UNIVERSIDADE FEDERAL DE MINAS GERAIS

MARCUS LEPESQUEUR FABIANO GOMES

ANÁLISE DE AGRUPAMENTOS DOS PARÂMETROS DE TRANSITIVIDADE LINGUÍSTICA

Belo Horizonte

2018

MARCUS LEPESQUEUR FABIANO GOMES

ANÁLISE DE AGRUPAMENTOS DOS PARÂMETROS DE TRANSITIVIDADE LINGUÍSTICA

Monografia apresentada ao Programa de Pós-Graduação em Estatística do Instituto de Ciências Exatas da UFMG, como requisito parcial para obtenção do título de Especialista em Estatística.

Orientadora: Profa. Dra. Ilka Afonso Reis

Belo Horizonte

2018

RESUMO

A transitividade tem um papel central em grande parte das teorias linguísticas principalmente pelo fato de que, em um número muito significativo de línguas, as estruturas sintática e semântica da transitividade se sobrepõem. Ainda permanece um desafio, para as teorias linguísticas, encontrar uma explicação para essa tendência, no mínimo *quasi*-universal da gramática. O desenvolvimento de métodos computacionais para a análise de grande quantidade de dados tem tornado uma realidade as análises semânticas de *corpus* e possibilita lançar nova luz sobre os fenômenos linguísticos. Esse trabalho visa a apresentar uma metodologia estatística capaz de permitir se analisarem padrões semânticos e sintáticos da oração, chegando a resultados semelhantes a categorias teoricamente conhecidas. Esse tipo de metodologia pode ser útil, sob uma nova perspectiva, na investigação, de maneira relativamente independente da sintaxe, dos padrões semânticos a que os falantes são expostos. Em um processo de amostragem simples sem reposição, foram selecionadas 690 unidades oracionais de um corpus de 23 entrevistas orais. Essas unidades oracionais foram analisadas em termos de nove parâmetros de transitividade e de sua sintaxe oracional. O objetivo foi identificar grupos de orações que compartilham semelhanças em termos desse conjunto de traços. Os grupos encontrados revelam um tipo de significado protoconceptual das orações, que inclui traços aspectuais e actanciais que se relacionam. Os resultados evidenciam três cenários micro-narrativos básicos sobre os quais se desenrola o evento expresso na oração. Ainda que esses cenários apareçam correlacionados a certos padrões sintáticos, eles não são exclusivos destes últimos, ou seja, diferentes padrões sintáticos podem acomodar um mesmo padrão semântico geral.
Palavras chave: transitividade, análise de agrupamento, semântica, sintaxe oracional.

ABSTRACT

Transitivity has a central role in most linguistic theories, especially because in a very significant number of languages the syntactic and semantic transitivity structures overlap. It is still a challenge for linguistic theories to find an explanation for this, at least quasi-universal, grammatical trend. The development of computational methods for the analysis of large amounts of data has made corpus semantic analysis a reality and allows projecting a new light onto the linguistic phenomena. This work aims to present a statistical methodology capable of allowing the analysis of semantic and syntactic clausal patterns, reaching similar results found in theoretically categories. This type of methodology may be useful, from a new perspective, for the investigation of semantic patterns to which the speakers are exposed, regardless of syntax,. In a simple sampling procedure without replacement, 690 oral units were selected from a corpus of 23 oral interviews. These sentence units were analyzed in terms of nine transitivity parameters and their clausal syntax. The goal was to identify groups of sentences that share similarities in terms of this set of traits. The groups found reveal a kind of proto-conceptual meaning of the sentences, which includes related aspectual and actantial traits. The results show three basic micro-narrative scenarios on which the event expressed in clausal unfolds. Although these scenarios appear correlated with certain syntactic patterns, they are not exclusive. Different syntactic structures can accommodate the same general semantic pattern.
Keywords: transitivity, cluster analysis, semantics, clausal syntax.

SUMÁRIO

1 TRANSITIVIDADE LINGUÍSTICA

Na gramática tradicional, a transitividade é comumente descrita como uma propriedade do elemento verbal e se relaciona às diferentes complementações, sintáticas e semânticas, que aparecem nas orações. Cunha e Cintra (1985), por exemplo, adotam a tipologia tradicional que diferencia verbos intransitivos, de verbos transitivos diretos, indiretos e bitransitivos. Essa classificação baseia-se, fundamentalmente, na presença ou na ausência de algum tipo de objeto sintático e na consequente interpretação semântica que acompanha o uso do verbo.

Na prática, a linguística contemporânea tem mostrado que fenômenos relacionados à transitividade parecem ser muito mais complexos do que faz crer a gramática tradicional: primeiro, porque os verbos variam as características da sua regência em diferentes contextos de uso; segundo, porque as definições tradicionais da transitividade tratam, comumente, da mesma forma, elementos sintáticos e semânticos, que não apenas são distintos, mas que também interagem de forma complexa. Um amplo número de pesquisas voltadas para a análise desses fenômenos tem mostrando, de modo geral, que a transitividade se manifesta a partir de fatores sintático-semânticos e discursivo-pragmáticos que são simbioticamente dependentes (LUCENA e CUNHA, 2011).

A transitividade tem um papel central em grande parte das teorias linguísticas, principalmente pelo fato de que um número muito significativo de línguas apresenta uma estrutura formal – a morfossintaxe transitiva – cuja principal função é expressar um conjunto específico de propriedades semânticas (NAESS, 2007). A questão principal em torno do fenômeno da transitividade é que as características semântica e sintática da transitividade tendem a covariar, sendo um fenômeno universal, ou ao menos *quasi* universal das línguas humanas. Givón (2001) aponta que, apesar das características transitivas de uma oração parecerem independentes, é um fato, na maioria das línguas, que as estruturas sintática e semântica da transitividade se sobrepõem, de forma que grande parte das orações semanticamente transitivas são também sintaticamente transitivas. De forma semelhante, Naess (2007) tenta demonstrar que, em muitas línguas, uma oração que é formalmente distinta da oração transitiva também se desvia dessa oração transitiva em termos das suas propriedades semânticas, ou seja, a escolha por uma estrutura linguística diferente reflete o desejo do falante de exprimir uma semântica diferente do protótipo da transitividade.

Ainda permanece um desafio para as teorias linguísticas encontrar uma explicação para essa tendência, no mínimo *quasi*-universal, da gramática. Naess (2007) apresenta esse desafio na forma de duas perguntas: 1) por que as línguas convergem um certo conjunto de propriedades semânticas em um tipo de oração, ao invés de possuírem critérios independentes? 2) por que justamente essas propriedades, aparentemente tão díspares, aparecem tão estritamente correlacionadas? Essas são instâncias específicas da questão mais geral sobre como as unidades linguísticas podem se integrar, não em padrões caóticos e idiossincráticos, mas em estruturas sentenciais regulares e, portanto, comunicáveis. Em outras palavras, dada a infinidade de combinações estruturais possíveis à linguagem, os estudos sobre a transitividade podem nos ajudar a entender como uma língua se estabiliza em uma arquitetura específica, finita e compartilhável.

Em um artigo seminal sobre o tema, Hopper e Thompson (1980) isolaram alguns dos componentes da noção de transitividade e estudaram a forma como eles são tipicamente codificados na gramática. A partir de um conjunto de evidências translinguísticas os autores propuseram 10 parâmetros semânticos e morfossintáticos relacionados ao fenômenos da transitividade. Estes parâmetros incluem características relacionadas ao número de argumentos sintáticos da oração (o parâmetro Participantes); às relações aspectuais (os parâmetros Cinese, Telicidade[1] e Pontualidade) e modais (os parâmetros Modalidade e Polaridade) do evento expresso na oração; às noções actanciais ligadas ao sujeito sintático (Agentividade e Intencionalidade) e ao objeto sintático (Afetação e Individuação)[2]. O Quadro 1 resume a categorização destes parâmetros em termos de alta e de baixa transitividade.

[1] Hopper e Thompson (1980) se referem a esse parâmetro como "Aspecto". Para evitar confusão com o sentido mais geral de aspecto, utilizarei aqui o termo telicidade.
[2] Os detalhes sobre a maneira como cada parâmetro é operacionalizado estão disponíveis em Lepesqueur (2017).

Quadro 1- Parâmetros da transitividade

Parâmetros	**Alta transitividade**	**Baixa transitividade**
Participantes	2+ participantes	1 participante
Cinese	Ação	Estado
Telicidade	Télico	Atélico
Pontualidade	Pontual	Durativo
Intencionalidade	Intencional	Não-intencional
Polaridade	Afirmativa	Negativa
Modalidade	*Realis*	*Irrealis*
Agentividade	Agentivo	Não-agentivo
Afetação	Afetado	Não afetado
Individuação	Altamente individuado	Não individuado

Hopper e Thompson (1980) argumentam que as gramáticas das línguas agrupam este conjunto de traços em função de uma escala de transitividade, i.e., em uma mesma sentença, um traço morfossintático ou semântico obrigatório que marca alta transitividade tende a não ocorrer com outro que marca baixa transitividade. Nas palavras os autores:

> A título de exemplo, suponhamos que uma língua tenha uma oposição, marcada em sua morfologia, entre verbos télico e atélico. Suponhamos também que o O [Objeto sintático] na presença de um verbo télico seja obrigatoriamente marcado, morfologicamente, como possuindo uma das características de transitividade relevantes para os Objetos, por exemplo a Individuação. A Hipótese da Transitividade agora prevê que, se o verbo é télico (isto é, está no lado alto da escala de Transitividade para Telicidade), então O será também marcado como estando no lado alto da escala relativa aos Objetos nessa língua, neste caso a Individuação. (Hopper e Thompson, 1980, p. 255, tradução nossa)[3]

A fim de exemplificar esse pareamento entre parâmetros obrigatórios, descrito pelos autores, tomemos os exemplos [1] a [3] a seguir.

