March 29, 1998

Research Project
Rhythmic patterns, parameter setting and language change

I. Goals and basic assumptions of the project

The basic aim of this project is to elucidate a long-standing linguistic question: How does language change proceed in time? The view of language that has developed in modern linguistics is that the set of possible grammars for human languages is finite and that each grammar in the set is characterized by the values it takes for each of a finite set parameters, each of which takes one of a small number of values. From this it would seem to follow that language change should occur as a jump from one set of parameter values to another, and hence be abrupt and discrete. Given the nature of language acquisition and the known resistance to change of the language of adults, one might further expect that such abrupt change would take place from one generation to the next.

Yet the brute facts of language change are otherwise. Over and over again, it has been observed that language change, as documented in historical texts, is gradual and often proceeds over several generations of speakers. This has often been thought to contradict the theoretical paradigm of modern linguistics. Work by Kroch (1989), however, has shown that the contradiction is only apparent. Indeed, the gradualness of change as seen in texts is compatible with a categorical theory of grammar. This is so because texts reflect mixed or bilingual usage, in which authors choose stochastically between the forms output by conservative and innovative grammars. This way of reconciling the facts of language change with linguistic theory, however, predicts that in the absence of interfering sociolinguistic effects, abrupt change should be observable.

The current proposal will attempt to establish the empirical case for abrupt grammatical change in the case of the history of the pronominal clitic system of Portuguese, which is a clear and well-documented case of what must be a discrete parametric change. In the grammar of Classical Portuguese (ClP), in root affirmative sentences clitic pronouns could appear either immediately before the verb (proclitic) or immediately after it (enclitic). Yet, in the grammar of modern European Portuguese (EP), such clitics may only surface as enclitics. There have been a number of linguistic analyses proposed to account for the distinction between these two grammars. Though they differ slightly in what they take to be the crucial difference between these two stages in the development of the language, these analyses share the property of reducing the difference to a single parametric difference between the grammatical systems.

In the case of the ClP clitic system, a further factor enters the discussion. The choice among relative orderings of clitic and verb appears to be sensitive to prosodic preferences; that is, to preferences in usage for certain patterns of alternation of stressed and unstressed syllables over other patterns. In consequence, Galves and Galves (1995) conclude that the gradual change in clitic syntax observed in texts reflects shifting prosodic preferences, leading eventually to a catastrophic reanalysis. It seems likely that the statistical patterns observed in texts are a mixture of variation due to bilingualism and variation due to prosodic preferences in usage. Our goal is to develop and test a model of prosody and of change that will allow us to distinguish these effects and to demonstrate that controlling for these factors allows the abruptness of true grammar change to become visible.

We propose to advance this research by focusing on the following four questions:

1. Can we show empirically that, in actual speech, usage frequencies are influenced by speakers' preferences for certain rhythmic patterns?

2. Can such preferences be detected in written texts through statistical analysis?

3. How do such preferences change over time?

4. What is the relationship between the time course of change in prosodic preferences and grammatical change?

In order to study the impact of distinct rhythmic patterns on written texts, we will focus on the difference between EP and Brazilian Portuguese (henceforth BP). We will make the working hypothesis that the rhythmic pattern of BP is, in the relevant respects, the same as that of ClP. Once the first two questions can be answered in the affirmative, we will be in a position to address the issue of how prosodic preferences have changed through the historical development of EP. Then we can study the interaction between changes in prosodic preferences and changes in the grammar. We suspect that prosodic preferences are not subject to the types of sociolinguistic pressures that contaminate syntactic change. Thus, we expect to be able to study precisely the relationship between the onset of the syntactic change and the time course of the prosodic change. In previous work, Galves and Galves (1995) have come to the conclusion that prosodic changes in ClP are causally related to the syntactic change. We aim to verify this causal relation, using our empirical studies to inform the development of a model of language acquisition in which the learner is guided by hints of a prosodic nature (Morgan 1986, Scarpa 1995). We make crucial use the Thermodynamical Formalism (Ruelle 1978) as a framework in which the notion of rhythmic pattern can be properly defined and articulated with a grammar in the definition of the probability measure governing the choice of the sample of positive evidence offered to a learning child.

On the practical side, we propose to create an annotated corpus of Classical Portuguese. Such a corpus will be used as a fundamental tool to the investigations described above. Furthermore, we plan to make the corpus available via www. This brings the following advantages:

1. Any other researcher in the world can look into the data used as a basis for our conclusions; in this way we hope to increase the reliability of our results by making our experiments reproducible.

2. With the creation of such a corpus, we hope our project will have long lasting effects, for we will be providing readily available data for future research by the international community.

II. Theoretical framework

The conception of language underlying this study is the one advocated by Chomsky in all his work, currently represented in the Model of Principles and Parameters (Chomsky 1985, and 1995 for the "Minimalist version" of this model). In this view, one must distinguish the notions of E(xternal)-Language and I(nternal)-Language. According to Chomsky the object of Linguistic Theory is I-language, *i.e.* the knowledge attained in the natural acquisition process and which allows any human being to produce and interpret sentences in his language.

In the Principle and Parameters model, specific grammars are obtained by fixing values of a finite number of parameters left unspecified by the Universal Grammar (UG). Therefore, from this point of view, two grammars or I-languages are different if they have at least one parameter fixed in a different way. Each parameter is responsible for a set of grammatical phenomena. In this framework, a grammatical change is the change in the value of at least one parameter, and we expect that this entail superficial changes in the set of syntactic phenomena dependent on the value assigned to this parameter.

