ANÁLISE DE TEXTO AUTOMATIZADA E ANÁLISE DE CONTEÚDO: ABORDAGENS COMBINADAS E APONTAMENTOS SOBRE A PRODUÇÃO LATINO-AMERICANA[1]

Lucy Oliveira[2]

RESUMO:

Este trabalho tem como objetivo apresentar uma discussão sobre a Análise Quantitativa Automatizada de Textos como lastro para a Análise de Conteúdo (AC) nas pesquisas em Ciência Política na América Latina. Como forma de minimizar o viés na produção de dados por meio da AC, diferentes ferramentas e usos computacionais têm surgido baseados no avanço de constructos epistemológicos da linguística, estatística, matemática e computação, com vistas a permitir a obtenção da unidade de análise ou mesmo de categorias com menor influência do pesquisador. Este é o caso da análise automatizada de Textos. Assim, neste trabalho, apresentaremos os fundamentos das duas análises, demonstrando seus pontos de encontro, bem como suas diferenças, além de suas contribuições para pesquisas em Ciência Política na região.

Palavras-chave*: Análise de texto, Análise de Conteúdo, Ciência Política, América Latina*

[1] Trabalho preparado para apresentação no Eixo Temático "Métodos de Investigación en Estudios Políticos y Sociales" do X Congreso Latinoamericano de Ciencia Política, da Associação Latino-americana de Ciencias Políticas (ALACIP), em coordenação com a Asociación Mexicana de Ciencias Políticas (AMECIP), organizado em colaboração com o Instituto Tecnológico de Estudios Superiores de Monterrey (ITESM), nos dias 31 de julho, 1, 2 e 3 de agosto de 2019.

[2] Pós-doutoranda pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) no Centro Brasileiro de Análise e Planejamento (CEBRAP). Doutora em Ciência Política pela Universidade Federal de São Carlos (UFSCar/Brasil) e pesquisadora nas áreas de comunicação política, agenda, eleições e interação Executivo-Legislativo. E-mail: lucyjorn.al@gmail.com

INTRODUÇ/ÃO

> Language is the medium for politics and political conflict. Candidates debate and state policy positions during a campaign. Once elected, representatives write and debate legislation. After laws are passed, bureaucrats solicit comments before they issue regulations. [...] Individual candidates and political parties articulate their views through party platforms and manifestos. [...] These examples, and many others throughout political science, show that to understand what politics is about we need to know what political actors are saying and writing. (GRIMMER & STEWART, 2013, p. 1)

Um dos principais métodos utilizados nos trabalhos que se debruçam sobre documentos e diferentes registros da política, quer seja para identificação de agendas, sentidos, posições ou para reconstruir o passado recente ou distante, é a Análise de Conteúdo. Esta serve para identificar, por meio de uma hermenêutica controlada, o que está sendo dito por um texto ou determinada comunicação. Assim, *content analysis is an observational research method that is used to systematically evaluate the symbolic content of all forms of recorded communications* (KOLBE E BURNETT, 1991, p. 243)

Apesar de sua importância, o método sofre fortes críticas no que se refere à replicabilidade das pesquisas e generalização dos resultados, bem como pelo viés subjetivo.

> Content analysts may disagree on the readings of a text. Coding instructions may not be clear. The definitions of categories may be ambiguous or do not seem applicable to what they are supposed to describe. Coders may get tired, become inattentive to important details, or are diversely prejudiced. Unreliable data can lead to wrong research results. (KRIPPENDORFF, 2004, p.1)

Além disso, o volume de informação e comunicação produzidos pelo campo político sempre andou passos à frente da capacidade de processamento e análise dos pesquisadores, permitindo a análise de *small data* e recorte muito específicos diante do tempo e recursos dispendidos para a realização da AC.

Entretanto, na última década, o desenvolvimento de ferramentas computacionais para coleta e processamento de dados, bem como a

aproximação mais intensa das Ciências Sociais com as áreas de computação, estatística e linguística, tem permitido a construção de modelos e o desenvolvimento da análise quantitativa de texto como uma das técnicas de AC, ampliando sua capacidade heurística por meio do tratamento de um volume considerável de dados, bem como da diminuição da subjetividade e arbitrariedade de parâmetros.

De fato, cabe destacar que a Análise de Conteúdo, desde o seu surgimento, sempre possuiu uma vertente estatística e matemática para a análise das comunicações humanas. Berelson (1971) por exemplo, estabelece já na década de 1950 que a análise de conteúdo é uma técnica de investigação que tem por finalidade a descrição objetiva, sistemática e quantitativa do conteúdo manifesto na comunicação. Entretanto, durante muito tempo, o tratamento estatístico era posterior ao recorte e codificação de dados gerados. Assim, o problema da objetividade e replicabilidade persistiriam.

Neste sentido, nesse trabalho caracterizaremos, a partir da retomada dos fundamentos da Análise de Conteúdo, as contribuições da Análise Textual, bem como suas conexões, de forma a apresentar, num segundo movimento, as contribuições que o uso das técnicas combinadas pode trazer aos trabalhos em Ciência Política.

De forma a ilustrar ainda a lacuna que o uso dessas metodologias combinadas na região latino-americana, iremos apresentar um breve panorama bibliométrico da produção da região amparado, em especial, nos trabalhos apresentados nos congresso da Associação Latinoamericana de Ciência Política (ALACIP) tomando como pressuposto que os encontros bianuais promovem a circulação ampla e diversa de pesquisas e trabalhos em andamento entre os países membros, se tornando um espaço importante de intercâmbio e vitrine privilegiada da produção da Ciência Política nos países das américas e Caribe[3].

