A unidade informacional de Tópico na *Language into Act Theory*

Teste de concordância e estudo de padrões melódicos

CAVALCANTE, Frederico[1]
RASO, Tommaso[1]
BARBOSA, Plínio A.[2]

[1]Universidade Federal de Minas Gerais
[2]Universidade Estadual de Campinas

Resumo: *Neste trabalho apresentamos os resultados de um estudo das formas prosódicos da unidade de Tópico (TOP), assim como a Language into Act Theory define a unidade. Apresentaremos os resultados de um teste Kappa realizado para estimar o grau de acordo na identificação do TOP em dados de fala espontânea com a participação de 4 anotadores. Em seguida tratamos da validação estatística, com base em dados do inglês americano e português europeu e brasileiro, do esquema de classificação proposto em estudos anteriores para as formas melódicas do TOP. Na validação, utilizamos a Análise de Dados Funcionais, Análise Funcional de Componentes Principais e ANOVA. Concluímos mostrando que o acordo obtido entre os anotadores é substancial e que as técnicas estatísticas empregadas corroboram a descrição do TOP realizada por estudos anteriores. Além disso, propomos modelos estatísticos para as formas melódicas do TOP.*
Palavras-chave*: Language into Act Theory; Tópico; Formas prosódicas.*

Abstract: *In this paper we present the results of a study of the prosodic forms of the Topic unit (TOP), as defined within the framework of the Language into Act Theory. We present the results of a Kappa test conducted to estimate the degree of agreement among 4 annotators in the identification of TOP based on spontaneous speech data. We then present a statistical validation of the classification scheme proposed in previous studies for the melodic forms of TOP, based on the analysis of spontaneous speech data from corpora of American English and European and Brazilian Portuguese. We utilized Functional Data Analysis, Functional Principal Components Analysis and ANOVA for the validation. We conclude by showing that the degree of agreement among the annotators is substantial and that the statistical techniques that we used corroborate the descriptions made in previous studies. Furthermore, we propose statistical models for the melodic forms of TOP.*
Keywords: *Language into Act Theory; Topic; Prosodic Forms.*

1 Introdução

Este trabalho apresenta um estudo estatístico das formas prosódicas de Tópico (TOP), assim como definido pela *Language into Act Theory* (L-AcT; Cresti, 2000), a qual tem no *enunciado*, i.e., a menor sequência do fluxo da fala com autonomia pragmática e prosódica, a unidade de referência para a análise da fala espontânea. Examinamos aqui dados provenientes de corpora de fala espontânea do inglês americano (IA), do português europeu (PE) e do brasileiro (PB).

A fronteira de um enunciado é sinalizada por uma quebra prosódica de tipo terminal. Todo enunciado necessariamente contém uma unidade prosódica, chamada de Comentário (COM), que veicula a força ilocucionária (Austin, 1962). O COM pode ser acompanhado por outras unidades informacionais opcionais, delimitadas entre si por quebras não-terminais. O TOP, uma dessas unidades, é encontrado em 9-15% dos enunciados, o que depende do quão interativo e dependente situacionalmente é o evento comunicativo. A função do TOP é fornecer um âmbito cognitivo para a interpretação da força ilocucionária, a qual em ausência de um TOP é interpretada com base em um domínio dado contextualmente.

No nível prosódico, indispensável para a veiculação do TOP, a pesquisa mostra que há, essencialmente, três padrões funcionalmente equivalentes de formas que codificam a função da unidade. Essas formas são normalmente caracterizadas por alongamento silábico e níveis mais altos de intensidade, sendo classificadas em três tipos, segundo o movimento de f0 que exibem, como se segue. Tipo 1: movimento ascendente-descendente na última tónica e eventuais postônicas da unidade; Tipo 2: movimento ascendente na última tônica e eventuais postônicas; Tipo 3: núcleo dividido em dois seminúcleos com possível interposição de sílabas sem função informacional; o primeiro seminúcleo apresenta valores normalmente altos de f0, enquanto que

o segundo apresenta valores mais baixos. A diferença de altura entre o primeiro e o segundo seminúcleo pode variar bastante. Os diagramas abaixo ilustram cada uma das formas.

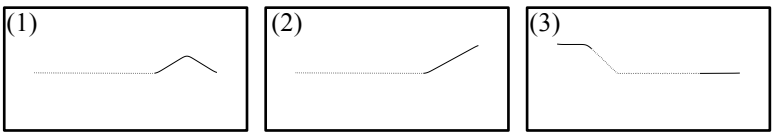

Figura 1: Diagrama das formas de TOP: (1) Tipo 1, (2) Tipo 2, (3) Tipo 3.