Parameters values are set during the acquisition process and since each of them is responsible for great portions of the grammar, their change is too catastrophic to be likely to happen during the lifetime of speakers. This is why grammatical changes can be argued to happen only in the acquisition process and correspond to the fixation of some parametric value in a way different from the preceding generation. Why this happens from time to time is one of the main questions of this study. In this sense, this historical question can be reformulated, in this framework, as an acquisition question.

As for E-language, it is defined by Chomsky (1985) as the set of actual utterances which are produced by a given I-language. But crucially, E-language is not the direct reflex of I-language since it is also also the result of performance aspects involved in the use of language. This distinction is particularly crucial for the studies in Historical Linguistics. In effect, the texts which are the object of our research are pieces of E-language, and reflect only indirectly the grammar which underlies them. This is particularly sensible in written texts, given their conservative character. Though our object is to understand what are the grammars (I-languages) involved in the history of a given language, and when and how these grammars change from one to another, we know that what we see is the picture of the evolution of a set of utterances belonging to E-language. This calls for a reflexion on the relationship between E-languages and I-languages. This project is concerned by two ways of understanding this relationship.

Kroch (1994, 1995 a) gives us a model of the relation of E-language and I-language in the historical changes as they appear in texts. These changes do not generally appear abruptly but gradually, as periods of variation between constructions evolving in time. According to him, what we see then is the reflex of *grammars in competition*. One grammar is the old one which has already been replaced in the speakers' mind but continues to be assigned more prestige by the community, and the other is the new grammar which imposes itself gradually, but certainly more slowly in written texts than in speech. It is important to emphasize that, from this point of view, when the variation begins to appear in texts, the grammar has already changed *i.e.* some parametric value has already be reset by a generation of children. This crucially differentiates this approach from others (cf. for instance Lightfoot 1991) which consider that the grammatical change occurs only when the variation reaches its end.

Kroch (1989, 1994) proposes two important statistical tools to study the time-course of language change as it appears in the evolutive processes of variation. One is the modelling of replacement of a grammar by another by the S-curve. If a process of variation does not follow this curve, we can say that it does not correspond to a change in progress, i.e. a competition between two grammars, but to a variation allowed by one grammar.

The other is the constant rate hypothesis, by which all the superficial changes corresponding to the same underlying grammatical change evolve at the same rate. This entails that if two variation processes in progress at the same period do not follow the constant rate hypothesis, this will show us that they do not correspond to the same grammatical change, or that one of them is not a change but, again, simply the effect of a variation produced by the same grammar.

The second aspect of the articulation between I-language and E-language brings us back to the pair competence/performance and concerns the relationship between prosody and grammatical acquisition and change. As was said above, written texts are not the direct reflex of I-languages, because they are also the result of performance aspects. We've already emphasized the role of normative considerations. But, in this study, we make the crucial claim that there is another aspect of performance which drives the choice of the utterances, in written as in as well as in oral texts. It is the rhythm associated to the expressions of the language at a given time. Differently from grammar, rhythmic patterns can change during the lifetime of speakers. Furthermore, they are not likely to be submitted to the norm.

In the acquisition process, children have no access to their parents' grammar, but only to the utterances generated by this grammar. Researchers agree on the fact that the data to which children are exposed are incomplete and come up mixed with a lot of noise.

Moreover, they do not seem to pay attention to corrections, which means that they have no access to negative data (cf among others, Marcus 1993). In generative grammar, the fact that children are able to acquire a language with such a poor stimulus is a crucial argument for the innateness of the Language Faculty, or Universal grammar. This gives us an explanation for the possibility of acquisition. But the question of how children come to fix the parameters in such or such way on the basis of incomplete and noisy data continues to be to a large extent unexplained. The fact that entire populations converge to the same grammar tells us that there must be robust "triggers" for a given parameter setting.

Grammars produce bracketed strings of words, not only sequences of morphological expressions. However the sample of positive evidence presented to the child is mainly constituted of ordered strings of words, with only indirect evidence about the structure of sentences. We refer to Morgan (1986) who extensively argues for the claim that the child must use her/his knowledge of the rhythmic pattern of the parental language, which, by hypothesis, is acquired first, to make guesses about the structures of the sentences s/he receives and subsequently identify the grammar which produced them.

The role of rhythm - which is part of performance - in grammatical change was already raised by Marianne Adams in her thesis about the loss of V2 in French (cf. Adams 1987). But she had no tools, at this point, to offer a model of interaction between a phenomenon of performance and competence.

To formally develop this point, we need a precise model of grammar selection in which prosody plays a role in the identification of the grammar. Using the Thermodynamical Formalism, prosody can be defined through a thermodynamical potential. With this potential we define the probability measure which governs, in a given discursive context, the choice among the sentences produced by the grammar. In technical terms this probability measure will be a *Gibbs state* with the potential given by prosody. Roughly speaking, the grammar provides a set of possible sentences, and given a discursive context, each one is assigned a probability by the prosody.

This approach addresses the question of the interaction between competence and performance in the Theory of Grammar. In the present state of the theory of generative grammar, the Phonological Form *PF* is an output of the computational system which builds the syntactic expressions. In the Minimalist version of the theory, the levels of representation *PF* and *LF* (*Logical Form*, the representation of the part of the meaning which depends on grammar) are instructions provided by the grammar for the Articulatory-Perceptual (A-P) and Conceptual-Intentional (C-I) systems respectively. The subsystem which computes *PF* is the place where the rules of the metrical theory apply to construct a representation of the accentual structure of the clause (Halle and Vergnaud, 1987). The analysis of the position of the main stress in Cinque (1993) and of the focus accent in Zubizarreta (1994) are proposals to describe some syntactically-based features of this representation. At the next step, these phonological instructions are interpreted by the Articulatory- Perceptual system. Our hypothesis is that the way the A-P system performs this task depends on the rhythmic pattern. By expliciting the role of the rhythmic pattern in the actual realization of utterances, we'll make a progress in our understanding of the interplay between grammar and performance, and bring elements of answer to the third question adressed by Chomsky (1985): "How is knowledge of language put to use?". No advance in the understanding of language acquisition and change can be attained without adressing this question.

