all signs, and its value declared to be of type *variant*. Given the working language variants assumed here for the sake of the evaluation experiment, its possible values are the ones presented in Figure 1.

This attribute is constrained to take the appropriate value in lexical items and constructions specific to one of the two varieties. For example, a hypothetical lexical entry for the lexical item *autocarro* (*bus*, exclusive to European Portuguese) would include the constraint that the attribute VARIANT has the value *ep-variant*,

[1] At the time of the experiments reported here, the grammar was of modest size, resulting from a year and a half of development.

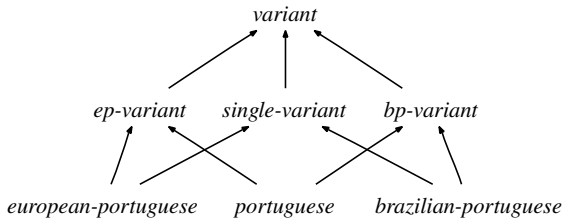

Figure 1: Type hierarchy under *variant*.

and the corresponding Brazilian Portuguese entry for *ônibus* would constrain the same feature to bear the value *bp-variant*. Items that are common to both European Portuguese and Brazilian Portuguese are left underspecified with respect to this feature. They do not have to be constrained with [VARIANT *variant*] because this constraint is defined in the type *sign*, from which all lexical types inherit.

Figure 2 shows examples of these cases, with simplified feature structures. 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 constrain grammar behavior, as explained below.

$$\begin{bmatrix} \text{STEM} \left\langle \text{“autocarro”} \right\rangle \\ \text{VARIANT } \textit{ep-variant} \end{bmatrix} \begin{bmatrix} \text{STEM} \left\langle \text{“ônibus”} \right\rangle \\ \text{VARIANT } \textit{bp-variant} \end{bmatrix} \begin{bmatrix} \text{STEM} \left\langle \text{“carro”} \right\rangle \\ \text{VARIANT } \textit{variant} \end{bmatrix}$$

Figure 2: Constraints on lexical items. Example of an European Portuguese item (*autocarro — bus*), a Brazilian Portuguese item (*ônibus — bus*) and an item common to both varieties (*carro — car*).

Lexical items are not the only elements that can have marked values in the VARIANT feature. Lexical and syntax rules can have them, too. Such constraints model constructions that markedly pertain to one of the dialects. Section 4 presents a small examination of these differences.

The feature VARIANT is structure-shared among all signs comprised in a full parse tree. This is achieved by having all lexical or syntactic rules unify their VARIANT feature with the VARIANT feature of their daughters (Figure 3).

Since this feature is shared among all signs, it will be visible everywhere, including the root node. It is possible to constrain the feature VARIANT in the root condition of the grammar so that the grammar works in a variant-“consistent” fashion: this feature just has to be constrained to be of type *single-variant* (in root nodes) and the grammar will accept either European Portuguese or Brazilian Portuguese. Furthermore, in the unnatural condition where the input string bears marked properties of both variants (e.g. from lexical items and syntax rules), that string will receive no analysis: the feature VARIANT will have the value *portuguese* in this case (the greatest lower bounds for *ep-variant* and the other *bp-variant*), and there is no unifier for *portuguese* and *single-variant*.

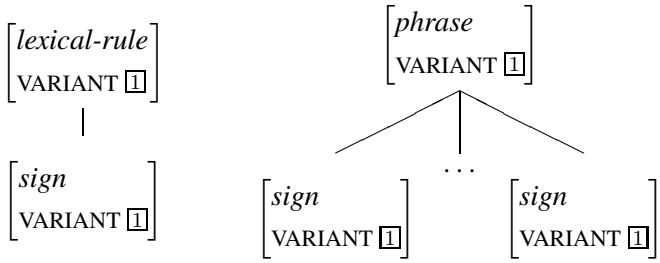

Figure 3: Constraints on rules. Lexical and syntax rules identify the VARIANT feature of the mother with the VARIANT features of all the daughters.

Figure 4 shows an example of this situation, where the marked Brazilian item *dezesseis* (sixteen) co-occurs with the marked European item *autocarros* (buses). This is specially useful in generation, where one may be interested in generating all relevant sentences in either European Portuguese or Brazilian Portuguese, but one does not want to generate sentences with phrases like the one in this example.

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 Brazilian Portuguese, since these will be marked to have a VARIANT of type *bp-variant*, which is incompatible with *european-portuguese* (there is no unifier for these two types according to the hierarchy in Figure 1). It is also possible to have the grammar reject European Portuguese sentences in detriment of Brazilian Portuguese ones (by using type *brazilian-portuguese*) or to ignore variation completely by assigning to VARIANT the *variant* value, thus not constraining the VARIANT feature in the start symbol.

The mechanism presented here has the following properties:

- Increased coverage and flexibility. The grammar can handle input from all variants under consideration if the VARIANT feature is constrained with a general type.
- Parameterization. The grammar can be tuned to a relevant dialect by constraining the feature VARIANT with a specific type. This is welcome in parsing, but specially desirable in generation, where the grammar can be configured to generate only in a given selected variant.
- Consistency. If VARIANT is constrained to be *single-variant*, the grammar can deal with all variants, but it will reject sentences with mixed characteristics.