[1] Ele dependurou a roupa	[Intencionalidade = alta; Participantes = alta]
[2] Ele quebrou o vaso	[Intencionalidade = ?; Participantes = alta]
[3] O vaso quebrou	[Intencionalidade = baixa; Participantes = baixa]

[3] No original: "By way of example, let us suppose that a language has an opposition, marked in its morphology, between telic and atelic verbs. Let us assume also that the O in the presence of a telic verb is obligatorily signaled in morphology as possessing one of the Transitivity features relevant for O's, e.g. Individuation. The Transitivity Hypothesis now predicts that if the verb is telic (i.e. is on the high side of the Transitivity scale for Aspect), then the O will also be signaled as being on the high side of the other scale relevant for O's in this Language, viz. Individuation."

Em [1], o sujeito sintático “Ele” é obrigatoriamente interpretado como um agente intencional (parâmetro Intencionalidade=alta), ao contrário de [2], que tem uma interpretação ambígua: o sujeito sintático de [2] pode ser tanto um agente intencional, como um agente não-intencional da ação. A estrutura oracional de [1] tem necessariamente dois participantes, a saber “Ele” e “A roupa” (parâmetro Participante=alto), e não pode ocorrer nas formas inacusativas (“A roupa dependurou”), semelhantes a [3], com apenas um participante. Por sua vez, o sujeito sintático de [3], “O vaso”, recebe, obrigatoriamente, uma interpretação não intencional (parâmetro Intencionalidade=baixa) e essa interpretação necessariamente ocorre em uma estrutura oracional com apenas um participante (parâmetro Participante=baixo). Tratando-se de parâmetros obrigatórios, Intencionalidade e Participante tendem a aparecer do mesmo lado da escala[4].

Hopper e Thompson (1980) mostraram que esse tipo de pareamento entre parâmetros obrigatórios, do mesmo lado da escala, ocorre em um número significativo de línguas. Em Hopper e Thompson (2001), os autores ampliam o debate ao tratar não apenas de parâmetros obrigatórios, mas também de traços não obrigatórios da unidade oracional. Nesse contexto, a interpretação da intencionalidade na oração [2] depende de fatores textuais e pragmáticos e, mesmo não sendo um traço obrigatório, a oração aparece em uma estrutura de dois participantes, a saber, “Ele” e o “vaso”.

Nesse modelo a transitividade passou a ser definida não como uma característica do elemento verbal, mas como um conjunto de componentes ligados à unidade oracional que se relacionam de maneiras específicas. Essa mudança de perspectiva alimentou uma série de pesquisas que investigaram tanto a maneira como as línguas codificam formalmente os parâmetros da transitividade, quanto as motivações semânticas e pragmáticas da variação na morfossintaxe transitiva.

No português do Brasil, Lepesqueur (2017) mostrou que apenas um dos parâmetros propostos por Hopper e Thompson (1980), a saber, a Afetação do objeto sintático, é um preditor positivo, estatisticamente significativo, da sintaxe transitiva. Dito de outra forma, a presença da afetação do objeto na oração é um indicador de alta probabilidade da ocorrência da estrutura oracional transitiva. Os demais parâmetros de transitividade encontram-se distribuídos de maneira mais ou menos homogênea entre

[4] O objetivo deste trabalho não é analisar a hipótese da transitividade de Hopper e Thompson (1980). Os parâmetros foram operacionalizados independente da sua obrigatoriedade, de acordo com a proposta de Hopper e Thompson (2001).

todas as estruturas oracionais, não compondo elementos distintivos da sintaxe transitiva. O autor sugere ainda que certos parâmetros, especialmente a Telicidade, podem estar associadas a padrões sintáticos não-transitivos.

A princípio isso não significa uma contraposição à teoria proposta por Hopper e Thompson (1980), uma vez que esses autores estavam interessados na correlação entre parâmetros obrigatórios, fora de uma teoria probabilística, e com um viés translinguístico. No entanto, os resultados apresentados por Lepesqueur (2017) apontam certas particularidades da organização da estrutura transitiva no português e sugerem possibilidades de se repensar a associação entre a semântica e a sintaxe transitiva.

Em uma análise de interface entre sintaxe e semântica, o caminho tradicional de investigação da transitividade tem sido agrupar padrões morfossintáticos a fim de se analisar uma estrutura semântica subjacente. Assim, por exemplo, pode-se distinguir a estrutura formal transitiva, como em [1] e [2], da intransitiva (inacusativa), como em [3], para, em seguida, tentar-se identificarem as diferenças semânticas nesses grupos. Mas o caminho inverso também é possível: primeiro identificar grupos de orações semanticamente semelhantes e posteriormente analisar a relação desse grupo com padrões formais da língua.

Por diversas razões, a primeira opção tem sido o caminho canônico de investigação. Uma das principais questões é o fato da língua agrupar uma quantidade a princípio ilimitada de informações conceituais em um número relativamente limitado de estruturas e regras gramaticais. Isso torna mais fácil agrupar as unidade oracionais a partir das suas características formais, que são em um número relativamente reduzido, do que agrupá-las a partir das suas distinções conceptuais.

Mas especialmente com o desenvolvimento de métodos computacionais para a análise de grande quantidade de dados, tem se tornado uma realidade as análises semânticas de *corpus*. Este trabalho parte, portanto, dessa via inversa e visa a identificar grupos de orações que compartilham semelhanças em termos do conjunto de parâmetros de transitividade como um todo, de maneira parcialmente[5] independente da estrutura formal da oração. Buscamos identificar grupos naturais de unidades oracionais a partir de um conjunto de técnicas estatísticas de agrupamento. Essa metodologia mostrou-se

[5] Digo parcialmente porque os parâmetros não são puramente semânticos. Por exemplo, o parâmetro Afetação refere-se a uma distinção semântica que ocorre em um certa posição sintática, a saber, a posição de objeto. Mas esse objeto pode ser, a princípio, preposicionado ou não, ou fazer parte de uma estrutura sintática transitiva ou bitransitiva.

capaz de permitir se analisarem empiricamente traços semânticos ou morfossintáticos em dados reais da língua em uso, chegando a resultados semelhantes àqueles esperados teoricamente.

Como os parâmetros de transitividade, independentemente da estrutura da unidade oracional, se agrupam no português do Brasil? Existe apenas uma semântica não transitiva, ou podemos esperar diferentes padrões semânticos fora da transitividade? Partindo desse conjunto de questões, esperamos que os resultados apresentados aqui possam elucidar as regularidades sintático-semânticas às quais os falantes estão expostos e sobre as quais emergem os fenômenos gramaticais.

2 METODOLOGIA

Uma das principais dificuldades para a compreensão da transitividade é que estamos lidando em um campo de interface entre a estrutura formal e a estrutura conceptual[6]. Apesar dos avanços recentes da linguística sobre a natureza dessa articulação, restam ainda muitas questões a respeito da maneira através da qual um item lexical se integra em uma sintaxe – e, mais ainda, em uma estrutura macrotextual e discursiva - e como isso pode produzir efeitos de significado.

Por consequência, a compreensão da transitividade depende, antes de tudo, de um tratamento de dois eixos distintos entre si: um eixo essencialmente semântico e outro essencialmente sintático. Por fim, além da descrição desses dois eixos, é preciso um modelo linguístico, talvez mais especificamente semiótico, que explique a maneira complexa e particular através da qual a sintaxe e a semântica transitiva interagem.

Existem diferenças teóricas e metodológicas significativas na descrição e na análise dos eixos conceptual e formal de uma língua. No campo da análise formal, dispomos de uma longa tradição gramatical, desde a linguística estrutural até a gramática gerativa, o que facilita o trabalho. Essa tradição, no entanto, especialmente com o advento das teorias gerativas nos anos 60, tirou grande parte do estatuto teórico da semântica, ao considerá-la uma espécie de epifenômeno ou subproduto de regras

[6] Aqui eu utilizo o termo conceptual, escrito com p, para destacar o caráter processual da estrutura semântica. Em geral, os teóricos da Linguística Cognitiva têm utilizado o termo *conceptualization* (traduzido normalmente como conceptualização) para se referir ao processo de construção de significado, destacando sua natureza dinâmica e processual. A conceptualização tem sido descrita como um processo imagético (em oposição à noção tradicional de estruturas proposicionais), interativo (porque envolve processos de negociação e interação entre os interlocutores), e imaginativo (porque envolve processos de simulação e mesclagens conceituais) (BROCCIAS, 2013).

transformacionais. Foi a partir da virada funcional/cognitiva da linguística, que a descrição e a análise desses padrões formais reaparecem articuladas às especificações de ordem semântico-pragmáticos (TAYLOR, 1995).