III. The steps of the project

1. The history of clitic-placement from the 16th c. to the 19th c.

The history of clitic-placement will be studied in correlation with the position of the verb in the sentence. These two syntactic features of languages are strongly related in the generative approach since the position of the clitic is argued to depend on the position of the verb in the clause (Kayne 1991). From an historical point of view, it was already argued by Salvi (1990) that the change in clitic placement is correlated with the loss of the *V2* properties of ClP (cf also Torres Moraes 1995, Galves 1996a.). Though presenting many features of V2 languages (cf. Ribeiro 1995b for the V2 characteristics of Old Portuguese), Portuguese, in all its history, is a rather atypical V2 language, since it displays a great proportion of V1 sentences. Its syntax of clitic-placement is atypical as well when compared with what happens in other Romance languages. It is likely that these two central aspects of its syntax are deeply correlated. A rigorous study of the correlation between clitic placement and the syntax of the verb, in main as well as in subordinate clauses was not done yet, and it is one of the aims of this project.

Clitic placement is one of the more salient and intriguing fact of the history of portuguese since at the two ends of the documented story of European Portuguese, enclisis is strongly dominant in V2 contexts (cf. Galves 1996b. in *Resultados de auxílios anteriores*, figura 1). In the middle, *i.e*, during the 16th century, it is the reverse: proclisis is almost categorical in V2 contexts (cf Lobo 1992). The period which interests us corresponds to the second part of the story, when enclisis becomes an option again. In this period, which concerns authors born from the second half of the 16th century to the first half of the 19th century, we observe a variation between enclisis and proclisis, which varies not only from author to author but also, in the same author's work, from gender to gender. This is the case with Francisco Mello and Pe Antonio Vieira, born the same year (1608).

According to Martins (1994), EP is already instantiated in Vieira. Her hypothesis is based on her analysis of Vieira's sermons, in which she finds 70% of enclisis in V2 constructions. In Mello, she finds only 2% of enclisis. She concludes that Vieira represents the innovative grammar, and Mello the old one. But a pilot study performed on Vieira's letters, in comparison with Mello's letters, showed that in this kind of texts, they have the same distribution of enclisis and proclisis (cf. *Resultados de auxílios anteriores*).

Furthermore, Table 2 of Galves 1996b. (cf *Resultados de auxílios anteriores*) shows that authors born in the first half of the 18th century are predominantly proclitic.

According to Kroch's hypothesis, the variation observed between the 17th and the 18th century between proclisis and enclisis could be the reflex a "bilingual" situation. Only a rigorous statistical analysis can give us a precise answer to this question since, if the variation is produced by a competition of grammars, the evolution of this variation in time is expected to correspond to a S-shaped curve. According to the previous studies on the 18th c. (Torres Moraes 1995), this is not likely to be the case, but a more systematic and accurate statistical analysis of the data will provide a more precise picture of what goes on in this period, which was rather little-studied until now.

In order to do this, we propose a systematic work of longitudinal description of clitic-placement involving four generations of authors per century, with two authors at least per generation, and, if possible, two different kinds of written work per author (see *Corpus Anotado "Tycho Brahe" - Organização Preliminar dos Dados* in *Cronograma*).

The use of the statistical methods mentionned above must allow us to draw a picture of the syntactic evolution of portuguese in co-relation with hypotheses about the grammatical changes undergone by the language. As a first step of this part of the project, we shall have to construct the corpus, which means taking decisions at both a qualitative level, choosing the texts in such a way that we obtain a representative sample of the variation at each point we consider in time, and a quantitative level, defining the amount of data we need to obtain reliable results.

Our working hypotheses about the grammars involved in the history of this period are the following (cf Galves and Galves 1995, Torres Moraes 1995, Galves 1996a.) :

- ClP allows the variation between proclisis and enclisis in V2 constructions. The former corresponds to a structure in which the phrase anteceding the verb, and the verb itself are in a specifier/head relation. Given the V2 caracteristics of ClP, we'll assume that the category hosting the verb is Comp . The structure corresponding to proclisis constructions can be thus represented as follows:

(1) $[_{CP}$ XP cl-V [....]]

As for enclisis, this grammar assigns it a structure which is similar to V1 constructions, which the verb is outside the boundaries of the clause, as in the following representation:

(2) XP [$_{CP}$ V-cl [....]]

According to this hypothesis, the increasing of enclisis can be initially explained by the increasing of the choice for structure (2) in detriment of the choice of structure (1). The reasons that lead authors to make this choice can be purely stylistic. But it will be argued that, at some point of the history of the language, they are essentially prosodic.

- Following Galves and Galves (1995), and Galves (1996a and b.), we'll work with the hypothesis that the change to EP involves a reanalysis of the enclitic structures in which the preverbal phrase is the subject. In EP, enclitic constructions with preverbal subjects have the following structure:

(3) [$_{CP}$ NP V-cl [....]]

This means that not only proclisis disappears from the language, but enclisis is assigned a new structure, which corresponds to a new parameter setting, sl i.e. to a new grammar. It is a strong hypothesis of this project that not only has prosody led people to choose enclisis in detriment of proclisis at a certain point of the history of ClP, but also led a generation of children to assign a new structure to a sub-class of these enclitic constructions, selecting a new grammar.

One of the main aspects of the syntactic part of this project is to study the syntax of order in order to obtain a precise description of the position of the verb with respect to the subject and the topicalized phrases in root and embedded sentences, and how this syntax evolves in time. The syntax of clitic-placement can be seen as a part of this study, and we predict that the change affecting the former will also affect the latter because they are dependent on the same parameter setting.