The ability to parse more variants means more coverage, which generally increases ambiguity. The last two properties above are ways to control this kind of ambiguity. If the input string contains an element that can only be found in variety v_1 and that input string yields ambiguity in a different stretch but only in varieties

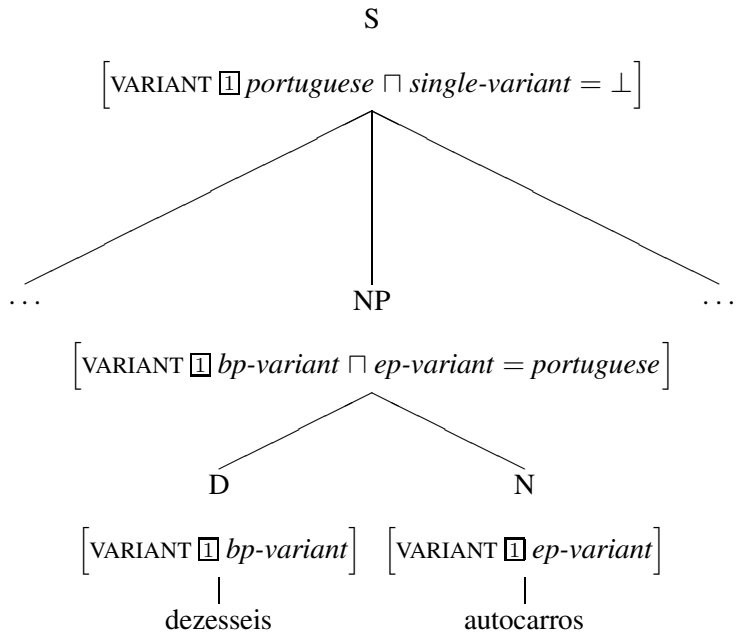

Figure 4: Example of an inconsistency. The noun phrase *dezesseis autocarros* (*sixteen buses*) is inconsistent. It should be either *dezesseis ônibus* (Brazilian Portuguese) or *dezasseis autocarros* (European Portuguese). The constraint on the VARIANT of the root node (to be *single-variant*) rejects the structure.

v_k other than v_1, this ambiguity will not give rise to multiple analyses if the grammar is constrained to accept strings with marked elements of at most one variety.

This can be illustrated with a simple example. The preposition *a* (to, at) is homonymous with a form of the definite article. In European Portuguese, in many contexts definite articles are obligatory before possessives, but in Brazilian Portuguese they are optional in these cases. In Brazilian Portuguese the string *a minha opinião* is ambiguous between the reading corresponding to *my opinion* and *to my opinion*, because of the lexical ambiguity of *a*. The interaction with pro-drop and various word-order possibilities multiplies possible parses as this and similar phrases can be the subject, the direct object, the indirect object or a PP adjunct. But in European Portuguese this string will not be ambiguous between an NP and a PP in contexts where the article is obligatory. In these contexts, only the reading corresponding to *my opinion* will be available.

In general, we can know whether a string is European Portuguese or Brazilian Portuguese if a marked item or construction occurs. Consider a similar example, but where the noun is specific to European Portuguese: for instance *a minha ideia* (*my idea*, the Brazilian Portuguese spelling of the word is *idéia*). If the root node is constrained to have a VARIANT of type *single-variant*, the PP reading is rejected (even when we do not know the specific variant of the input in advance), since the PP analysis is only available in Brazilian Portuguese where the noun is spelled differently. That PP will have a VARIANT of type *portuguese*, which does not unify with *single-variant* in the root node, as was seen before. Figure 5 depicts the corresponding computations.

Variant Detection

With this grammar design it is also possible to use the grammar to detect to which variety the input happens to belong. This is done by parsing that input and placing no constraint on the 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 properties specific to European Portuguese or Brazilian Portuguese respectively; the value *variant* or *single-variant* (depending on the constraint on the root node) indicates that no marked elements were detected and the text can be from both variants.

Also in this case where the language variant of the input is detected by the grammar, the desired variant-"consistent" behavior of the grammar is enforced if the feature VARIANT is set to *single-variant*. The examples in Figure 5 also illustrate this functioning: the constraint on the feature VARIANT of the marked item *ideia* is propagated throughout the syntactic structure.