O caminho canônico de investigação da transitividade tem sido através da sintaxe: no geral, as teorias linguísticas têm tentado analisar padrões gramaticais (morfossintáticos e lexicais) buscando inferir uma estrutura conceptual subjacente. Este é o raciocínio básico por trás dos trabalhos de Hopper e Thompson (1980), Givón (2001) ou Naess (2007). Um caminho alternativo é tomar a estrutura conceptual como um dado, perceptível pelos falantes, a fim de, posteriormente, estabelecerem-se relações simbólicas com a estrutura formal da língua. A proposta de Halliday (1985), que compreende o sistema da transitividade como uma função gramatical que organiza, com seus próprios modelos e esquemas, é um exemplo da tentativa de focalizar, inicialmente, a maneira como a informação conceptual é estruturada para, posteriormente, identificar sua manifestação formal na língua.

Tomar a estrutura conceptual como um dado *a priori*, no entanto, pode ser uma tarefa complicada. A semântica é muito mais difícil de caracterizar: primeiro, por sua complexidade; segundo, pelas dificuldades de observação de um fenômeno que é subjetivo, e, finalmente, pela deficiência do nosso conhecimento, comparativamente ao que temos da morfossintaxe[7]. Mas os embaraços relacionados à descrição semântica não podem, no entanto, impedir que se produzam análises dessa natureza, em especial pela importância dada à semântica dentro do paradigma cognitivo não gerativista.

Este trabalho pretende, portanto, investigar a relação entre os parâmetros de transitividade, inicialmente de maneira parcialmente independente da estrutura argumental onde ele ocorre, para posteriormente tentar estabelecer uma relação entre os parâmetros e a estrutura oracional do português do Brasil. Para isso, uma análise de agrupamentos pode auxiliar, a partir de uma base empírica, a identificar grupos de orações que compartilham semelhanças em termos dos seus parâmetros. A análise de agrupamento permite descobrir uma estrutura natural dos dados sem a necessidade *a priori* de uma hipótese a respeito desses grupos.

2.1 Composição do *corpus*

[7] Como afirma Mario Perini, em comunicação pessoal.

O *corpus* desta pesquisa é composto de relatos orais produzidos por 23 participantes, publicados em Lepesqueur (2017)[8]. As narrativas orais produzidas por esses participantes foram gravadas e transcritas. Para facilitar a importação e o tratamento dos dados pelo programa computacional de análise estatística, cada linha da transcrição contém o trecho correspondente a uma única unidade oracional, definida como uma predicação centralizada pela unidade verbal. As transcrições foram realizadas usando-se as convenções ortográficas, sem, no entanto, atenção especial às questões fonéticas, uma vez que não possuem relevância para a pesquisa. Apesar disso, conforme proposto por Tenuta (2006), foram respeitados os padrões de pronúncia, em especial a ausência de morfema de plural e reduções como "tá", para "está" e "cê" para "você".

Do total de 7939 unidades oracionais da transcrição, 5690 fizeram parte da análise, uma vez excluídos trechos do entrevistador, unidades oracionais abandonadas ou parcialmente incompreensíveis, expressões idiomáticas e estruturas não sentenciais. Em um processo de amostragem simples sem reposição, foram selecionadas 690[9] unidades oracionais (30 por participante), analisadas em termos dos parâmetros de transitividade e sua sintaxe oracional. Os dados analisados foram tabulados no Software Estatístico R (R Development Core Team, 2012) de forma a conter, para cada unidade oracional[10] observada, a classificação dos parâmetros de Hopper e Thompson (1980, 2001), em termos de alta (1) ou baixa (0) transitividade. Foram incluídos todos os parâmetros apresentados no Quadro 1 (ver pag. 11) , com exceção da Individuação do objeto, que se refere a um conjunto variado de traços, o que inclui aspectos da referencialidade e definição/indefinição do objeto sintático. Os autores operacionalizam esse parâmetro em uma escala própria, distinta dos demais. A Tabela 1 apresenta as 5 primeiras observações do banco de dados.

[8] A pesquisa de Lepesqueur (2017) teve o objetivo de investigar o fenômeno da transitividade em uma população clínica. Parte do *corpus*, portanto, é comporto de entrevistas produzidas por pacientes com diagnóstico de esquizofrenia paranoide. O referido trabalho não identificou algum tipo de correlação especial intra-parâmetros na população clínica, apenas a maior probabilidade de ocorrer o parâmetro Afetação na fala dos pacientes. Uma vez que trata-se de um parâmetro pouco frequente no *corpus*, não há evidências de que os agrupamentos apresentados neste trabalho não possam ser generalizados.

[9] Os dados são originais da pesquisa de Lepesqueur (2017), que definiu o tamanho da amostra respeitando o número mínimo de observações sugeridas por Hair Jr. et al. (2009) para análise de regressão logística. O autor também considerou um desenho experimental balanceado em termos do número de observações por participantes.

[10] Os critérios de delimitação da unidade oracional são descritos com detalhes em Lepesqueur (2017).

Tabela 1: Primeiras linhas do banco de dados[11]

ID	Sujeito	Unidade Oracional	Parâmetros de transitividade								
			Afet.	Ag.	Int.	Part.	Cin.	Tel.	Pont.	Mod.	Polar.
3	suj01	Aqui, eu fui pega de refém,	0	0	0	0	1	1	1	1	1
13	suj01	porque a polícia foi junto comigo, o corpo de bombeiro.	0	1	1	1	1	0	0	1	1
14	suj01	Aí resgatô a ambulância	1	1	1	1	1	1	0	1	1
16	suj01	[Dizem que] quatro pessoas morreu.	0	0	0	0	1	1	1	0	1
18	suj01	e morreu.	0	0	0	0	1	1	1	1	1

2.2 Análise de agrupamentos

A análise de agrupamentos (também conhecida como análise de conglomerado ou de *cluster*) é um conjunto de algoritmos e de técnicas analíticas multivariadas que visa a agrupar os elementos de uma amostra ou população a partir da similaridade desses elementos quando os comparamos em uma série de variáveis (MINGOTI, 2017). O objetivo desse tipo de técnica é realizar agrupamentos que maximizem as semelhanças entre observações que pertençam a um mesmo grupo, o que torna o grupo mais homogêneo, ao mesmo tempo em que minimizem as semelhanças entre grupos diferentes, o que torna os grupos heterogêneos entre si. No campo dos estudos linguísticos a análise de agrupamentos tem sido utilizada para descrever uma ampla gama de fenômenos que vão desde diferenças dialetais até polissemias (DIVJAK e FIELLER, 2014).

Uma questão central desse tipo de análise refere-se à métrica utilizada para se decidir o grau de similaridade (ou inversamente, de dissimilaridade) entre os elementos observados. No caso de variáveis qualitativas, tais como os parâmetros binários de transitividade analisados aqui, foi utilizado o coeficiente de concordância simples (s_{ij}) (SOKAL e SNEATH, 1963). Trata-se de um coeficiente simétrico, ou seja, que considera o mesmo peso para as concordâncias positivas ou negativas. O coeficiente é calculado pela soma do número total de concordâncias entre os atributos dos elementos *i* e *j*, dividido pelo número total de atributos.

$$s_{ij} = \frac{Número\ de\ atributos\ concordantes}{Número\ total\ de\ atributos}$$

[11] Na tabela, os parâmetros de transitividade foram abreviados e são apresentados na seguinte ordem: Agentividade (Ag.), Afetação (Afet.), Itencionalidade (Int.), Participante (Parti.), Cinese (Cin.), Telicidade (Tel.), Pontualidade (Pont.), Modalidade (Mod.), Polaridade (Polar.).

O valor de s_{ij} pode variar entre 0 e 1. Para o par de orações 3 e 13 da Tabela 1, por exemplo, teríamos um valor de s_{ij} = 4/9 = 0,44. O coeficiente s_{ij} foi calculado, para cada i≠j, através do coeficiente geral de Gower (1971) que permite integrar também, se necessários, variáveis quantitativas ou ordinais.

Considerando o coeficiente de similaridade s_{ij}, a matriz de dissimilaridade dos dados será composta pelo índice de dissimilaridade d_{ij}, calculado pelo complementar de s_{ij} para cada par de orações do *corpus*.

$$d_{ij} = 1 - s_{ij}$$

A matriz de dissimilaridade produzida a partir do coeficiente de concordância simples tem as seguintes características, importantes na técnica de agrupamento utilizadas aqui: 1) é simétrica, pois $d_{ij} = d_{ji}$; 2) é positiva, pois $d_{ij} \geq 0$, se i≠j; 3) é reflexiva, pois $d_{ij} = 0$, se i = j.

A partir dessa matriz de dissimilaridade diferentes algoritmos de agrupamento são utilizados para encontrar a melhor partição dos dados. As técnicas de agrupamento são classificadas entre hierárquicas e não hierárquicas. As técnicas não hierárquicas são métodos que buscam identificar a melhor partição a partir de um número pré-determinado de grupos, levando em conta tanto a coesão interna quanto a separabilidade dos grupos formados. Essas técnicas permitem a formação de novos grupos através da junção e combinação de grupos formados em passos anteriores. As técnicas hierárquicas, por sua vez, são tipicamente utilizadas em análises exploratórias, pois não dependem de um número pré-estabelecido de grupos. A técnica hierárquica pode ser aglomerativa ou divisiva (MINGOTI, 2017).

A técnica aglomerativa começa com *n* grupos, sendo *n* o número de elementos no banco de dados. Cada observação é separada em um *cluster* específico e o algoritmo de agrupamento tenta encontrar os valores mais semelhantes para formar os grupos. Inversamente, a técnica divisiva assume inicialmente todos os elementos em um único grupo e inicia a divisão dos elementos mais distantes em grupos diferentes. A similaridade entre dois conglomerados foi definida pelo método de ligação completa, ou seja, a partir da comparação da maior distância entre os ponto de dois grupos. Esse método tende a formar grupos mais compactos e sem a tendência de longas cadeias[12] .