The project will also consider the evolution of the so-called phenomenon of interpolation, by which the clitic and the verb are separated by some phrase. This phenomenon, which is typical of the Medieval Romance languages only disappeared during the 17th century in Portuguese. An accurate description of the loss of this construction in correlation with the evolution of proclitic and enclitic structures will certainly help us understanding better the grammatical changes underlying the history of clitic-placement in the language.

2. The identification of the rhythmic patterns of BP and EP

2.1 Phonological assumptions

By rhythmic pattern, we understand the way languages organize rhythmic segments around syllables which have been assigned a positive prominence value at some level (lexically or post-lexically). In current phonological theory, the notion of rhythmic segment is an I-language notion, closely related to the notion of *foot* (cf among others Halle and Vergnaud 1987). Feet are constructed either lexically (by attribution of primary stress), or post-lexically (by attribution of secondary stress) within the prosodic domain of the phonological word. According to this view, rhythmic patterns are a result of the interaction among stressed and non-stressed syllables within a metrical grid which, once constructed, can be subject to readjustments that are eurhythmically motivated.

In this project we shall conduct an analysis of stress and rhythm within the framework of the minimalist view of the interaction between grammar and performance systems, in which it is possible to consider rhythm as an interpretation, by the articulatory-perceptual system, of the phonological output. The phonological output already contains instructions about lexical and phrasal prominences, as well as information about prosodic domains, as proposed by current phonological theories. Within the minimalist framework, however, it seems possible to view secondary stress not as attributed by an algorithm grammatically defined, but, instead, as a product of phonetic implementation. Our hypothesis is that this implementation consists of assigning phonetic prominence to originally unstressed syllables occurring between lexically stressed ones, in such a way that utterances are segmented in optimal rhythmic units.

It is clear that BP and EP do not differ with respect to primary stress. We also assume that both languages do not differ significantly also with respect to phrasal stress assignment. However, as was already shown by Carvalho (1989), they differ with respect to the way they assign secondary stress. Our hypothesis is that they crucially differ in the way they organize optimal rhythmic units at the level of phonetic implementation (cf. Abaurre 1995 and Abaurre and Galves 1998).

In this general framework, we propose that the difference between BP and EP derives from the way the phonological output is segmented into rhythmic units. In other words, syllables that do not receive stress at the lexical level and do not belong to lexical feet must either receive secondary prominence at the level of phonetic implementation, or be attached to some prominent syllable, eventually yielding reduction processes.

Our working hypothesis is based on Carvalho (1989)'s analysis of the rhythmic difference between BP and EP. From our point of view, his analysis can be rephrased in the following terms: only in BP is the rhythmic segmentation bound by the limits of the phonological word. In EP, rhythmic units are constructed by assigning a rhythmic boundary to the left of any stressed syllable. The results of our previous analyses show that this process is not bound by the *Phonological Phrase*, either. We are assuming, at this point of the investigation, that only Intonational Phrase boundaries block rhythmic segmentation in EP (cf. cf *Resultados de auxílios anteriores* for a first implementation of this analysis). From this point of view, the reduction of pre-tonic syllables derives from the fact that they are at the end of the rhythmic unit.

One of the basic steps of the project is to do a systematic description of large speech corpora. We'll perform a comparative study of vowel reduction phenomena and the attribution of secondary stress in EP and in BP aiming to understand their role in the implementation of the rhythm of the utterances. The methodology to annotate big speech corpora in such a way that acoustic informations such as vowel duration and fundamental frequency will be associated with phonetic informations about the segments and phonological informations such as prosodic domains, lexical stress, as well as perceptual informations like secondary stress is presented in the next section.

2.2 The phonetic description of rhythm

The production of speech obviously involves some sort of rhythmic regularity and factors of a durational order are associated to some extent with this rhythm. An initial question arises with respect to the way in which we find the clues in the speech signal which permit us to describe this regularity better. One suggestion based on experiments with synthetic speech is furnished by Marcus (1981), who observes that listeners tend to perceive the "rhythmic pulse" of speech at a point very close to the beginning of the vowel in CV syllables. This result is not surprising if one considers that the onset of the vowel involves a sharp increase in amplitude in relation to that of the preceding consonant, since the ear is extremely sensitive to such abrupt amplitude variations.

There is also experimental evidence that in CV syllables the listener extracts information about the vowel before the initial consonant (Wood and Day, 1975), suggesting that the focus of listener attention is effectively centered on the vowel. If perceptually relevant indicators of rhythmic regularity in the speech signal are the target, it seems to make little sense to measure duration on the basis of syllable onset alone, since syllable onset does not necessarily define a pulse for the listener. The present proposal studies rhythm on the basis of temporal markers anchored to the onset of the vocalic nuclei of syllables.

The first problem is to find a working definition for "syllable". Since the initial objective is to describe certain rhythmic differences between Brazilian and European Portuguese, the definition used must be capable of reflecting known differences. Since these differences develop largely due to reduction, or even erasing of certain segments or syllables; it seems logical to adopt a syllable reference which is independent of such an influence. The use of syllabic division based on the phonologic structure of the lexical item in isolation is thus proposed, and any phenomena of elision, reduction, or erasure observed in the phonetic realization of speech will always be described in relation to this canonical syllable division, as this division is similar in the two varieties of Portuguese.

It is evident that the temporal markers proposed here do not coincide with the limits of the canonical syllable except in special cases of vowel-initial syllables, but reference to such canonical syllables is necessary since there is a one-to-one relationship between each marker and the canonical syllable. The metric unit of temporal marking will be called, rather freely, the "rhythmic syllable" (RS).