Evaluation

It is important to gain insight on the quality of the performance of this method. This is addressed in the next sections. The question we want to find an answer

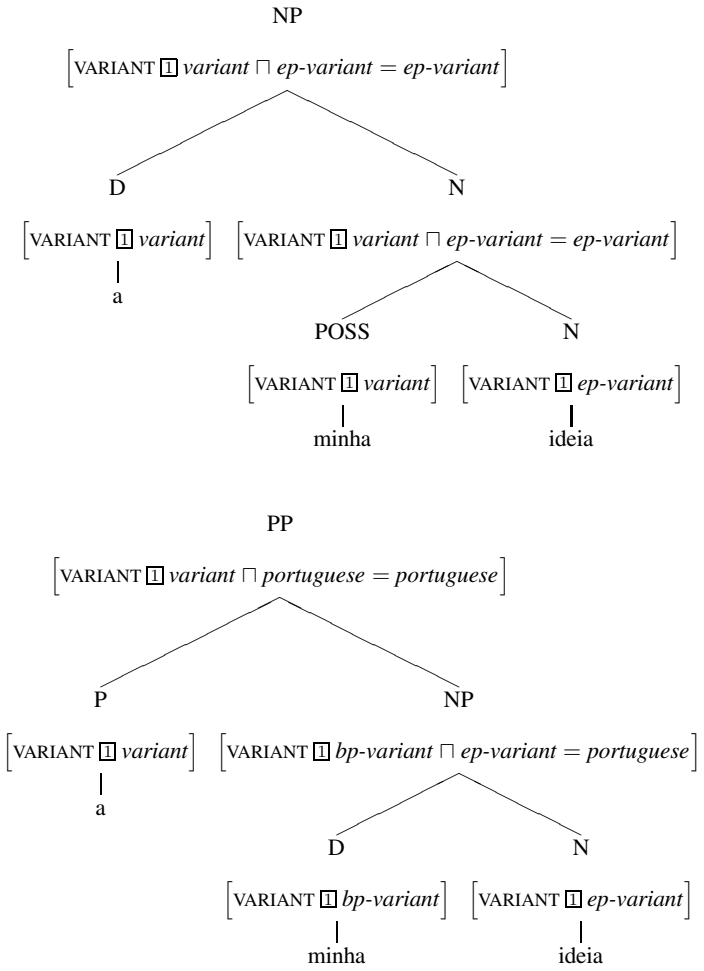

Figure 5: Example of ambiguity specific to one variety. When such ambiguous forms co-occur with items of another variety, they can be resolved by constraining the start symbol with a VARIANT feature of type *single-variant*.

to is: how appropriate is this design for the handling of variation? A simple way to evaluate this design is to parse sentences whose original dialect is known and check whether the grammar can consistently detect the right dialect, by reading off the value of the feature VARIANT in the feature structure for the sentence.

3 Data

To evaluate the approach to accommodate variation presented above, two corpora of newspaper text were used, CETEMPublico (204M tokens) and CETENFolha (32M tokens). The first contains text from a Portuguese newspaper, and the latter from a Brazilian one. These corpora are only minimally annotated (paragraph and sentence boundaries, *inter alia*), but are very large.

Some preprocessing was carried out: XML-like tags, such as the <s> and </s> tags marking sentence boundaries, were removed and each individual sentence was put on a single line. Some heuristics were also employed to remove loose lines (parts of lists, etc.) so that only lines ending in ., ! and ? and containing more than 5 tokens (whitespace delimited) were considered. Other character sequences that were judged irrelevant and potentially misguiding 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 (in an interview, for instance) were removed, since they are frequent and the grammar used is not intended to parse name plus sentence strings.

From each of the two corpora, 90K lines were selected, with the smallest length sentences. Of the resulting 90K+90K, 26% were shown to be fully parsable by the grammar and set apart. From these 26%, $1800 + 1800$ sentences were randomly chosen.

If a sentence is found in the European corpus, one can be sure that it is possible in European Portuguese, but one does not know if it is Brazilian Portuguese, too. The same is true of any sentences in the American corpus — these can also be sentences of European Portuguese in case they only contain lexical items and structures that are common to both variants.

In order to address this, a native speaker of European Portuguese was asked to manually decide from sentences found in the American corpus whether they were markedly Brazilian Portuguese. Conversely, a Brazilian informant detected markedly European Portuguese sentences from the European corpus. Thus a three-way classification is obtained: every sentence was classified as being markedly Brazilian Portuguese, European Portuguese or common to both variants.

As a result, 5KB of text (140 sentences) from each one of the three classes were selected for testing, and another 5KB (also around 140 sentences each) for training (development).

Many more sentences were classified as possible in both dialects than as sentences specific to either one. We only kept a subset of the sentences judged to

be common, in order to have a uniform distribution of the three classes in the data. 16% of the sentences in the European corpus were considered impossible in Brazilian Portuguese, and 21% of the sentences in the American corpus were judged exclusive to Brazilian Portuguese. Overall, 81% of the text was common to both varieties. Since a single marked item or construction in a sentence causes it to be classified as marked, we see that a very large part of the language variants overlap (very likely more than 81%).