[12] Que ocorre quando um *cluster* incorpora, a cada interação, um único elemento próximo.

Para decidir sobre o número *k* de grupos da partição final dos dados analisados, utilizamos algumas medidas de avaliação da qualidade dos agrupamentos, analisando tanto a *compacidade* (a máxima similaridade intra-grupo) quanto a *separabilidade* (a mínima similaridade entre grupos).

Inicialmente utilizamos duas medidas de avaliação de todas as partições de 2 a 30 *clusters*[13], tanto na técnica aglomerativa quanto na divisiva. A primeira medida foi uma generalização da soma de quadrados dos desvios intra-*cluster* (tipicamente utilizada na métrica euclidiana) e a segunda medida a largura de silhueta (ROUSSEEUW, 1987).

A soma de quadrados dos desvios intra-cluster (SQ_k) é uma estimativa da compacidade de um dado cluster *k* e se refere, aqui, à metade da soma dos quadrados das dissimilaridades intra-*cluster* dividido pelo tamanho do cluster. SQ_k é definido como:

$$SQ_k = \frac{1}{2\,n_k} \sum_{i,j \in C_k} {d_{ij}}^2$$

onde n_k é o número de elementos no *cluster* C_k e d_{ij} é o valor da dissimilaridade entre o elemento i e j do cluster C_k. Quanto maior o valor de SQ_k, menor será a compacidade do cluster *k*. A soma de quadrados dos desvios intra-cluster é uma medida particular do *cluster* k. Na partição final com K *clusters*, cada um desses *clusters* apresenta um valor próprio de SQ_k. A soma de quadrados dos desvios intra-cluster da partição é dado, portanto, pela média dos valores de todos os SQ_k da partição final.

A largura média de silhueta (L) oferece uma estimativa da separabilidade dos agrupamentos ao comparar a similaridade de uma observação amostral com as demais observações do próprio *cluster* e do *cluster* vizinho mais próximo. A largura média de silhueta é calculado a partir do coeficiente de silhueta (S_i) da observação amostral *i*, dado por:

$$S_i = \frac{b_i - a_i}{MAX(a_i, b_i)}$$

onde a_i é a média da dissimilaridade (d_{ij}) da observação amostral *i* com todos os membros do *cluster* ao qual pertence e b_i, a dissimilaridade (d_{ij}) mínima da observação

[13] A principio, não esperamos que haja mais de 30 grupos teoricamente importantes para explicar o fenômeno da transitividade. Mas não se trata de uma restrição da técnica. Ainda que computacionalmente demorado, é possível analisar até *n-1* agrupamentos, sendo *n* é o número de observações no banco de dados.

i com todos os demais dados que não pertencem ao seu *cluster*. O coeficiente de silhueta S_i varia no intervalo de [-1,1] e se aproxima de -1 quando o elemento *i* está, em média, mais próximo dos elementos de um *cluster* vizinho do que dos elementos do seu próprio *cluster* (caso em que $b_i<a_i$). O coeficiente aproxima-se de 0 na medida em que b_i seja semelhante a a_i, sugerindo que o elemento *i* encontra-se em um ponto intermediário entre dois *clusters*. O coeficiente aproxima-se de 1 quando o elemento *i* está em média mais próximo dos elementos do próprio *cluster* do que do *cluster* vizinho (caso em que $b_i>a_i$).

A largura de silhueta (S_k) de um cluster *k* é dada pela média dos coeficientes de silhueta de todas as observações pertencentes ao cluster *k*.

$$S_k = \frac{1}{n_k} \sum_{i \in C_k}^{K} S_i$$

Com a finalidade de compor um índice da partição final, foi calculado a largura média de silhueta das partições que continham de 2 a 30 *clusters*. A largura média de silhueta (L) da partição foi calculado pela média dos valores de S_k dessa partição. Dessa maneira, maiores valores da largura média de silhueta sugerem uma boa separabilidade entre os grupos que compõem a partição final.

$$L = \frac{1}{K} \sum_{k=1}^{K} S_k$$

2.3 Escolha da melhor técnica de agrupamento

A partir dos índices apresentados, definimos a melhor partição obtida na técnica aglomerativa e a melhor partição obtida na técnica divisiva. Para auxiliar na comparação entre essas duas partições finais, iniciamos um segundo passo de análise das medidas da qualidade da partição a partir do Índice Dunn2 (Halkidi et al, 2001) e Índice WB. Ambos os índices são calculados a partir da dissimilaridades intra-*cluster* $d(C_k)$ e a dissimilaridade entre *clusters* $d(C_k, C_l)$.

Quanto menor as dissimilaridades intra-*cluster*, maior o compacidade da partição. A dissimilaridade intra-*cluster* do *cluster k* é dado por:

$$d(C_k) = \frac{2}{n_k\,(n_k - 1)} \sum_{i \in C_k, j \in C_k} d_{ij}$$

Quanto maior as dissimilaridades entre *clusters*, maior a separabilidade da partição final. A dissimilaridade entre o *cluster k* e *l* é dado por*:*

$$d(C_k, C_l) = \frac{1}{n_k n_l} \sum_{i \in C_k, j \in C_l} d_{ij}$$

O Índice WB (*whitin*/*between*), I_{wb}, é calculado pela razão entre as médias de $d(C_k)$ e $d(C_k, C_l)$ para todos os *clusters* da partição final. Quanto menor o índice WB, melhor a relação entre compacidade (numerador) e separabilidade (denominador). I_{wb} é dado por :

$$I_{wb} = \frac{\frac{1}{K}\sum_1^K d(c_k)}{\frac{2}{K(K-1)}\sum_{c_k, c_l, k<l} d(C_k, C_l))}$$

onde K é o numero total de agrupamento formados, $d(C_k)$ é a dissimilaridades intra-*cluster* do cluster k e $d(C_k, C_l)$ as dissimilaridade entre os *clusters k e l.*

O Índice Dunn2 é dados pela razão entre a menor dissimilaridade entre dois *clusters* e a maior dissimilaridade intra-cluster da partição final. Quanto maior o índice, melhor a relação entre a separabilidade (numerador) e a compacidade (denominador). O índice Dunn2 é dado por:

$$dunn2 = \frac{\min\limits_{k \neq l} d(C_k, C_l)}{\max\limits_{k} d(C_k)}$$

A partição final, depois de comparadas as técnicas aglomerativa e divisiva, foi representada graficamente utilizando a técnica de escalonamento multidimensional (MDS). O método MDS faz a decomposição espectral de uma matriz relacionada à matriz de dissimilaridade entre os elementos amostrais. Assim, ao se construir novas dimensões e grafar seus valores num gráfico de dispersão, conserva-se aproximadamente as dissimilaridades que os elementos amostrais apresentam entre si. Em suma, essa técnica permite representar espacialmente a matriz de dissimilaridade dos elementos sintetizando essa matriz em um certo numero de componentes utilizadas como coordenadas de um gráfico de percepção. Neste gráfico, as relações geométricas correspondem, de maneira aproximada, às relações de dissimilaridade dos dados observados (MINGOTI, 2017).

3 RESULTADOS E DISCUSSÃO

3.1 Análise descritiva dos dados

As unidades oracionais do *corpus* foram analisadas em termos dos 9 parâmetros propostos por Hopper e Thompson (1980, 2001), sendo cada um destes parâmetros caracterizado como de alta (1) ou baixa (0) transitividade. A Tabela 2, a seguir, mostra a distribuição, no *corpus*, da frequência absoluta e relativa dos parâmetros, segundo o seu grau de transitividade.

Tabela 2 - Distribuição de frequência das 690 orações segundo a transitividade (alta ou baixo) dos nove parâmetros analisados

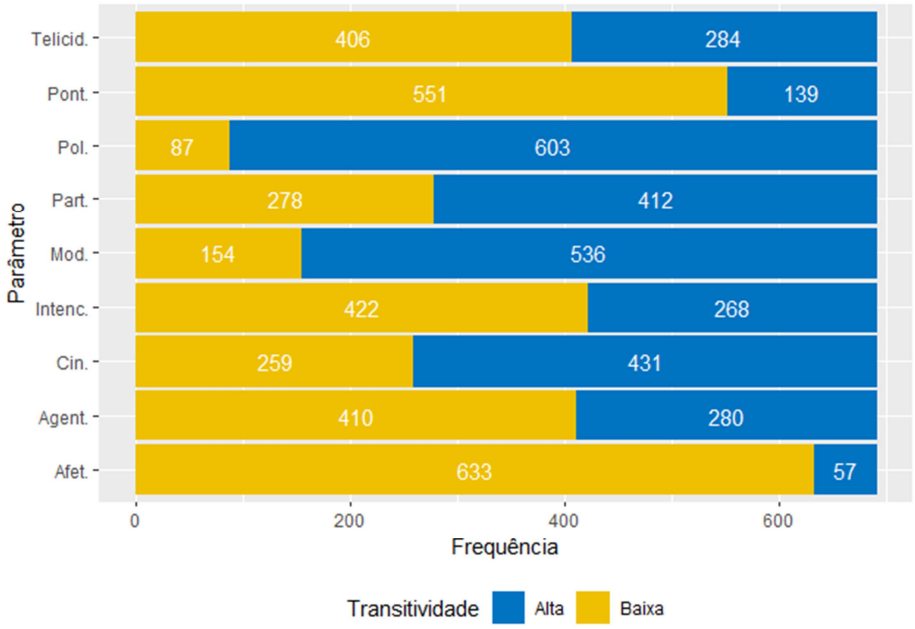
Parâmetro	Transitividade	Frequência absoluta	Frequência relativa
Agentividade	Alta	280	0,406
	Baixa	410	0,594
Afetação	Alta	57	0,083
	Baixa	633	0,917
Intencionalidade	Alta	268	0,388
	Baixa	422	0,612
Participantes	Alta	412	0,597
	Baixa	278	0,403
Cinese	Alta	431	0,625
	Baixa	259	0,375
Telicidade	Alta	284	0,412
	Baixa	406	0,588
Pontualidade	Alta	139	0,201
	Baixa	551	0,799
Modalidade	Alta	536	0,777
	Baixa	154	0,223
Polaridade	Alta	603	0,874
	Baixa	87	0,126

Alguns traços de transitividade são especialmente raros na amostra analisada: a afetação do objeto sintático (Afetação=Alta) ocorre em menos de 10% do *corpus* e a baixa polaridade da oração (Polaridade=Baixa) em apenas 12,6%[14].