Each RS will be associated with a duration, corresponding to the distance in time between the onset of one vocalic nucleus and that of the following canonical syllable. Most of the time, there will be no greater difficulty in identifying the RS, since the onset of a vocalic nucleus is generally a point which is easily detected in the speech signal. Certain problems may, however, arise in phonetic contexts involving the presence of some voiced consonants, such as laterals and nasals; in these more difficult cases, reading a wide-band spectrogram in conjunction with the waveform will make it possible for the trained observer to define the limits between successive sounds with reasonable precision. Even in unfavorable situations, a maximum error in the neighborhood of one to two periods is expected, corresponding to approximately 8-16 milliseconds for the male voice.

Cases involving some sort of syllabic restructuring in production will require rather arbitrary decisions. External Sandhi variation involving contiguous elements with the same vowel quality, such as "na analise" being produced in speech as "nanalise", will be difficult to interpret since there is no way of determining if the change involved total erasure of one of the vowels (CV1V2C – CV1C... or CV2C...?) or time compression of the two (in which case there would have been no syllabic restructuring as such). If it is impossible to determine that a specific vowel has actually been eliminated, the second analysis seems preferable, and the temporal markers for the onset of the syllabic V nucleus will be effected in the middle of the vocalic segment. When elision involving different vowels is present, such as the case of "de algum" produced as "dalgum", it is obvious that the vowel /e/ was suppressed. It will, however, be obligatorily marked in the proposed system, although with zero length. In conclusion, even if there is erasure/suppression of a vowel, this vowel will still be associated with a definite marker of the onset of the RS. It is hoped that this procedure will facilitate the visualization of certain differences between Brazilian and European Portuguese related to the time reduction of segments in certain environments (phonetic, phonologic, syntactic, etc.) better.

Rhythm depends on a relatively complex interaction between length, pitch and intensity (Cf. Allen 1968), although intensity has been considered not to be especially important for the perception of accent (Fry 1955). In fact, most systems of speech synthesis do not treat variations of intensity very rigorously, dealing with it merely as a redundant parameter, with variation determinable from other factors (Nagle and Chiquito, 1993). On the other hand, it is known that the absence of information about F0 even affects the perception of basic temporal aspects, such as speech rate (Cf. den Os, 1985). It thus seems convenient to associate a measurement of F0 with each RS . This measurement should be taken directly from the waveform as the inverse of the time interval of one period extracted from the most stable region of the vocalic nucleus of the RS. In principle, a single F0 value for each vowel should be sufficient to describe the contours.

All measurements of length and F0 will be made using the CSL 4300-B system of the Computerized Speech Lab of Kay Elemetrics. This system permits the creation of short routines (macros) which facilitate and help standardize measurement. The measurement will be based on pre-programmed screens which simultaneously present three different windows showing a) the wave sampled at 10 Khz, b) the corresponding wide-band spectrogram, and c) a signal of constant intensity in which different types of clicks can be inserted at whatever points are considered relevant. This parallel control signal has exactly the same length as the original speech signal and will always be taped as a second channel independent of the original signal, yet associated with it. There is thus no danger of loosing the information from time markers and F0. Whenever the signal is loaded into CSL 4300-B, the signals from the two channels are dissociated so that the operator can easily modify the control signal, including new markers and/or erasing old ones.

Figure 1 shows a typical screen, illustrating the pronunciation of the word "fonetica". The contents of the windows would be as follows:

Window A: Waveform;

Window B: Reference signal with markers;

Window E: Wide-band spectrogram associated with the waveform and the marked control signal;

Windows C and D: supplementary windows for operations of automatic correction.

Time markers must always be made based on short segments of the signal. The example in Figure 1, for example, displays a signal of 0.786 seconds; this makes it possible to visualize significant acoustic events clearly for an analysis. In Window B, the intensity of the markers inserted in the control signal can be modified, thus providing signals of either greater or lesser intensity. Each of the markers illustrated here has a greater intensity than that of the reference intensity and corresponds to the onset of a vocalic nucleus.

Markers for the measurement of F0 are indicated using markers with a lower intensity than that of the reference signal. Figure 2 shows two such markers, one at the beginning and the other at the end of the period in the stable region of the vowel /E/ in "fonetica" in addition to that of greater intensity associated with the onset of the vocalic nucleus. It can be seen that the periods can easily be detected in the waveform. The inclusion of a short stretch of the signal (approximately 250 ms in the example given in Figure 2) in the window improves visualization for marking the periods.

Figure 3 again shows the word "fonetica", now with the markers of both vocalic onset and F0. It can be seen that a marker at the onset of a vocalic nucleus and two markers for F0 are associated with each RS. The file containing these markers of contrasting intensity is then converted into text format. Table 1 shows a short stretch of a text file with the two successive markers of the onset of RSs (samples a1 and a2 - with markers of greater intensity than that of the reference) and two markers indicating F0 (samples b1 and b2 - with markers of lower intensity). The existence of two modified samples for each mark is due to a limitation of the CSL 4300-B system, but this limitation does not interfere with accuracy since distances are always computed on the basis of the first modified sample.

Table 1 shows a reference intensity of 15000, with an intensity of 30000 used for the marker of the onset of the vocalic nucleus and 3000 for that at the beginning and end of the F0 period. (The table obviously does not represent true distances, which would involve a long list of numbers). Since the sampling rate is 10 Khz, each line shown represents only a tenth of a milisecond; hence, the accuracy of measurement is guaranteed.