4 Differences Between European Portuguese and Brazilian Portuguese Found in the Training Corpora

We proceed to an analysis of the training data resulting from the manual classification described in Section 3. A brief typology of the markedly Brazilian elements found in the American training corpus is presented. We also present the relative frequency of these phenomena based on the same data. We do not present the marked items found in the European corpus, because, being native speakers of European Portuguese, we could not always determine the reason why the Brazilian informant marked sentences as specific to European Portuguese.[2]

0. Differences due to lack of orthographic harmonization (33.3%)
   (a) Phonetic or phonological differences reflected in spelling (9.3%)
   e.g. BP *irônico* vs. EP *irónico* (*ironic*)
   (b) Pure spelling differences, no phonemic difference (24%)
   e.g. BP *ação* vs. EP *acção* (*action*)
1. Lexical differences (26.9% of all differences found)
   (a) Different form, same meaning (22.5%)
   e.g. BP *time* vs. EP *equipa* (*team*)
   (b) Same form, different meaning (4.4%)
   e.g. *policial*: BP *police officer*, EP *criminal novel*

[2] Although we were able to extract a large amount of information from the European Portuguese training data as well, by checking possible candidates in dictionaries and web searches, we cannot quantify the different phenomena at stake precisely, as in some cases a decision could not be made. We should have asked the informants to paraphrase the marked sentences in a way that sounded acceptable to them, so that we could have detected the markedly European items and constructions consistently.

2. Syntactic differences (39.7%)

   (a) Co-occurrence of definite articles and prenominal possessives (12.2%)
   BP: Meu pai cuida de tudo.
   my father takes care of everything
   EP paraphrase: O meu pai cuida de tudo.
   the my father takes care of everything
   My father takes care of everything.

   (b) Different subcategorization frames (9.8%)
   Progressive auxiliary *estar* selects for a gerund in Brazilian Portuguese, and preposition *a* plus infinitive in European Portuguese (5.4%)
   BP: O gravador está funcionando?
   the tape recorder is working.GER
   EP paraphrase: O gravador está a funcionar?
   the tape recorder is PREP work.INF
   Is the tape recorder working?

   (c) Clitic placement (6.4%)
   BP: Tommy se apaixona por Betsy.
   Tommy CLITIC falls in love for Betsy
   EP paraphrase: Tommy apaixona-se por Betsy.
   Tommy falls in love CLITIC for Betsy
   Tommy falls in love with Betsy.

   (d) Bare NPs headed by singular count nouns (5.4%)
   BP: Médico também é ser humano.
   doctor also is being human
   EP paraphrase: Um médico também é um ser humano.
   a doctor also is a being human
   A doctor is a human being, too.

   (e) Different subcategorization frame and different word sense (1.9%) e.g. BP *fato* (*fact*, with a sentential complement) vs. EP *fato* (*suit*, no complements)

   (f) Co-occurrence of prenominal *todo* and definite articles (0.9%)
   BP: Todo mundo aqui gosta deles.
   all world here likes of them
   EP paraphrase: Todo o mundo aqui gosta deles.
   all the world here likes of them
   Everyone here likes them.

   (g) Contractions of prepositions and articles (0.9%)
   BP: Eles estão em uma creche da cidade.
   they are in a kindergarten of the city
   EP paraphrase: Eles estão numa creche da cidade.
   they are in a kindergarten of the city.
   They are in one of the city's kindergartens.

(h) Matrix wh-questions without subject-verb inversion or *é que* (0.9%)
BP: O que ele veio fazer aqui?
what he came to do here?
EP paraphrase: O que é que ele veio fazer aqui?
what is that he came to do here
What did he come here for?

(i) Postverbal sentential negation (0.5%)
BP: Mas, felizmente, isso não existe não, bonitinha .
but fortunately that not exists not foxy
EP paraphrase: Mas, felizmente, isso não existe, bonitinha .
But fortunately that not exists foxy
But fortunately that doesn't exist, foxy.

(j) other (0.5%)
BP: Enquanto isso, de dia, trabalhava de alfaiate.
while that at day I worked as tailor
EP paraphrase: Enquanto isso, de dia, trabalhava como alfaiate.
while that at day I worked as tailor
Meanwhile, I worked as a tailor during the day.

Figure 6 presents a pie chart of these differences.

One third of the differences found would be avoided if the orthographies were harmonized (0). Differences that are reflected in spelling can be modeled by the grammar via multiple lexical entries, with constraints on the feature VARIANT reflecting the variety in which the lexical item with that spelling is used. In some cases, a different solution would be preferable. When the difference is systematic (e.g. the European Portuguese sequence *ón* always corresponds to a Brazilian Portuguese sequence *ôn*, with an example in (0a)), it would be better to have a lexical rule that affects only spelling and the VARIANT feature producing one variant from the other.[3]

Orthographic differences, which account for 33.3% of all differences appear in 47.9% of the sentences (in the American training corpus). This means that, by simply looking at lexical items, almost 50% recall could be obtained on these data, assuming perfect lexical coverage.

Some differences cannot be detected by the grammar. This is the case of (1b), which would require word sense disambiguation. When word sense differences are accompanied by different syntax, they can be detected by the grammar (2e) in limited circumstances (e.g., in that example, the difference is detected only if the complement is expressed). This places the upper bound for recall for Brazilian Portuguese between 95.6% and 93.7%, judging by these frequencies.