Vale salientar que, para a L-AcT, o TOP é uma unidade de natureza prosódico-pragmática, para a qual não é estabelecida restrições sintáticas ou semânticas *a priori*. Sintagmas de diferentes categorias já foram observados preenchendo o TOP, inclusive diversos casos em que é impossível postular uma relação sintática entre o conteúdo do TOP e o do COM. A dimensão propriamente pragmática do TOP refere-se à sua relação com a ilocução (e não com a predicação).

O TOP vem sendo objeto de análise de diferentes estudos realizados nos quadros da L-AcT – cf. Raso, Cavalcante e Mittmann (2017), para a síntese mais recente. Esses estudos, baseados em dados do italiano (Firenzuoli & Signorini, 2003), português europeu e brasileiro (Mittmann, 2012; Rocha, 2012) e inglês americano (Cavalcante, 2016), identificaram 3 diferentes formas prosódicas para o TOP. A identificação dessas formas se deu através da observação pormenorizada e manipulação de parâmetros acústicos da unidade, através de uma metodologia que não incluiu uma validação estatística.

Tendo isso em vista, o presente trabalho tem como objetivo estabelecer o grau de acordo entre quatro anotadores numa tarefa de identificação do TOP em dados de fala espontânea do português brasileiro, utilizando a estatística Kappa (Fleiss, 1971) e submeter a classificação das formas prosódicas do TOP a uma validação estatística, aplicando técnicas de Análise de Dados Funcionais e de Componentes Principais (Gubian et al., 2015; Ramsay & Silverman 2005) a 137 contornos melódicos extraídos de corpora de fala do IA, PB e PE, compilados de acordo com a metodologia proposta pela L-AcT.

2 Metodologia

Os contornos de TOP aqui analisados provêm de corpora de fala espontânea coletados de acordo com a metodologia estabelecida pela L-AcT. Trata-se dos minicorpora da família C-ORAL para o PB (Mittmann & Raso, 2011) e IA (Cavalcante& Ramos, 2016) e do componente do PE do C-ORAL-ROM (Cresti & Moneglia, 2005). A amostra contém 143 contornos melódicos pertencentes às três formas descritas acima, sendo 56 do IA, 59 do BP e 28 do EP, este com um número reduzido de formas devido ao fato de o corpus não apresentar anotação informacional.

Para o teste de concordância, utilizamos a estatística Kappa e uma amostra contendo 100 enunciados complexos amostrados aleatoriamente a partir das seções *Telefônicas* (20 enunciados), *Mídia* (40) e *Contexto Natural* (40) da seção formal do C-ORAL-BRASIL (Raso & Mello, 2012).

Utilizamos a *Análise de Dados Funcionais* (FDA, da sigla em inglês), para a suavização e alinhamento temporal dos contornos melódicos, e a *Análise de Componentes Principais de dados funcionais* (F-PCA), para revelar os principais modos de variação da amostra e para obter modelos, a fim de alcançar uma representação fidedigna dos dados e das classes que representam. Além disso, utilizamos os PC escores obtidos por meio da F-PCA para testar a separabilidade das curvas e determinar se elas de fato se agrupam de acordo com a classificação proposta pelos estudos anteriores. Os procedimentos estatísticos foram conduzidos sem

qualquer informação no que diz respeito à classificação dos contornos, informação esta que só foi introduzida na fase final da análise, para verificar se os padrões revelados pelas técnicas corroboravam a nossa classificação.

3 Análises e resultados

Com relação ao teste de concordância, em média (e desvio padrão), cada anotador identificou 52 (2.7) TOPs na amostra, que contém 541 unidades prosódicas. Houve acordo total em 36 casos e acordo parcial em 16, dos quais 11 foram identificados como TOP por três anotadores, e 5 por dois anotadores. Os coeficientes *k* obtidos são 0.79 (Geral), 0.80 (Mídia), 0.79 (Contexto Natural), 0.66 (Telefônicas). Vê-se que, independentemente do modo como os dados são particionados, o acordo obtido é *substancial*.

Passando à análise das formas prosódicas, a figura abaixo mostra todos os contornos melódicos de TOPs da amostra suavizados e alinhados temporalmente através da FDA.

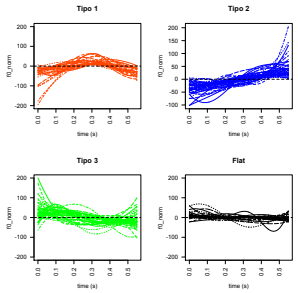

Figura 2: Contornos entonacionais suavizados e alinhados temporalmente. Curvas separadas de acordo com a classificação prévia para uma melhor visualização. Para *Flat*, ver texto abaixo.

Vemos acima que ao submeter os dados à suavização e ao alinhamento, tornando-os diretamente comparáveis, os padrões que emergem já sugerem a separabilidade das curvas de acordo com a classificação de Raso et al. (2017) (cf. Figura 1).