A simple routine can be used to select the desired time intervals automatically on the basis of the text file, as shown in Table 1. The measurements of length of the RS and the F0 of the vocalic nucleus can then be associated with the orthographic transcriptions of the RS, as well as a set of categorial variables (C1, C2, ... Cn) which may be relevant for the examination of specific questions of a phonological or syntactic order (RS structure, word stress, word position, etc), as shown in Table 2.

All these data will be submitted to statistical analysis, searching for correlations between the various levels of description, so that we can empirically confirm our conjectures about the rhythmic patterns of BP and EP.

Given our conjectures about the role of the implementation of the rhythmic pattern and the morpho-syntactic choices, we shall work preferentially with corpora of recorded authentic speech, without excluding the possibility of controlling some phonetic or phonological factors with constructed sentences.

3. Modelling rhythmic patterns and grammatical change.

If we succeed in justifying our hypothesis about how EP and BP rhythmically segment utterances, we can work with some chance of success on the following conjecture:

A rhythmic pattern induces the presence in written texts of templates of a well-defined type.

If this conjecture suggested by Anthony Kroch holds, then it follows that identifying the presence of templates of a given type in written texts will provide precise information about the rhythmic pattern of the corresponding spoken language.

In order to detect these templates, statistical tools like the measure of the entropy of the possible segmentations will be crucial. We expect that the sequences of rhythmic units produced by each rhythmic pattern have remarkable properties, em particular that the right segmentation corresponds to less dispersion, *i.e.* to less *entropy*, and conversely the wrong segmentation leads to a greater entropy. Based on our hypothesis about the different algorithm of segmentation in EP and BP, and making use of the phonetic correlations that the statistical study on oral texts will evidence, we shall be able to associate with written texts two kinds of possible segmentation together with their phonetic properties. This means that we shall be able to make hypotheses about the phonetic form of the segments as assigned by the corresponding rhythmic pattern. This will enable us to study their entropy.

We thus expect that using this method we shall be able to detect the change of rhythmic pattern in Portuguese, at some point of the 18th century. This will be when the segmentation corresponding to ClP (by hypothesis the same as BP) will be no more the less entropic one. The exact calculus of the measure of entropy is the basic question to be solved in this part of the project.

The modelling of the relationship between the rhythmic change and the grammatical change in the process of grammar selection by children acquiring their native language is the other aspect of the project in which mathematical formalisation is crucial for linguistic analysis.

In previous work, Galves and Galves (1995) have come to the conclusion that rhythmic changes in ClP are causally related to the syntactic change. We aim to verify this causal relation, using our empirical studies to inform the development of a model of language acquisition which the learner is guided by hints of a prosodic nature (Morgan 1986, Scarpa 1995). We make crucial use the Thermodynamical Formalism (Ruelle 1978) as a framework in which the notion of rhythmic pattern can be properly defined and articulated with a grammar in the definition of the probability measure governing the choice of the sample of positive evidence offered to a learning child.

The idea that a phonological change may drive a syntactical jump is not new. As far as we know it appears for the first time in an informal way in the framework of generative grammar in the study of the change from Old to Modern French proposed in Adams (1987). Kroch (1989) shows that this point of view is supported by the quantitative analysis performed in Fontaine (1985) and is coherent with the constant rate hypothesis.

Galves and Galves (1995) make an effort to develop in a systematic and rigorous way this point of view, by defining a model of language acquisition in which prosody plays a role in the reconstruction of the syntactic structures of the sentences presented to the learning child. The Thermodynamical Formalism provides a framework in which syntax and prosody can be put together, as parameters of the probability measure which governs the choice of the sample of sentences presented to the learning child. Using this model, they argue that the change in the stress pattern which presumably occurred in Portugal at the end of the 18th century (Teyssier 1980, Révah 1954) induced a parametric change in the grammar of Portuguese, as evidenced by the change in clitic-placement in root affirmative clauses.

Collet, Galves and Lopes (1995) present a general discussion of Maximum Likelihood and Minimal Entropy prosody driven identification models in the abstract mathematical setting of sub-shifts of finite type. An algorithm to perform this procedure is given in Cassandro and Galves (1995 a and b) and Cassandro, Galves and Galves (1996) who show how the stochastic acquisition model introduced by Gibson and Wexler (1994) (henceforth basic GW model) can be generalized in the framework of *simulated annealing processes* in which the identification principle is defined by a fitness function that can be associated to prosody. With respect to the basic GW model, we also refer to Frank and Kapur (1997) for a critical analysis of the notion of *trigger* considered by Gibson and Wexler and Berwick and Niyogy (1994) for a discussion of the rate of convergence of the model.

The type of model we are considering uses ideas coming from Statistical Physics, and has been widely used in different fields of pattern recognition (cf Geman, 1990 for a survey). Modeling the identification procedure of linguistic acquisition as a statistical inference makes it possible to use classical statistical criteria like the Maximum Likelihood Principle. The statistical inference is based on the hypothesis that the learning child knows what the structure of the probability measure is which governs the choice of the sample of positive evidence to which he is exposed. This probability measure is defined by two parameters.(Here we are using the word *parameter* in its standard statistical meaning, *i.e.* as an index which distinguish an element of a family. In particular, we **are not** saying that prosody is a parameter in the Chomskyan sense.) The first parameter is prosody, which we assume was previously acquired by the learner. The second parameter is the parental grammar, which is not known by the learner and is precisely the parameter to be estimated.

As pointed out by Kroch (1995 b), the conjecture that the acquisition procedure aims to adjust the prosodic potential and the grammatical parameter setting in an optimal way is a candidate to explain language change from inside and suggests the possibility of a new tool to identify the generation whose grammar change, if we are able to detect the new prosody in texts. This is the ultimate aim of our project.

4. Obtaining data

The above comments about the necessity of extending and controlling the data under analysis naturally lead us to an auxiliary project of constitution of a historical corpus on one hand, and of a speech data bank on the other hand.