Interestingly, 40% of the differences are syntactic. These cases are not expected to be difficult to detect by a grammar, but it may be difficult to take advantage of them with shallower methods. Consider the example of clitic placement, illustrated

[3]This was not implemented, because string manipulation is limited in the LKB.

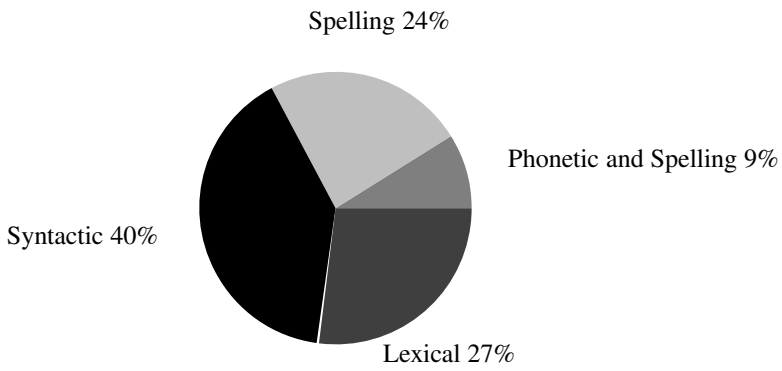

Figure 6: Breakdown by type of the differences detected in the Brazilian Portuguese training corpus.

in (2c). It is not a simple matter of clitics preceding the verb in Brazilian Portuguese and following it in European Portuguese, because they can also precede the verb in European Portuguese depending on the syntactic context (e.g. in finite subordinate clauses they must do so). Therefore, syntactic information is crucial to detect some of the differences found.

Another interesting example is the co-occurrence of definite articles and possessives (2a). Recall from one of the examples in Section 2 that the feminine singular form of the definite article, *a*, is homonymous with a preposition. Syntactic context can disambiguate this situation in several circumstances (e.g. after a preposition that does not introduce an infinitival clause it can only be an article; as an article it cannot introduce an NP headed by a noun that is masculine or plural, etc.).

5 Grammar Preparation

The evaluation experiments were carried out with a computational HPSG for Portuguese developed with the LKB platform (Copestake, 2002) that uses MRS for semantic representation (Copestake et al., 2001). At the time of the experiments reported here, this grammar was of modest size. In terms of linguistic phenomena, it covered basic declarative sentences and basic phrase structure of all categories, with a fully detailed account of the structure of NPs. It contained 42 syntax rules, 37 lexical rules (mostly inflectional) and a total of 2988 types, with 417 types for lexical entries. There were 2630 hand-built lexical entries, mostly nouns, with 1000 entries. It was coupled with a POS tagger for Portuguese, with 97% accuracy (Branco and Silva, 2004; Silva, 2007).

In terms of the sources of variant specificity, this grammar was specifically designed to handle the co-occurrence of prenominal possessives and determiners and most of the syntactic constructions related to clitic-verb order. As revealed by the study of the training corpus, these constructions underlie almost 20% of marked

sentences, and they are the bulk of the syntactic differences.

We present a simplified description of how word-order of complement clitics was controlled by the grammar at the time of the experiments. Basically, several binary versions of Head-Complement rules are used. In the feature structure for these rules there is a boolean feature PROCLISIS indicating whether proclisis (clitics before the verb) or enclisis/mesoclisis (clitics after or in the middle of the verb) is expected according to European Portuguese.[4] The value for this feature is determined by other elements in a sentence. An example: since in finite subordinate clauses proclisis is enforced, complementizers select for a complement with a PROCLISIS feature with the value $+$ (the start symbol is constrained with the value $-$ for this feature, because the unmarked order in matrix clauses is enclisis). There is a Head-Complement construction that ignores this feature and projects a non clitic complement.

The nature of clitics is represented by a feature WEIGHT: clitics have the value *clitic* for this feature, other syntactic constituents have the value *non-clitic*, and there is no unifier for these two types. The value of WEIGHT is lexically specified and always *non-clitic* for phrases.[5] The Head-Complement schema that projects non-clitics has constraints like:

$$\left[\begin{array}{l} \text{HEAD-DTR } \boxed{1} \\ \text{NON-HEAD-DTR } \boxed{2}\left[\text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \mid \text{WEIGHT } \textit{non-clitic}\right] \\ \text{ARGS } \left\langle \boxed{1}, \boxed{2} \right\rangle \end{array}\right]$$

The feature ARGS has as its value the list of daughters of a syntactic rule. The order of the elements in this list correlates with word order. The actual value of ARGS is determined by general types in the Matrix (*head-initial* and *head-final*), from which specific syntactic rules inherit, but we present the constraints on ARGS here instead of the relevant supertypes, in order for the word-order patterns in these rules to be visible.

There is a Head-Complement rule that projects a clitic to the left of the verb in proclisis contexts:

[4]The choice between enclisis and mesoclisis depends only on verbal tense and mood and is not relevant for our purposes. The opposition is between proclisis contexts and non proclisis contexts.