A distribuição de frequência dos parâmetros pode ser bem visualizada no Gráfico 1. Uma vez que cada parâmetro soma 690 observações (tamanho da amostra), o gráfico representa também, visualmente, a proporção relativa dos traços de baixa e alta transitividade.

[14] A maior proporção de orações afirmativas é, provavelmente, uma consequência do gênero textual do *corpus*, composto por entrevistas orais.

Gráfico 1 - Gráfico de barras da frequência absoluta dos parâmetros



De maneira geral, a presença de traços de baixa transitividade são mais comuns na amostra do que traços de alta transitividade[15]. Esta característica já era esperada uma vez que a bibliografia especializada tem afirmado que o gênero conversação tende a ser de baixa transitividade, como sugerem Hopper e Thompson (2001), para o inglês, Rozas (2004), para o espanhol, Shahrokhi e Lotfi (2012), para o persa, e Lima (2013), para o português. Bois (2003), analisando a preferência no discurso pelo uso de certas configurações sintáticas, mostrou que, em diversas línguas (a saber, Hebrew, Sakapultek, Papago, Inglês e Goonyandi), 50 a 62% das unidades oracionais não possuem nenhum argumento nominal. De maneira geral, as orações de baixa transitividade parecem ser mais úteis no contexto de comunicação interpessoal e de aspectos subjetivos do que as orações de alta transitividade (ROZAS, 2004).

3.2 Análise de agrupamento por técnica hierárquica

Optamos pelo uso das técnicas hierárquicas, em uma análise exploratória dos dados, uma vez que não temos um número pré-estabelecido de grupos e buscamos identificar uma estrutura natural dos dados. As análises foram conduzidas utilizando-se

[15] Especialmente se retiramos o parâmetro Polaridade, que deixou de ser considerado relevante na descrição do fenômeno da transitividade em Hopper e Thompson (2001).

tanto técnicas hierárquicas aglomerativas[16] quanto a divisivas[17] (Kaufman e Rousseeuw, 1990).

Para a investigação do melhor agrupamento dos dados, iniciamos com a análise do nível de fusão dos aglomerados. À medida que o número de *clusters* da partições aumenta, a média da dissimilaridade intra-*cluster* decresce. Os gráficos 1 e 2, a seguir, mostram os valores da média de SQ_k de todos os *clusters* que formam cada partição, tanto na técnica hierárquica aglomerativa quanto na divisiva.

Gráfico 2 - Média da soma de quadrados da dissimilaridade intra-*cluster*
Técnica Aglomerativa

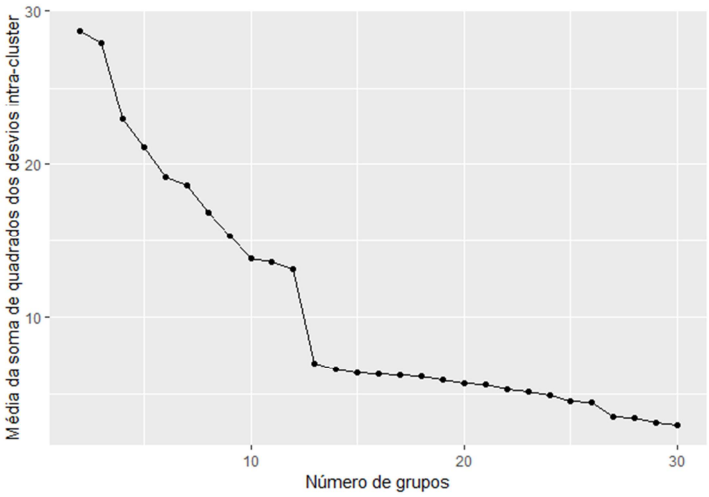
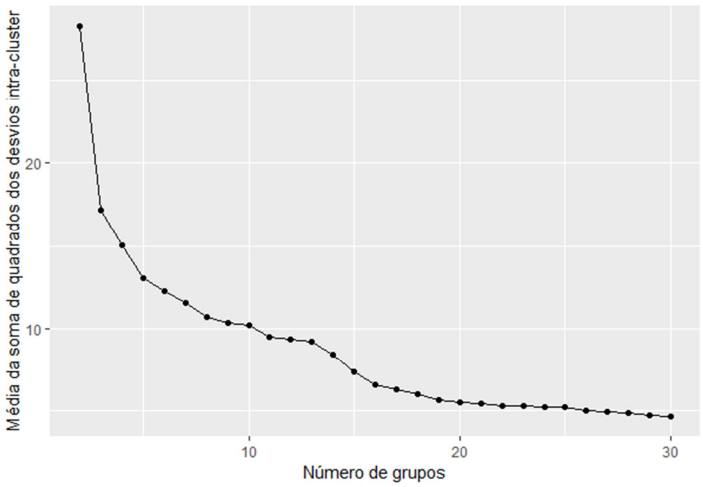


Gráfico 3 - Média da soma de quadrados da dissimilaridade intra-*cluster*
Técnica Divisiva



[16] Função hclust, implementada no pacote stats do software r
[17] Função diana, implementada no pacote cluster do software r

Buscamos identificar nos gráficos mudanças significativas (“quedas bruscas”) na média da soma de quadrados da dissimilaridade, o que representa ganhos importantes na homogeneidade ou compacidade dos agrupamentos. No uso da técnica aglomerativa, no Gráfico 2, destaca-se o ganho de consistência interna na partição k=13. No uso da técnica divisiva, no Gráfico 3, a queda significativa na soma de quadrado dos desvios ocorre na partição k=3.

Os gráficos 4 e 5, a seguir, mostram a média das larguras de silhueta de todos os *clusters* que compõem as partições com 2 a 30 grupos. Buscamos aqui as partições com maiores médias da largura de silhueta, o que representa maior separabilidade entre grupos vizinhos. Com o uso da técnica aglomerativa, o salto na média da largura de silhueta ocorre novamente no agrupamento de k=13, com ganhos poucos significativos depois dessa partição. Na técnica divisiva, o pico ocorre na partição com k=3.

Gráfico 4 - Média da Largura de silhueta: Técnica Aglomerativa

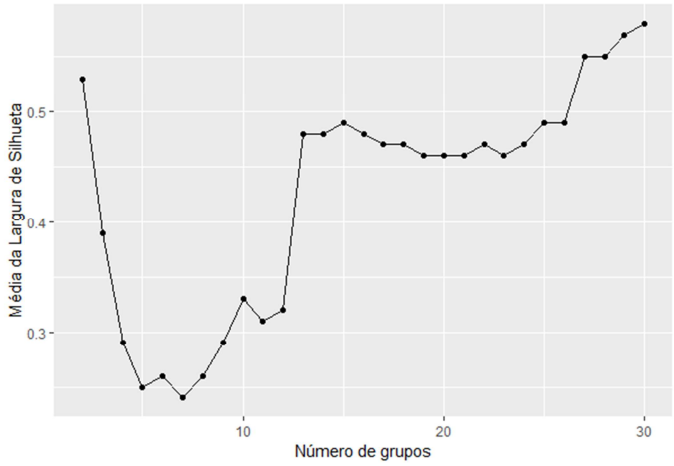
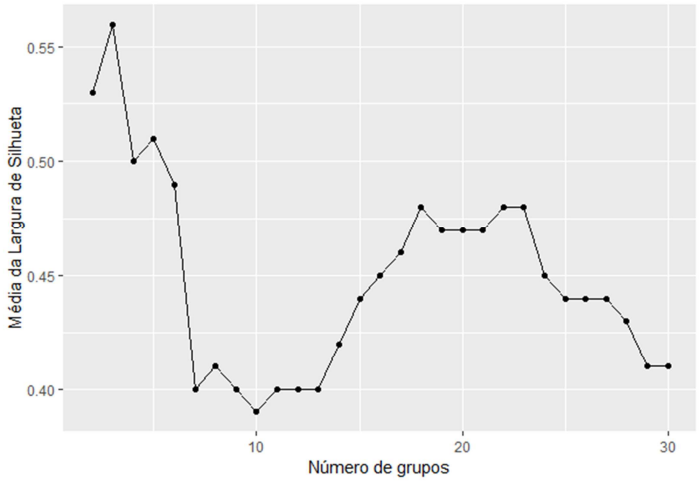


Gráfico 5 - Média da Largura de silhueta: Técnica Divisiva



As duas medidas de avaliação da qualidade dos agrupamentos sugerem, portanto, uma partição com k=3, no uso da técnica divisiva, ou com k=13, no uso da técnica aglomerativa. A Tabela 3 apresenta a comparação das duas partições através de outros índices de avaliação da qualidade dos agrupamentos.