1. The Tycho Brahe parsed corpus of historical Portuguese

A by-product of the project is the Tycho Brahe Parsed Corpus of Historical Portuguese, consisting of texts written by Portuguese authors born between 1550 and 1850. This electronic corpus, developed in the lines of the Penn-Helsinki Parsed Corpus of Middle English, will be hosted at the network of the Institute of Mathematics and Statistics of the University of São Paulo, being available to scholars for educational and research purposes. Our goal is to reach 1.000.000 words within four years.

The annotation of the Penn-Helsinki Parsed Corpus of Middle English is currently done with the help of two automatic analysers: a morphological tagger and a syntactic parser. The adaptation of these tools for Portuguese is part of the project (see *Resultados de Auxílios anteriores* for the adaptation of the Tagging scheme). The computational aspects of this adaptation will be coordinated by Marcelo Finger, (Department of Computer Sciences, IME-USP). The linguistic part of the adaptation as well as the correction of the outputs of the automatic parsers will be coordinated by Ilza Ribeiro at UEFS, and Charlotte Galves and Helena Britto at UNICAMP.

The texts which will be put in the corpus will be choosen mainly in function of the objectives of the project. This means that the privileged period will be from the second half of the 16th century to the first half of the 19th century and will concern Portuguese authors. The corpus will also include non litterary documents written by Portuguese and Brazilian people, currently edited in Brazil by the PROHPOR "Projeto para a história da língua portuguesa" coordinated by Rosa Virginia Mattos e Silva.

Our objective is to reach 1.000.000 words at the end of the four years of the thematic project. For that we will need automated tools to help us in the several activities needed to construct the corpus. The texts will have to be scanned and transferred to electronic media.

After that, we will do a first annotating step, which consists of tagging each word in the corpus with its associated part-of-speech. For that task, we plan to study how to adapt to Portuguese an automated tagger recently developed by Eric Brill for English. This tagger cannot provide 100% accuracy, so the result of the tagging process will have to be manually verified, but we hope that the use of automatic tools will considerably speed up the process of constructing the parser. Tools to help tagging and changing tags will also be developed. The tagger is being developed in the following steps:

1. Preliminary definition of tags to be used in the corpus

2. Manual tagging of 3000 words

3. Criticism of the preliminar tags on the light of the 3000 words manually tagged and consolidation of the tags.

4. Adaptation of Brill's algorithm to Portuguese.

5. In parallel with the previous activity, some further 17000–18000 word will be manually tagged. The 20000–21000 manually tagged words will be used to 'train' the tagger.

6. The tagger will be applied to new texts, that will be then manually corrected. Further improvements in the tagger may be suggested on the light of the kind of errors that are detected in this process.

The final stage in constructing the corpus consists of the development of an automated parser of portuguese, ie a collection of tools that will add syntactical tags to the text. This is a considerably more complex task than part-of-speech tagging, and studies on how to do that are still on course. But again, we shall take advantage of the work performed for English by Mike Collins, for the Penn-Helsinki project, and shall adapt it for Portuguese.

2. The speech data bank

The objective of the speech data bank is to provide a great and diversified corpus of comparative oral data from BP and EP. Besides providing the basis for a comparative description of the rhythm of both variety, this corpus will have to be constructed in such a way that it also allow us to bring evidence for the conjecture that, inside the limits of the restrictions imposed by the grammar of the language and given a discourse situation, lexical and morpho-syntactic choices are triggered by the implementation of the rhythmic pattern of the language. This will require texts which are instanciations of the same underlying discursive and semantic frame.

Two types of speech data will be used:

reading of a prose passage

spontaneous speech

For European Portuguese, reading data will be taken from television news broadcasts as transmitted in Portugal, since the sound quality of these transmissions is adequate for the analyses to be conducted. Text-dependent comparisons with BP will be based on these same texts, read by professional BP speakers, although sentences with non-compatible syntactic structure will be excluded. Text-independent comparisons can be made with other texts in the same style, extracted from Brazilian TV news broadcasts. Spontaneous speech data for both varieties (BP and EP) will also be obtained from television interviews on talk shows, as these are an excellent source for this type of material.

It is hoped that socio-stylistic variables will have little effect since for both types of television program utilized (interviews and news broadcasts), the situational context is practically the same in the two countries.

It is estimated that 45 minutes of speech of each style for each language will be adequate for the extraction of a corpus containing a sufficient number of RSs. At an average rate of 5 syllables per second, this would result in a total of over 50,000 observations (45 x 60 x 2 x 5 = 54,000).

5. References.

Abaurre, M.B. 1995. "Padrões fonéticos do ritmo em português", talk at the *2nd Workshop on Statistical Physics, Pattern Recognition and Grammar Selection*, São Sebastião.

Abaurre, M.B and C. Galves. 1998 "As diferenças rítmicas entre o português europeu e o português brasileiro: uma abordagem minimalista e otimalista", to appear in *D.E.L.T.A.*

Adams, M. ,1987. *Old French, null subjects and verb second phenomena*, PhD thesis, UCLA.

Allen, G.D. 1968. "The place of rhythm in a theory of language", *Working Papers in Phonetics - UCLA* **10**, 60-84.

Berwick, R. and P. Niyogi, 1994. "Formalizing triggers: a learning model for finite spaces", *A.I. Memo* No.1449, MIT, Cambridge, MA.

Carvalho, J. Brandão, 1989. "Phonological conditions on Portuguese clitic placement: on syntactic evidence for stress and rhythical pattern", Linguistics **27**, 405-436.

Cassandro, M. and A. Galves 1995a. "Convergence time and language change. Comments on the GW model", *Colloque Langues et Grammaires 2*, Université de Paris VII.