[5]The feature WEIGHT is reminiscent of the same feature in Abeillé and Godard (2003), but here different values are used. An equivalent treatment would be to enrich the type hierarchy under *synsem*, so that the distinction between clitics and non clitics is represented via subtypes of *synsem*, as in Miller and Sag (1997). Contrary to much HPSG work on Romance clitics, we chose to have them combine with verbs in syntax rather than in morphology for practical reasons that relate to orthography: the resulting string includes a space whenever the clitic precedes the verb. When clitics follow the verb, a hyphen is used instead, which is removed in a preprocessing step.

$$\left[\begin{array}{l}\text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \left[\begin{array}{l}\text{HEAD } \textit{verb} \\ \text{PROCLISIS } + \end{array}\right] \\ \text{HEAD-DTR } \boxed{1} \\ \text{NON-HEAD-DTR } \boxed{2} \left[\text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \mid \text{WEIGHT } \textit{clitic}\right] \\ \text{ARGS} \left\langle \boxed{2}, \boxed{1} \right\rangle \end{array}\right]$$

In order to account for variation with respect to clitic placement, there are thus two versions of Head-Complement rules for clitics in enclisis contexts that are marked with respect to the VARIANT feature and resort to the feature PROCLISIS:

$$\left[\begin{array}{l}\text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \left[\begin{array}{l}\text{HEAD } \textit{verb} \\ \text{PROCLISIS } - \end{array}\right] \\ \text{HEAD-DTR } \boxed{1} \\ \text{NON-HEAD-DTR } \boxed{2} \left[\text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \mid \text{WEIGHT } \textit{clitic}\right] \\ \text{ARGS} \left\langle \boxed{1}, \boxed{2} \right\rangle \\ \text{VARIANT } \textit{ep-variant} \end{array}\right]$$

$$\left[\begin{array}{l}\text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \left[\begin{array}{l}\text{HEAD } \textit{verb} \\ \text{PROCLISIS } - \end{array}\right] \\ \text{HEAD-DTR } \boxed{1} \\ \text{NON-HEAD-DTR } \boxed{2} \left[\text{SYNSEM} \mid \text{LOCAL} \mid \text{CAT} \mid \text{WEIGHT } \textit{clitic}\right] \\ \text{ARGS} \left\langle \boxed{2}, \boxed{1} \right\rangle \\ \text{VARIANT } \textit{bp-variant} \end{array}\right]$$

Turning now to the issue of prenominal possessives, in order to parse items that are not preceded by articles in Brazilian Portuguese, we just added determiner versions of possessives that have a marked VARIANT feature, with the value *bp-variant* (see Figure 5 above).

Finally, the lexicon contained lexical items specific to European Portuguese and specific to Brazilian Portuguese. They were taken from the Portuguese Wiktionary (http://pt.wiktionary.org), where this information is available. Namely, the Portuguese Wiktionary contains the categories "Portuguese spelling" ("grafia portuguesa") and "Brazilian spelling" ("grafia brasileira"), associated with items with specific spellings, and it is possible to list all the items in these categories. Leaving aside items judged to be very infrequent (e.g. *aniónico* / *aniônico* — *anionic*), around 740 marked lexical items were coded. Lexical items that are variant specific that were found in the training corpora (80 more) were also entered in the lexicon.

Known class	*Predicted class*			
	EP	**BP**	**Common**	Recall
EP	53	1	86	0.38
BP	6	61	73	0.44
Common	14	1	125	0.89
Precision	0.73	0.97	0.44	

Table 1: Confusion matrix for variant detection.

6 Results

The results obtained are presented in Table 1. When the grammar produced multiple analyses for a given sentence, that sentence was classified as markedly European, respectively Brazilian, Portuguese if all the parses produced VARIANT with type *ep-variant*, respectively *bp-variant*. In all other cases, the sentence would be classified as common to both. Every sentence in the test data was classified, and the figure of 0.57 was obtained as overall precision and recall.

The results in Table 1 concern the test corpus, of which all sentences are parsable. Hence, actual recall over a naturally occurring text is expected to be lower, given the development status of the grammar used in the experiment. Using the estimate that only 26% of the input sentences receive a parse by the grammar that was employed in these experiments (see Section 3), the actual figure for recall would lie near 0.15 (= 0.57 x 0.26).

Good recall was achieved for Common (89%), which means that the system erroneously commits to one of the variants only 11% of the time.

In contrast, recall for European Portuguese and Brazilian Portuguese was very low (38% and 44% respectively). What has the most negative impact on the recall values for European Portuguese and Brazilian Portuguese is a very high number of European Portuguese and Brazilian Portuguese test items being classified as "Common" (61% of all European Portuguese test sentences and 52% of all Brazilian Portuguese test sentences), because no marked item or construction was found. We believe that this is a consequence of a lack of lexical coverage (see Section 7) of items that are specific to one of the dialects and may also be a consequence of using only two syntactic cues (regarding clitics and possessives). Therefore, improving lexical coverage and taking advantage of more syntactic differences between the two variants should improve recall in this respect. These errors are also responsible for the low precision for the Common class (44%).