A Figura 3 mostra os modelos obtidos a partir da F-PCA. O se vê são formas que emergem dos dados, pois a informação sobre a classe à qual cada curva pertence é introduzida após a aplicação da FDA e da F-PCA. As curvas geradas são essencialmente idênticas àquelas dos diagramas da Figura 1, o que corrobora a classificação descrita em Raso et al. (2018).

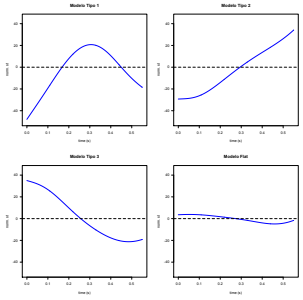

Figura 3: Modelos das formas de TOPs segundo a equação ($f(t) \approx \mu(t) + s1 * PC1(t) + s2 * PC2(t)$), obtidos a partir das medianas dos PC escores dos contornos de cada tipo para os 2 primeiros PCs.

Ao submetermos os escores dos componentes principais, obtidos via F-PCA, à ANOVA, tem-se que os escores do primeiro componente principal por si só já bastam para que possamos assumir a separabilidade das formas: H(2)=107.18, p < .001 – exceto para o par 1&2, para o qual $p < 0.01$).

Há um grupo de curvas que apresentam pouca variação de f0 e que estão representados na Figura 3 como *Flat*. Como as curvas de Tipo 3, as *Flat* apresentarem a possibilidade de núcleos descontínuos e um movimento global descendente, mas com uma variação de f0 bem menor. Estamos atualmente buscando verificar se a forma *Flat* de fato constitui uma classe individual ou se devemos considerá-la como uma variação do Tipo 3. Para tanto, estamos verificando o padrão duracional da unidade.

4 Considerações finais

Neste trabalho, mostramos que, numa tarefa realizada com dados de fala espontânea, o acordo na identificação do TOP é substancial. Através de técnicas estatísticas para a análise de dados funcionais, pudemos corroborar o esquema de classificação das formas de TOP sintetizado em Raso et al. (2017) e mostramos como podemos chegar a representações estatisticamente válidas de cada uma das classes. A pesquisa sobre o TOP ainda precisa responder de modo mais completo a questões sobre padrões de duração exibidos pela unidade e sobre a forma *Flat*. Esta constitui a atual direção da pesquisa.

REFERÊNCIAS

Austin, John L. *How to do things with words*. Oxford: Oxford University Press, 1962.

Cavalcante, F.; Ramos, A. The American English spontaneous speech minicorpus: architecture and comparability. *CHIMERA: Romance Corpora and Linguistic Studies*, v. 3.2, p. 99–124, 2016.

Cavalcante, Frederico Amorim. *The topic unit in spontaneous american english: a corpus-based study*. 2016. UFMG, 2016.

Cresti, Emanuela. *Corpus di Italiano parlato*. Firenze: Accademia della Crusca, 2000.

Cresti, Emanuela; Moneglia, Massimo (Org.). *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2005.

Firenzuoli, Valentina; Signorini, Sabrina. L'unitá informativa di topic: correlati intonativi. In: Marotta, G. (Org.). . *in Atti delle Giornate del Gruppo di Fonetica Sperimentale - XIII, Pisa, Novembre 2002 ETS, Pisa*. [S.l: s.n.], 2003. p. 177–184.

Fleiss, Joseph L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, v. 76, n. 5, p. 378–382, 1971.

Gubian, Michele; Torreira, Francisco; Boves, Lou. Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, v. 49, p. 16–40, 1 mar. 2015.

Mittmann, Maryualê Malvessi. *O C-ORAL-BRASIL e o estudo da fala informal: um novo olhar sobre o Tópico no Português Brasileiro*. 2012. 240 f. Universidade Federal de Minas Gerais, 2012.

Mittmann, Maryualê Malvessi; Raso, Tommaso. The C-ORAL-BRASIL informationally tagged minicorpus. In: MELLO, HELIANA R.; PANUNZI, A.; RASO, TOMMASO (Org.). . *Pragmatics and Prosody. Illocution,Modality, Attitude, Information Structure and Speech Annotation*. Amsterdam/Philadelphia: John Benjamins, 2011. p. 151–183.

Ramsay, J. O. (James O.); Silverman, B. W. *Functional data analysis*. 2nd ed. ed. [S.l.]: Springer, 2005.

Raso, Tommaso; Cavalcante, Frederico; Mittmann, Maryualê M. Prosodic forms of the Topic information unit in a cross-linguistic perspective: A first survey. 2017, Napoli: Aracne editrice, 2017. p. 473–498.

Raso, Tommaso; Mello, Heliana. *C-oral-Brasil. I : corpus de referência do português brasileiro falado informal*. Belo Horizonte: UFMG, 2012.

ROCHA, Bruno Neves Rati de Melo. *Características Prosódicas do Tópico em PE e o uso do pronome lembrete*. 2012. 269 f. UFMG, 2012.