Cassandro, M. and A. Galves 1995b. "Language acquisition and change in a generalized GW model", *Fourth Meeting on Mathematics of Language*, University of Pennsylvania.

Cassandro, M., A. Galves and C. Galves 1996. "Structure recognition and language change in a thermalized GW model", *ICML 96* (2^{nd} *International Conference on Mathematical Linguistics*), Tarragona.

Chomsky, N., 1985. *Knowledge of Language. Its Nature, Origin and Use*, New York:Praeger.

Chomsky, N. 1995. *The Minimalist Program*, Cambridge, MA: The MIT Press.

Cinque, G., 1993. "A null theory of phrase and compound stress", *Linguistic Inquiry* **24**, 239-297.

Collet, P., A. Galves and A. Lopes, 1995." Maximum likelihood and minimum entropy identification of grammars", *Random and Computational Dynamics* **3**, 241-256.

Fontaine, C.,1985. *Application de méthodes quantitatives en diachronie: l'inversion du sujet en français.* M.A. thesis,Université du Québec à Montreal.

Frank, R. and S. Kapur, 1997. "On the use of triggers in parameter setting", *Linguistic Inquiry* **28** .

Fry, D. 1955. "Duration and intensity as physical correlates of linguistic stress", *JASA* **27**, 765-768.

Galves, C. 1996a. "Clitic-placement and Parametric Change in Portuguese", in M. Saltarelli (ed.) *Selected Papers from the 24th LSRL*, Georgetown University Press, 227-240.

Galves, C. 1996b. "A história da colocação de clíticos em português" talk presented at the 12th Meeting of the APL, Braga.

Galves, A. e C. Galves 1995. "A case study of prosody driven language change. From Classical to Modern European Portuguese", manuscript, USP/UNICAMP.

Geman, D., 1990. "Random fields and inverse problem in imaging", in *18e Ecole d' été de probabilités de St Flour*, Lecture notes in mathematics, Berlin, Heidelberg, New York: Springer.

Gibson, E. and K. Wexler, 1994. "Triggers", *Linguistic Inquiry* **25**, 407-454.

Halle, M. and J.R. Vergnaud, 1987. *An essay on stress.* Cambridge, MA:MIT Press.

Kayne, R., 1991. "Romance clitics, verb mouvment and PRO", *Linguistic Inquiry* **22**, 647-686.

Kroch, A., 1989. "Reflexes of grammar in patterns of language change", *Language Variation and Change* **1**, 199-244.

Kroch, A. 1994. " Morphossyntactic variation" to appear in K. Beals et al. , eds. *Papers from the 30th regional meeting of the Chicago Linguistics Society: parasession on variation and linguistic theory.*

Kroch, A. 1995 a. "The statistical modeling of language change", manuscript, University of Pennsylvania.

Kroch, A. 1995 b. *Advocatus diavoli* talk at the *2nd Workshop on Statistical Physics, Pattern Recognition and Grammar Selection*, São Sebastião.

Kroch, A. 1996. "The time course of language change", manuscript, University of Pennsylvania.

Lightfoot, D. 1991. *How to set parameters*, Cambridge, MA: The MIT Press.

Lobo, T. 1992. *A colocação de clíticos em português, duas sincronias em confronto* unpublished Master thesis, Lisbon.

Marcus, G. 1993. "Negative Evidence in Language Acquisition", *Cognition* **46**.

Marcus, P. 1981. "Acoustic determinants of perceptual-center (p-center) location", *Perc. Psychoph.* **30**, 247-256.

Martins, A.M. 1994. *Clíticos na história do português*, unpublished PhD thesis, Lisbon.

Morgan, J. 1986. *From simple imput to complex grammar*, Cambridge, MA:MIT Press.

Nagle, E.J. e J.G. Chiquito 1993. "Sintese de sinais de fala usando o sintetizador de formantes de Klatt", *Anais do XI Simp. Bras. Telecom.*, 718-723.

den Os, E. 1985. "Perception of speech rate of Dutch and Italian utterances", *Phonetica* **42**, 124-134.

Révah, I. S., 1958. L'évolution de la prononciation au Portugal et au Brésil du XVI^e siècle à nos jours. In *Anais do primeiro Congresso Brasileiro da língua falada no teatro.*

Ribeiro, I. 1995a. "Questões sobre a ordem dos constituintes no português arcaico e no português clássico", talk at the *2nd Workshop on Statistical Physics, Pattern Recognition and Grammar Selection*, São Sebastião.

Ribeiro, I. 1995b. "Evidence for a V2 fase in Old Portuguese", in A. Battye and I. Roberts (eds.) *Clause Structure and Language Change*, New York: Oxford.

Ruelle, D., 1978. "Thermodynamic formalism". In *Encyclopedia of Math. and its Appl.* **5**, Reading, MA:Addison Wesley.

Salvi, G. 1990. " La sopravvivenza della legge di Wackernagel nei dialetti occidentali della peninsola iberica", *Medioevo Romanzo*, **15**, 177-210.

Scarpa, E. 1995. "When syllables are utterances", talk at the *2nd Workshop on Statistical Physics, Pattern Recognition and Grammar Selection*, São Sebastião.

Teyssier, P., 1980. *Histoire de la langue portugaise.* Paris: P.U.F.

Torres Moraes, M.A. 1995. *Do português clássico ao português europeu moderno: um estudo da cliticização e movimento do verbo.* Doctoral dissertation, UNICAMP.

Wood, C.C. e R.S. Day 1975. "Failure of selective attention to phonetic segments in consonant vowel syllables", *Perc. Psychoph.* **17**, 346-350.

Zubizarreta, M.L., 1994. "Word order, prosody and focus", manuscript, USC.