Very good precision was obtained for Brazilian Portuguese (97%): the cues used to classify a sentence as Brazilian Portuguese thus seem to be very robust (proclisis in contexts where European Portuguese shows enclisis, absence of definite articles preceding prenominal possessives, marked lexical items).

Precision for European Portuguese was lower (73%). As can be seen from Table 1, most of these errors originate from the system classifying as European Portuguese sentences that the gold standard says are common to both variants.

This situation arises because enclisis is correlated with European Portuguese by the grammar, but this correlation is not very strong in the test sentences (more about this in Section 7).

7 Error Analysis

Limited lexical coverage is responsible for a large proportion of errors: at least 40% of the cases of sentences incorrectly classified were due to lexical items specific to one of the two variants that were not in the lexicon. We used a POS-tagger to guess the category of unknown words, so problems of lexical coverage often did not have an impact on parse coverage. However, the POS-tagger cannot guess whether a word is specific to Brazilian Portuguese or European Portuguese, so these items were underspecified with respect to their VARIANT feature.

Many of these missing lexical items are interesting or challenging. Some involve derivation. The adverbs *tranqüilamente* (Brazilian Portuguese) and *tranquilamente* (European Portuguese) — *calmly* — were not in the lexicon, although their adjectival bases were (Brazilian *tranqüilo*, European *tranquilo* — *calm*). In some cases the morphological process involved seems less productive: Brazilian *gringolândia* (*a place filled with foreigners*) from Brazilian *gringo* (*foreigner*). There is also a case of a noun derived from an acronym, with the acronym showing up in the derived form with a phonetic spelling: *peemedebista* (a member of the Brazilian political party PMDB). Some other missing lexical entries involve multi-word expressions or idioms: European *de jeito* (*of acceptable quality*, literally *of skill*); European *a cores* vs. Brazilian *em cores* (*in color*, using different prepositions).

In some cases the differences are difficult to detect via dictionaries, as they involve only grammatical features. One example is the noun *ioga* (*yoga*), which is feminine in Brazilian Portuguese and masculine in European Portuguese. Also, some differences in spelling only show up in inflected forms (not in the lemma): European *europeia(s)* vs. Brazilian *européia(s)* — *European*, feminine singular (plural), the lemma being *europeu* in both dialects.

It is worth noting that 20 sentences (14 with the class Common and 6 with the class Brazilian Portuguese) were misclassified by the grammar as European Portuguese. 70% of these errors (11/14 for the Common class and 3/6 for the Brazilian Portuguese class) are due to clitic placement according to European syntax. The point here is that clitic placement according to European syntax appears in Brazilian newspaper text as well. In fact, three sentences in the Brazilian Portuguese class presented enclisis (and also characteristics specific to Brazilian Portuguese) and were misclassified as European Portuguese by the grammar for this reason and because the Brazilian Portuguese characteristics were not detected. 11 sentences in the Common class also presented enclisis, and were misclassified by the grammar as European Portuguese because of this. Some of these sentences came from the American corpus, and some from the European one. The justification we find for

enclisis appearing in the Common class (in sentences from the European corpus) is that, since enclisis is possible in Brazilian newspaper text, it is not considered markedly European when it is seen in European newspaper text, so the Brazilian informants did not classify sentences with enclisis as markedly European. This means that there is some interference of genre in these results. While proclisis in contexts where enclisis is expected in European Portuguese is a so good indicator of Brazilian Portuguese text, enclisis in European enclisis contexts is not a good indicator of European Portuguese, as it can also be found in Brazilian Portuguese text.

The remaining sentences misclassified as European Portuguese are due to misspellings in Brazilian text that unexpectedly conform to European orthography. In Brazil a diaeresis is used on *u* (*ü*) when it follows *q* or *g*, precedes *e* or *i* and is pronounced. The errors were due to spellings like *aguentar* (*to bear*) and *tranquilo* (*calm*), instead of *agüentar* and *tranqüilo*.

A very small number of errors ($<1\%$) was due to the lack of case sensitivity in the LKB (month names are capitalized in European Portuguese and not capitalized in Brazilian Portuguese) and word sense differences.

8 Related Work

There is a considerable amount of literature on grammar specialization and grammar porting (Kim et al., 2003).

With the architecture presented here, it is still possible to specialize a grammar to one of the dialects. In fact this can be done automatically by traversing the source files with the lexical entries and the syntactic/morphological rules and eliminating those that are marked to be specific to all but the desired dialect. This can be done for efficiency reasons. If one wants to parse or generate in a specific variant and this elimination is not performed, the constructions and lexical items specific to all others will only be ruled out when the root node is reached. Therefore, it can be much more efficient to eliminate them in the source files altogether. On the other hand, our experiments showed a large amount of overlap between the two dialects under consideration, so we expect that items that are specific to only one of them should not be frequent in practice. Therefore, the added cost of considering both dialects at run time may not be too detrimental as far as efficiency is concerned, but we have not measured the impact of this.