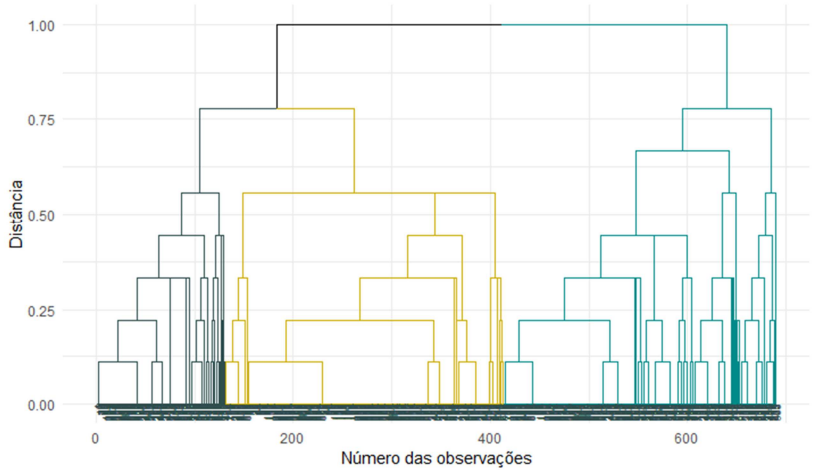
Tabela 3 - Medidas de avaliação da qualidade dos agrupamentos
(k=3, técnica divisiva; k=13, técnica aglomerativa)

Número de *clusters*	k=3	k=13
Média de SQ_k	17,16	**7,01**
Média de $d(C_k)$	0,18	**0,09**
Média de $d(C_k, C_l)$	**0,5**	0,43
I_{wb}	0,37	**0,22**
Índice Dunn2	**1,75**	0,93
Média da Largura de Silhueta	**0,56**	0,48

A Tabela 3 mostra um melhor desempenho da partição k=3 (técnica divisiva) nos índices Dunn2 e média da largura de silhueta (sendo ambos os índices uma estimativa da relação entre compacidade e separabilidade), além do melhor desempenho na média da dissimilaridade entre grupos (Média de $d(C_k, C_l)$). Apesar da partição final com treze grupos apresentar menor dissimilaridade intra-grupo (Média de $d(C_k)$), e consequentemente, melhor desempenho na razão I_{wb}, esse ganho não acompanha a perda em parcimônia no uso de um número tão grande de grupos.

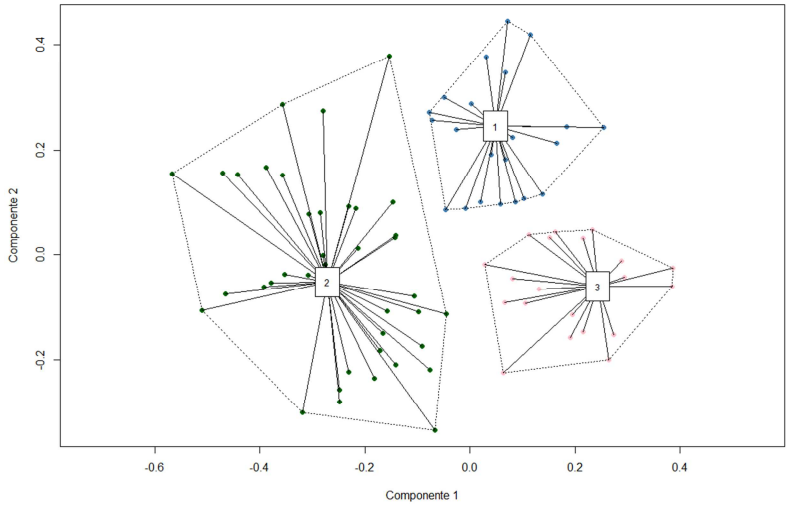
Optou-se, portanto, pela partição final com três grupos, utilizando-se a técnica hierárquica divisiva. O Gráfico 6 mostra o dendograma do agrupamento final, destacando em cores diferentes cada um dos três grupos.

Gráfico 6 - Dendograma: técnica divisiva



Utilizando a técnica de Escalonamento Multidimensional (MDS), é possível representar espacialmente, em um gráfico de percepção, a matriz de dissimilaridade dos elementos e a partição final dos dados. O Gráfico 7 representa as observações e o agrupamento final de maneira que a distância entre os pontos corresponde aproximadamente à dissimilaridade entre as observações.

Gráfico 7 - Gráfico de Percepção – Técnica divisiva k=3



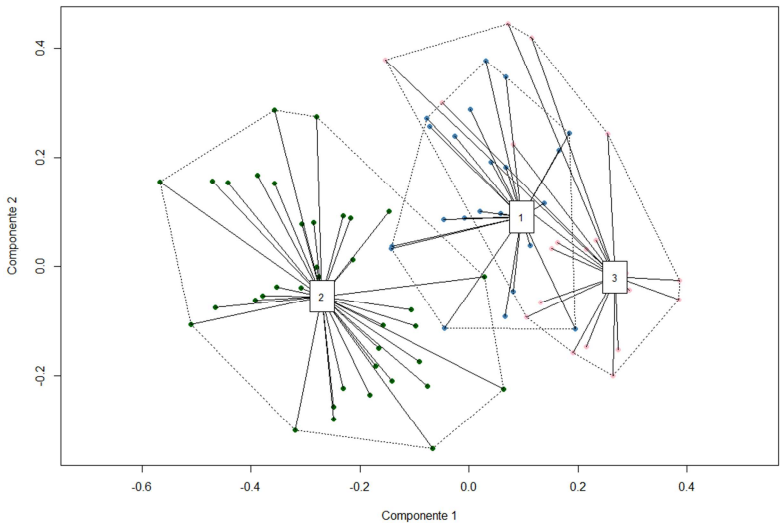
A proporção da variância explicada pelas duas dimensões obtidas através do escalonamento é de 0,47. É importante notar que não pretendemos aqui realizar a análise de agrupamento a partir do MDS, mas apenas representar graficamente a distribuição dos *clusters*. Apesar da relativa baixa qualidade de ajuste obtido via MDS,

o gráfico permite visualizar bem a correspondência entre a partição final obtida no uso da técnica divisiva e a representação espacial das observações.

O mesmo procedimento pode ser utilizado para mostrar o agrupamento obtido pela técnica aglomerativa. Se k=3 representa realmente uma partição natural dos dados, esperamos obter resultados parecidos independentemente da técnica utilizada.

O Gráfico 8, a seguir, apresenta a mesma distribuição espacial das observações mas, dessa vez, a classificação dos cluster é feita a partir da técnica aglomerativa.

Gráfico 8 - Gráfico de Percepção – Técnica aglomerativa k=3



Comparando o Gráfico 7 e 8, é possível observar que o *cluster* 2 é estável, sendo identificado pelas duas técnicas de maneira muito semelhante. No entanto, há uma diferenças de classificação entre os *clusters* 1 e 3.

A Tabela 4 é uma matriz de confusão que sintetiza a diferença entre os agrupamentos feitos pelas técnicas aglomerativa e divisiva.

Tabela 4 - Matriz de confusão entre técnicas aglomerativa e divisiva

		Classificação: Tec. Divisiva		
		1	2	3
Classificação: Tec. aglomerativa	1	119	4	90
	2	0	272	4
	3	11	1	189

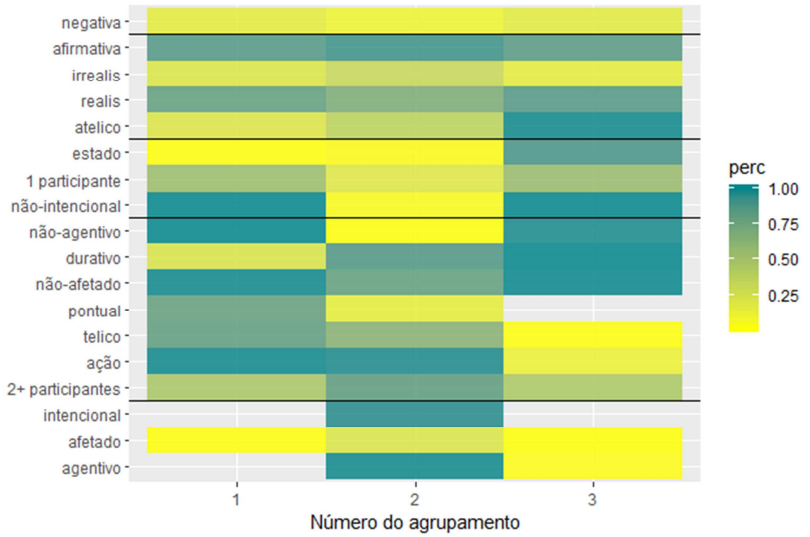
A Tabela 4 mostra que das 690 observações, 580 (84%) foram agrupadas da mesma maneira pelas duas técnicas hierárquicas. Tomando como referência a técnica divisiva, das 130 observações pertencentes ao *cluster* 1, apenas 11 observações (8,5%) foram agrupadas de maneira diferente pela técnica aglomerativa. De maneira semelhante, das 277 observações pertencentes ao *cluster* 2, apenas 5 observações (1,8%) obtiveram outro tipo de classificação na técnica aglomerativa. No que se refere ao *cluster* 3, das 283 observações 94 (33,2%) obtiveram um agrupamento divergente pela técnica aglomerativa. A única divergência significativa entre as partições vem, portando de um grupo de 90 observações que saíram do *cluster* 3 e foram incorporadas ao *cluster* 1 pela técnica aglomerativa. Isso provocou o deslocamento do centroide[18] do *cluster* 1 (representado no gráfico 7 e 8 pelo quadrado que contém o número do *cluster)*.

