Projeto de parceria

entre

Laboratório de Fonética do Centro de Estudos da Linguagem (Onset-CEL),

Fac. Letras, Universidade de Lisboa,

Núcleo de Modelagem Estocática e Complexidade (NUMEC)

da Universidade de São Paulo,

Projeto Tycho Brahe

[suportado pelos Projetos Temáticos Fapesp « Padrões prosódicos, fixação de parâmetros e mudança linguística » (coordenado por Charlotte Galves do Instituto de Estudos da Linguagem da Universidade de Campinas), e « Comportamento estocástico, fenômenos críticos e identificação de padrões rítmicos em línguas naturais» (coordenado por Antonio Galves, do Instituto de Matemática e Estatística da Universidade de São Paulo)]

Foi conjecturado por linguistas que as classes rítmicas são caracterizadas pelo fato de certos domínios prosódicos e/ou certa informação fonético-fonológica serem ou não relevantes. Por exemplo, entre as propriedades conducentes ao ritmo acentual estão a maior complexidade silábica e maior irregularidade da distribuição de vogais e consoantes, a distinção forte entre silába acentuada e não acentuada e, presumivelmente, um papel mais claro das fronteiras de palavra enquanto elementos delimitativos de sequências de sons; já entre as propriedades conducentes ao ritmo silábico, estão a maior simplicidade silábica e maior regularidade da distribuiçãode vogais e consoantes, a não distinção entre sílaba acentuada e não acentuada no que respeita à composição e duração da sílaba e um papel demarcativo das fronteiras de constituintes prosódicos superiores à palavra, designadamente do sintagma fonológico. A meta deste projeto de parceria é a obtenção de evidências estatísticas em corpora escritos que dêem suporte à essa conjectura. Com essa meta, de um ponto de vista prático, nosso objetivo é fazer modelos para cadeias simbólicas obtidas codificando características fonológicas de textos escritos do Português Brasileiro contemporâneo, do Português Europeu contemporâneo e do Corpus do Tycho Brahe.

A codificação será feita marcando características fonológicas conjecturadas como sendo relevantes para a implementação do «ritmo» da língua (por exemplo, a distribuição de Cs e Vs, as fronteiras silábicas, as fronteiras de palavra prosódica, o acento de palavra), utilizando as funcionalidades do Programa FreP, em desenvolvimento pelo Centro de Pesquisa Onset.

Os modelos pesquisados deverão satisfazer o princípio de Comprimento Mínimo de Descrição (Minimum Description Length – MDL), introduzido por Jorma Rissanen nos anos 70. Isto é, trata-se de encontrar o modelo, numa classe de modelos dada, que minimize a soma dos seguintes comprimentos: comprimento da sua descrição e comprimento da compressão da cadeia simbólica comprimida usando o modelo. O segundo ponto é equivalente a pedir que o modelo maximize a verosimilhança da sequência.

Uma classe de cadeias simbólicas que tem se revelado suficientemente flexível e econômica e tem sido aplicada com sucesso tanto a dados biológicos quanto linguísticos são as Cadeias de Memória Variável. Essas cadeias são descritas por árvores de contextos probabilísticas.

Há atualmente algoritmos, como o algoritmo Contexto ou o algoritmo das Florestas probabilísticas, que permitem estimar de maneira consistente as árvores de contextos probabilísticas subjacentes a uma cadeia simbólica dada. Isso significa o seguinte : se a cadeia for efetivamente gerada utilizando-se como mecanismo de geração uma árvore de contexto probabilística, e se a amostra for suficientemente grande, então o algoritmo de estimação identifica exatamente essa árvore. Acrescente-se a isso que há evidências empíricas de que o algoritmo das Florestas probabilísticas é também robusto. Isso significa o seguinte: se uma cadeia for engendrada majoritariamente por uma árvore probabilística dada, e em seguida sofrer pequenas sub-sequências simbólicas engendradas por outras árvores, ou simplesmente perturbadas por ruído na manipulação dos dados, então o algoritmo das Florestas probabilísticas identifica aparentemente a árvore majoritariamente responsável pela geração dos dados. Essa propriedade do algoritmo está em fase de demonstração.

Um dos principais interesses das árvores de contexto probabilísticas é que elas podem ser interpretadas linguisticamente. Isso abre um campo de interação frutuosa entre matemáticos e linguistas. Com efeito, os contextos probabilísticos definem em cada passo a porção do passado pertinente para a escolha do próximo símbolo. Em proteómica, é um fato experimentalmente comprovado que os contextos probabilísticos descrevem bem domínios biológicos na cadeia de amino-ácidos definindo uma proteina. Trata-se então de entender se em linguística também a noção de contexto probabilístico fornece uma formalização adequada para os domínios prosódicos ou a informação prosódica relevantes. Se assim for, teremos encontrado uma ferramenta estatística para detectar características prosódicas relevantes em cadeias simbólicas obtidas codificando-se corpora escritos.

A distribuição das tarefas entre linguistas e matemáticos fica agora clara.

1. Em primeiro lugar cabe aos linguistas propor as questões e conjecturas relevantes a serem tratadas.
2. Segundo, a partir dessas questões cabe aos linguistas propor a codificação pertinente dos dados linguísticos.
3. Cabe aos matemáticos propor classes de modelos que potencialmente têm as características desejáveis, e também propor métodos estatísticos para identificação de modelos que melhor se ajustem aos dados empíricos codificados.
4. Cabe aos linguistas identificar o possível significado linguístico das características dos modelos estimados em 3.
5. Cabe aos matemáticos identificar um quadro matemático mais geral no qual as características do tipo das identificadas em 4. sejam verificadas.

O ítem 5 explica porque um projeto interdisciplinar como o nosso pode ser interessante, de um ponto de vista estritamente matemático. Com efeito a construção de modelos probabilísticos, capazes de descrever bem o comportamento dos dados linguísticos e as propriedades desses modelos, podem sugerir questões matemáticas novas, interessantes independentemente da sua motivação linguística inicial. Esse quadro matemático mais abstrato é também interessante para os linguistas na medida em que ele põe em evidência as características essenciais dos modelos linguísticos e os seus possíveis desenvolvimentos.

Constituirão outputs desta parceria:
1. A codificação das características fonológicas do Corpus do Tycho Brahe, com recurso à ferramenta FreP.
2. O levantamento das questões relevantes sobre as propriedades rítmicas do Português Europeu actual e do Português Clássico, bem como o estabelecimento de uma hipótese da deriva histórica do ritmo do Português.
3. A modelagem estocástica das sequências simbólicas linguisticamente codificadas.
4. A interpretação linguística dos resultados da modelagem estocástica, designadamente tendo em consideração as questões e hipóteses referidas em 2.
5. Os resultados de 1 a 4 darão origem a artigos em parceria a apresentar em encontros científicos e a submeter a revistas científicas das especialidades envolvidas.

Lisboa, 2 de Maio de 2007

Sónia Frota, pelo Centro de Estudos da Linguagem (Onset-CEL)

Charlotte Galves, pelo NUMEC e pelo Projeto Tycho Brahe