Søgaard and Haugereid (2005) present a proposal similar to ours. They seek to model variation within Scandinavian languages, by resorting to a LANGUAGE feature. Stymne (2006) goes even farther and uses a LANG feature in a grammar for two rather different languages: English and Swedish.

9 Conclusions

In this paper we presented an architecture to model language variation with typed-feature formalisms. The design that was proposed here can allow for parameterization of a grammar to parse or generate only in a given dialect, or parse input consistently only in one dialect even when the language variant of the input is unknown beforehand. At the same time, consistency of analysis can be enforced, and ambiguity controlled. Moreover, this approach also allows the grammar to function as a dialect classifier, as it can be used to detect the language variant at stake.

We proceeded to evaluate this design, using a grammar for Portuguese that accommodates both European Portuguese and Brazilian Portuguese. Our results are promising, and the grammar achieved very high precision in some cases (97% precision when classifying the input as belonging to Brazilian Portuguese). When the grammar classified the input as European Portuguese, it was right 73% of the time, which is another encouraging result. 89% of the sentences that displayed no dialectal characteristics were also correctly classified as common to European Portuguese and Brazilian Portuguese.

In other cases, the results can be improved. Many European Portuguese characteristics were not recognized (resulting in 38% recall for European Portuguese), and neither were several Brazilian Portuguese characteristics (with 44% recall for Brazilian Portuguese). This means that large improvements can be obtained by extending the grammar with more dialect specific lexical items and constructions. In addition, from the several sources of variant specificity, the grammar used here was prepared to cope only with grammatical constructs that are responsible for at most 20% of them. Also the lexicon, that included a little more than 800 variant-distinctive items, can be largely improved.

There are some interesting challenges, too. We came across the classical problems of lexical coverage, like multi-word expressions and new words.

Some differences between variants are not absolute in practical scenarios. An example of this that affected our results is the spelling oscillations between *u* and *ü* after *q* and *g* in Brazilian Portuguese.

Also, textual genre seemed to affect the results, as Brazilian newspaper text presents some syntactic properties of European Portuguese, like clitic word order.

Besides, there are problems beyond a grammar's capacity, like word sense distinctions. Although word sense differences were frequent in the training data (present in 6.3% of all marked Brazilian Portuguese lexical items found), they turned out to be negligible in the errors found in the test data.

These are issues over which more acute insight will be gained in future work, which will seek to improve the contributions put forth in the present paper.

Given the 97% precision achieved for the Brazilian Portuguese class (with a somewhat lower precision for the European Portuguese class, of 73%), we think that our results are the proof-of-concept that an informed approach can produce very good results in this task, using the architecture we presented.

Summing up, a major contribution of the present paper is a design strategy

for type-feature grammars that allows them to be appropriately set to the specific variant of a given input. Concomitantly, this design allows the grammars to identify the variety used in the input.

References

Abeillé, Anne and Godard, Danièle. 2003. The Syntactic Flexibility of French Degree Adverbs. In Stefan Müller (ed.), *Proceedings of the HPSG-2003 Conference, Michigan State University, East Lansing*, pages 26–46, Stanford: CSLI Publications.

Bender, Emily M., Flickinger, Dan and Oepen, Stephan. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In John Carroll, Nelleke Oostdijk and Richard Sutcliffe (eds.), *Procedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Branco, António and Silva, João. 2004. Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, pages 507–510, Paris: ELRA.

Copestake, Ann. 2002. *Implementing Typed Feature Structure Grammars*. Stanford, California: CSLI Publications.

Copestake, Ann, Flickinger, Dan, Pollard, Carl and Sag, Ivan A. 2001. Minimal Recursion Semantics: An Introduction. *Language and Computation* 3, 1–47.

Kim, Roger, Dalrymple, Mary, Kaplan, Ron, King, Tracy Holloway, Masuichi, Hiroshi and Ohkuma, Tomoko. 2003. In Emily Bender, Dan Flickinger, Frederik Fouvry and Melanie Siegel (eds.), *Proceedings of ESSLLI 2003 Workshop on Ideas and Strategies for Multilingual Grammar Development*, pages 49–56, Vienna, Austria.

Miller, Phillip H. and Sag, Ivan A. 1997. French Clitic Movement without Clitics or Movement. *Natural Language and Linguistic Theory* 15(3), 573–639.

Pollard, Carl and Sag, Ivan. 1994. *Head-Driven Phrase Structure Grammar*. Chicago University Press and CSLI Publications.

Silva, João Ricardo. 2007. *Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization*. MSc Dissertation, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal.

Søgaard, Anders and Haugereid, Petter. 2005. Implementing Dialectal Variation in Typed Feature Structure Grammars. Unpublished Manuscript.

Stymne, Sara. 2006. *Swedish-English Verb Frame Divergences in a Bilingual Head-driven Phrase Structure Grammar for Machine Translation*. MSc Dissertation, Linköpings Universitet.