Estas análises sugerem que, apesar dos três *clusters* serem estáveis quando comparamos as duas técnicas, há um grupo de observações que possivelmente partilham características tanto do *cluster* 1 quanto do *cluster* 3[19].

Para melhor identificar as características típicas de cada *cluster*, apresentamos também a frequência relativa das variáveis em cada um deles. O gráfico 9 mostra a frequência relativa de ocorrências dos traços de transitividade em cada grupo da técnica divisiva k=3, destacando em verde quando o traço ocorre em aproximadamente 100% das unidades oracionais do grupo em questão e, em amarelo, quando os parâmetros correm em aproximadamente 0% das unidades oracionais pertencentes àquele grupo. A ordem de apresentação das categorias nesse gráfico foi escolhida de modo a facilitar a visualização dos conjuntos de traços mais frequentes (blocos em verde) e os menos frequentes (blocos em amarelo) em cada grupo.

[18] O centroide de um determinado cluster é definido como o valor que minimiza a soma da dissimilaridade das observações do *cluster* em questão.
[19] Essas observações merecem uma descrição a parte que vai além dos objetivos deste trabalho.

Gráfico 9 - Frequência relativa dos parâmetros em cada *cluster*



Percebe-se que os parâmetros Polaridade (negativa e afirmativa) e Modalidade (*realis* e *irrealis*) são relativamente distribuídos de forma homogênea entre os grupos. Os demais parâmetros se agrupam de forma bem definida, mostrando um padrão semântico específico de cada cluster, representado nos blocos em verde e amarelo.

O *cluster* 3 possui uma estrutura aspectual bem definida. Por estrutura aspectual, entendemos as diferenças da estrutura temporal interna, não relacionais, do evento expresso na oração (COMRIE, 1976). Este grupo apresenta predicados que expressam estados, sendo durativos (o evento apresenta certa extensão temporal) e atélicos (não apresentam um ponto de conclusão). Eles são igualmente não agentivos e não intencionais. Distintamente, os c*lusters* 1 e 2 expressam ações (são eventos não-estativos). O *cluster* 1 expressa eventos não-agentivos e não-intencionais, tipicamente pontuais, enquanto o *cluster* 2 expressa eventos tipicamente agentivos, intencionais e durativos, podendo ou não apresentar um ponto télico.

3.3 Relação entre os agrupamentos e a estrutura sintática da oração

Especialmente no âmbito da Linguística Cognitiva, um dos conceitos chaves para a compreensão da estrutura semântica é a categorização. O processo de aquisição da linguagem envolve não apenas aprender quais categorias são relevantes para nós, em nosso ambiente, mas também aprender um número limitado de estruturas e regras

gramaticais utilizadas para se expressar um número ilimitado de experiências (DIVJAK e FIELLER, 2014).

A categorização é o resultado de uma capacidade cognitiva humana geral de realizar abstrações e reconhecer um núcleo comum de aspectos da experiência corpórea e social.[20]. A experiência envolve padrões recorrentes, ou *gestalts,* no sentido de uma organização coerente, que são fundamentais para o processo de significação e estão na origem de certos pontos de referência do nosso sistema conceitual (JOHNSON, 1987, LAKOFF, 1987). Em resumo, nosso sistema conceitual ancora-se em certos padrões de interação sensório-motoras, que servem de base para a significação.

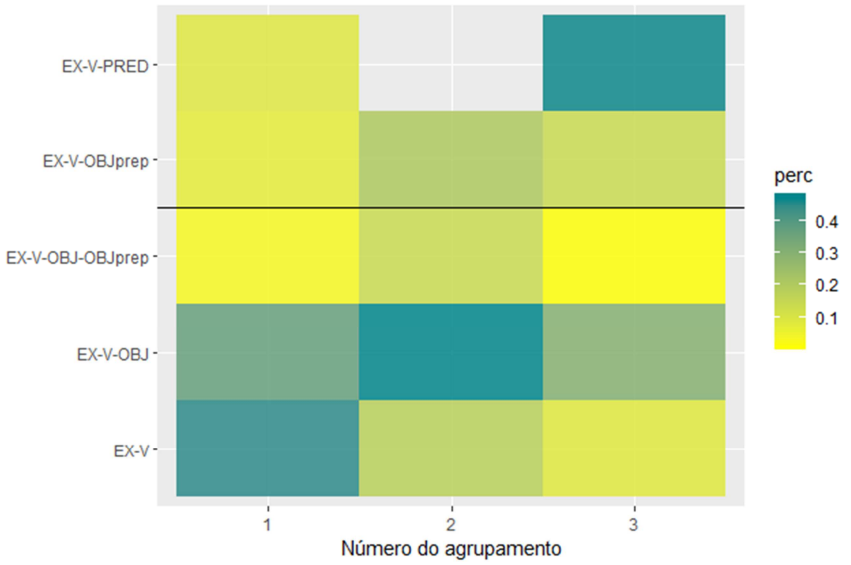
O conjunto dos dados analisados neste trabalho revela que as características semânticas da transitividade podem ser agrupadas em padrões relativamente bem definidos em termos de suas características. Esses grupos parecem instanciar três cenas prototípicas ou microcenários narrativos sobre as quais a unidade oracional se organiza.

Nós reencontramos aqui um agrupamento semelhante à distinção tradicional das classes acionais de Vendler (1967). O primeiro *cluster* aproxima-se do que Vendler denominou de *achievements*: eventos pontuais que expressam tipicamente uma mudança, mais ou menos súbita, de um estado para outro. As orações desse grupo, no *corpus,* ocorrem tipicamente associadas a sujeitos sintáticos não-agentivos e não intencionais. O segundo *cluster* agrupa o que Vendler denominou de Atividade e *Accomplishment*. Essas duas classes denotam processos que se desenvolvem no tempo, seja sem ou com um ponto télico (um ponto final ou de culminância do evento). Nos dados, eles ocorrem tipicamente associados com sujeitos sintáticos agentivos e intencionais. O *cluster* 3 denota o que Vendler chama de estado, o que equivale a uma eventualidade que se mantém inalterada em um determinado intervalo temporal.

A partir dessa partição, é possível verificar a frequência relativa da sintaxe oracional em cada *cluster*. O Gráfico 10 mostra que cada grupo pode ser caracterizado pela predominância de determinadas formas sintáticas.

[20] Contrariamente à visão clássica que compreende os conceitos como representações de estados de um mundo objetivo e, portanto, não sujeitos à experiência subjetiva, estudos empíricos têm mostrado que os conceitos são definidos e compreendidos dentro de um quadro conceitual que depende da natureza da experiência humana. Esta concepção denominada de actuação *(enaction)* ou corporeidade *(embodiment)* foi especialmente tratada por Johnson (1987) e por Varela et. al. (1991), dentro das Ciências Cognitivas, e se resume na afirmação de que a cognição não pode ser compreendida fora de nossa história social e de ações corporalizadas. Por ação corporalizada, entende-se, primeiro, que a nossa cognição é inseparável da forma como experienciamos processos sensoriais e motores (percepção e ação) decorrentes de termos um corpo como o nosso e, segundo, que essa experiência encontra-se mergulhada em um contexto biológico, psicológico e cultural mais abrangente. (VARELA et. al., 1991)

Gráfico 10 - Frequência relativa da sintaxe oracional em cada *cluster*



No Gráfico 10, a sintaxe é representada por um tipo de notação que agrupa unidades oracionais que compartilham certas características e comportamentos sintáticos. Utilizou-se a notação “EXT”(argumento externo) como uma variável que representa o que é identificado classicamente como o sujeito sintático, independentemente da posição que ocupa na oração. Isso também inclui a desinência verbal, que em Português marca as noções gramaticais de sujeito, de pessoa e de número. O argumento externo pode representar também um sintagma fora do escopo da unidade oracional que tem um papel semântico associado ao verbo e à sua construção. O símbolo “V” representa uma unidade verbal, o que inclui não apenas o verbo, mas também perífrases aspectuais e modais, assim como construções compostas por verbos leves. Por fim, o símbolo “PRED” representa um sintagma predicativo, “OBJ” objeto direto não preposicionado e “OBJprep”, objeto indireto ou preposicionado.

O primeiro *cluster* (eventos pontuais que expressam uma mudança, mais ou menos súbita, de um estado para outro, tipicamente não agentivos e não intencionais) apresenta predominantemente estruturas do tipo EX-V como em [4] e [5]. Mas pode ocorrer também formas EX-V-OBJ, como em [6], [7] e [8] especialmente envolvendo verbos de percepção (como ver e ouvir):

[4] quatro pessoas morreu.
[5] depois a intimação estourô,

[6] ela também viu ele.
[7] Já ouví passá uma sombra
[8]Eu ganhei trinta mil reais

O segundo *cluster* (processos que se desenvolvem no tempo, com ou sem um ponto télico, tipicamente agentivos e intencionais) apresenta predominantemente estruturas do tipo EX-V-OBJ como em [9] e [10], mas ocorrem também com objetos preposicionados, como em [11] e [12]. Em orações bitransitivas, como [12], o objeto preposicionado frequentemente marca o ponto télico do evento:

[9] fiquei apertando esse ossinho
[10] Aí eu preparei minhas mala toda.
[11] eu tratava das criação
[12] que ele me levô pro interior

Por fim, o terceiro *cluster* (eventualidade que se mantém inalterada em um determinado intervalo temporal, tipicamente não agentivas e não intencionais) é composto principalmente por orações com predicativos do sujeito como em [13] e [14], mas também com algumas ocorrências de estruturas do tipo EX-V-OBJ, principalmente com o verbo "ter", como em [15] e certos verbos psicológicos como em [16]:

[13] Eu tô doida
[14] E ele era evangélico,
[15] eu tenho marido,
[16] a psicóloga que sabe tudo,

4 CONCLUSÃO

Os resultados quantitativos apresentados nessa pesquisa mostram que as unidades oracionais, no português do Brasil, podem ser agrupadas em termos de parâmetros da transitividade, revelando a presença de três microcenários narrativos, semanticamente específicos, sobre os quais se desenrola o evento expresso. Apesar de não haver uma associação perfeita entre sintaxe e esses microcenários, é possível perceber a predominância relativa de certas estruturas sintáticas associadas a cada padrão semântico. Esse tipo de análise corrobora a hipótese adotada por diversos autores da Linguística Cognitiva (BRANDT, 2004; GOLDBERG, 1995, 2006; RADDEN e

DIRVEN, 2007) de que existe uma relação entre o núcleo conceitual de um determinado evento e a forma como ele é expresso em construções gramaticais.

Cada *cluster* analisado revela um tipo de significado protoconceptual, o que inclui traços aspectuais e actanciais próprios, que introduz as categorias lexicais da oração em uma cena ou cenário dinâmico. Essa noção de cenas predicativas, que vem desde Tesnière (1965), tem sido amplamente reconhecida no âmbito da Linguística Cognitiva:

> Em particular, construções envolvendo estruturas argumentais básicas parecem estar associadas a cenas dinâmicas: gestalts experienciais ancoradas, tais como alguém volitivamente transferindo alguma coisa para alguém, alguém causando alguma coisa se mover ou mudar de estado, alguém experienciando alguma coisa, alguém se movendo e assim por diante.[21] (Goldberg, 1995, p.5, tradução nossa.).

Não existe um consenso na literatura, mesmo com o extenso debate produzido sobre o assunto, em relação a quais seriam essas cenas associadas à sintaxe oracional e como elas podem ser descritas em termos de valores semânticos. O desafio teórico é a demonstração de regras gerais das operações sintáticas, uma vez que os efeitos de significação que elas produzem são enormemente variados. Mas se tomarmos o caminho inverso, ao analisar a semântica de maneira relativamente independente da sintaxe, fica evidente que esses cenários micro-narrativos existem enquanto um grupo de certos traços associados. A questão central é que esses cenários aparecem correlacionados a certos padrões sintáticos, mas não são exclusivos destes últimos. Diferentes padrões sintáticos podem acomodar um mesmo padrão semântico geral, impondo a este ultimo, possivelmente, certas particularidade.

A metodologia estatística adotada aqui mostrou-se uma ferramenta útil para se captarem esses padrões semânticos, chegando a resultados semelhantes às categorias aspectuais teoricamente conhecidas e mostrando, além disso, como essas categorias aspectuais se relacionam com categorias actanciais de agentividade e intencionalidade. Esse tipo de metodologia pode ser útil, sob uma nova perspectiva, na investigação, de maneira relativamente independente, dos padrões semânticos a que os falantes são expostos e sugere uma arquitetura específica sobre a qual a língua se organiza.

[21] No original: "In particular, constructions involving basic argument structure are shown to be associated with dynamic scenes: experientially grounded gestalts, such as that of someone volitionally transferring something to someone else, someone causing something to move or change state, someone experiencing something, something moving, and so on."

5 REFERÊNCIAS BIBLIOGRÁFICAS

BOIS, J. W. D. Argument structure: Grammar in use. In: BOIS, J. W. D. **Preferred argument structure:** Grammar as architecture for function. Amsterdam: John Benjamins, 2003. p. 11–60.

BRANDT, P. A. **Dynamic schematism and the cognitive semantics of language**, 2004. Disponivel em: <http://www.case.edu/artsci/dmll/larcs/documents/Dynamicschematismandthecognitivesemanticsoflanguage.pdf>. Acesso em: 03 nov. 2016.

BROCCIAS, C. Cognitive Grammar. In: TROUSDALE, G.; HOFFMANN, T. **The Oxford Handbook of Construction Grammar**. Oxford: Oxford University Press, 2013. p. 191-210.

COMRIE, B. **Aspect:** an Introdurtion to the Study of Verbal Aspect and Related Problems. Cambidge: Cambridge University Press, 1976.

CUNHA, C. F. D.; CINTRA, L. F. L. **Breve gramática do português contemporâneo**. Rio de Janeiro: Nova Fronteira, 1985.

DIVJAK, D.; FIELLER, N. Cluster Analysis.Finding Structure in Linguistic Data. In: GLYNN, D.; ROBINSON, J. *(Ed)* **Corpus Methods for Semantics:** Quantitative Studies in Polysemy and Synonymy, 405–441. Amsterdam: John Benjamins, 2014.

GIVÓN, T. **Syntax:** an introduction. Amsterdam: John Benjamins, v. 1, 2001.

GOLDBERG, A. **Constructions:** A construction grammar approach to argument structure. Chicago: University of Chicago Press, 1995.

GOWER, J. C. A general coefficient of similarity and some of its properties. **Biometrics**, 27: 857-874, 1971.

HAIR, J. F. et al. **Multivariate Data Analysis**. 7ª. ed. Upper Saddle River: Prentice Hall, 2009.

HALKIDI, M., BATISTAKIS, Y., VAZIRGIANNIS, M. On Clustering Validation Techniques, **Journal of Intelligent Information Systems**, 17, 107-145, 2001

HALLIDAY, M. A. K. **An introduction to functional grammar**. 3ed, London:Arnold, 1985.

HOPPER, P.; THOMPSON, S. A. Transitivity in Grammar and Discourse. **Language** , v. 56, n. 2, p. 251-299, 1980.

HOPPER, P.; THOMPSON, S. A. Transitivity, Clause Structure, and Argument Structure: Evidence from Conversation. In: BYBEE, J. L.; HOPPER, P. J. **Frequency and the Emergence of Linguistic Structure.** Amsterdam: John Benjamins, 2001. p. 27-60.

JOHNSON, M. **The body in the mind:** the bodily basis of meaning, imagination, and reason. Chicago: University of Chicago Press, 1987.

Kaufman, L., Rousseeuw, P.J. **Finding groups in data**. John Wiley & Sons, New York, 1990.

LAKOFF, G. **Women, fire, and dangerous things*:*** What categories reveal about the mind. B. Chicago: University of Chicago Press, 1987.

LEPESQUEUR, M. **Transitividade na esquizofrenia: comparação dos relatosorais de eventos psicóticos entre grupos clínico e não clínico.** Tese de Doutorado em Linguística Teórica e Descritiva: Universidade Federal de Minas Gerais, 2017.

LIMA, L. C. D. O. **A transitividade na conversação:** uma abordagem cognitivo-funcional. Universidade Federal do Rio Grande do Norte: Dissertação de Mestrado em Programa de Pós Graduação Em Estudos da Linguagem, 2013.

LUCENA, N. L. D.; CUNHA, M. A. F. D. Relações de Herança em Orações Transitivas: O mecanismo de extensão metafórica. **Letras & Letras**, v. 27, n. 1, p. 85-96, 2011.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2017.

NAESS, A. **Prototypical Transitivity**. Amsterdam: John Benjamin, 2007.

R DEVELOPMENT CORE TEAM. **The R Project for Statistical Computing**, 2012. Disponivel em: <http://www.r-project.org/>.

RADDEN, G.; DIRVEN, R. **Cognitive English Grammar**. Amsterdam: John Benjamins, 2007.

ROUSSEEUW, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **J. Comput. Appl. Math**., **20**, 53—65, 1987.

ROZAS, V. V. Transitividad prototípica y uso. **Boletín de Lingüistica**, n. 21, p. 92 - 115, 2004.

SHAHROKHI, M.; LOTFI, A. R. Manifestation of transitivity parameters in persian conversations: a comparative study. **Procedia - Social and Behavioral Sciences**, 2012, v. 46, p. 635 – 642.

SOKAL, R.; SNEATH, P. H. **The Principles of Numerical Taxonomy**. *W. H.* Freeman, San Francisco and London, 1963.

TAYLOR, J. R. **Linguistic Categorization:** Prototypes in Linguistic Theory. 2ª. ed. Oxford: Clarendon Press, 1995.

TENUTA, A. M. **Estrutura narrativa e espaços mentais**. Belo Horizonte: Faculdade de Letras da UFMG, 2006.

TESNIÈRE, L. **Eléments de Syntaxe Structurale**. Paris: Klincksieck, 1959.

VARELA, F.; THOMPSON, E.; ROCH, E. **A mente corpórea**. Lisboa: Instituto Piaget, 1991.

VENDLER, Z. **Linguistics in Philosophy**. Ithaca: Cornell, 1967