[3] Neste sentido, a proposta inicial deste trabalho era investigar os periódicos de maior impacto na Ciência Política da região. Entretanto, duas questões impediram que mantivéssemos o objetivo original: há diferentes métricas em cada país para definir quais os periódicos de maior impacto e há uma circulação de autores não apenas das américas nestas revistas. Assim, para um primeiro movimento, tomamos como recorte os encontros da ALACIP que, mesmo tendo circulação aberta a pesquisadores de outras regiões do globo, tem foco na América Latina e reúne uma produção qualificada e profícua em suas atividades. Vale destacar que os trabalhos aprovados passam, assim como nas revistas, pela aprovação de pares, o que permite que possamos nos focar em algum grau de seleção do material a ser analisado, manter o recorte regional e ainda ter uma amostra ampla e diversa da produção da Ciência Política latino-americana.

Nossos dados tiveram assim como fonte primária os programas de 6 dos 9 congressos realizados pela associação[4].

Assim, este trabalho está dividido em duas partes, além dessa introdução. Na seção a seguir apresentaremos as bases da Análise de Conteúdo e sua relação com a análise quantitativa textual e lexicométrica. Além disso, apresentaremos os principais pressupostos desta segunda. Na segunda parte, demonstraremos como está a produção latino-americana a partir da análise dos programas, bem como apontaremos possíveis objetos e aplicações das técnicas combinadas. Por fim, concluímos que a análise quantitativa textual automatizada tem se difundido na última década no *mainstream* da Ciência Política internacional e tem se tornado um campo metodológico relevante a ser considerado também nas pesquisas sobre a América Latina.

1. Análise de Conteúdo e Análise textual: semelhanças e diferenças

As origens da Análise de Conteúdo (AC) remontam tempo como a análise de hinos e textos religiosos nos séculos XVII e XIX (BARDIN, 2009; KRIPPENDORFF, 2004). Entretanto, é apenas com Lasswell, na primeira metade do século XX, que aplicou técnicas para o estudo da imprensa e da propaganda na 1ª Guerra Mundial que a Análise de Conteúdo se desenvolve e torna-se um instrumento importante e diversificado de análise das comunicações orais, textuais ou imagéticas.

Neste sentido, Bardin (2009) destaca que a análise de conteúdo pode ser aplicada, desde à análise de comunicações individuais até coletivas, bem como a mensagens de diferentes naturezas e suportes. Ou seja, mesmo expressões não linguísticas, como gestos, comportamentos, sintomas patológicos ou espaços podem ser passíveis de análise de conteúdo. Isto porque, estes códigos possuem significações, tornando-se passível de inferência e técnicas para identificá-la.

[4] Isso se deve ao fato de que nos detivemos no material disponível online no portal da associação e passível de comparação. Nos dois primeiros congressos, os programas nem as memórias estão disponibilizadas online e no caso do 7º evento, apenas um relatório numérico está acessível. Assim, constam em nosso corpus de análises os programas gerais do 3º, 4º, 5º, 6º, 8º e 9º congressos, realizados respectivamente em Campinas (Brasil), San Jose (Costa Rica), Buenos Aires (Argentina), Quito (Equador), Lima (Peru) e Montevidéo (Uruguai).

FIGURA 1: Quadro resumo dos possíveis domínios da AC

Domínios possíveis da aplicação da análise de conteúdo

Código e suporte	*Número de pessoas implicadas na comunicação*			
	Uma pessoa «monólogo»	*Comunicação dual, «diálogo»*	*Grupo restrito*	*Comunicação de massa*
LINGUISTICO				
Escrito	Agendas, maus pensamentos, congeminações, diários íntimos.	Cartas, respostas a questionários, a testes projectivos, trabalhos escolares.	Ordens de serviço numa empresa, todas as comunicações escritas, trocadas dentro de um grupo.	Jornais, livros, anúncios publicitários, cartazes, literatura, textos jurídicos, panfletos.
Oral	Delírio do doente mental, sonhos.	Entrevistas e conversações de qualquer espécie.	Discussões, entrevistas, conversações de grupo de qualquer natureza.	Exposições, discursos, rádio, televisão, cinema, publicidade, discos.
ICÓNICO (sinais, grafismos, imagens, fotografias, filmes, etc.).	Garatujas mais ou menos automáticas, grafitos, sonhos.	Respostas aos testes projectivos, comunicação entre duas pessoas através da imagem.	Toda a comunicação icónica num pequeno grupo (p. ex.: símbolos icónicos numa sociedade secreta, numa casta...).	Sinais de trânsito, cinema, publicidade, pintura, cartazes, televisão.
OUTROS CÓDIGOS SEMIÓTICOS (i. é, tudo o que não sendo linguístico, pode ser portador de significações; ex.: música, código olfactivo, objectos diversos, comportamentos, espaço, tempo, sinais patológicos, etc.).	Manifestações histéricas da doença mental, posturas, gestos, tiques, dança, colecções de objectos.	Comunicação não verbal com destino a outrem (posturas, gestos, distância espacial, sinais olfactivos, manifestações emocionais, objectos quotidianos, vestuário, alojamento...), comportamentos diversos, tais como os ritos e as regras de cortesia.		Meio físico e simbólico: sinalização urbana, monumentos, arte...; mitos, estereótipos, instituições, elementos de cultura.

Fonte: Bardin, 2009, p. 37

Assim, o que define a Análise de Conteúdo? Para Bardin (2009), ela é um conjunto de técnicas voltadas a identificar as significações humanas. Krippendorff () também foca o seu conceito de AC nos fundamentos metodológicos do processo de análise, acrescentando ainda a importância da replicabilidade, já que não há leitor independente de um texto, nem mesmo um conjunto de comunicações alheia aos seus contextos. Ao considerar isso, é preciso então, para uma análise de conteúdo confiável, o controle dos mecanismos metodológicos.

Para tanto, Bardin (2009) sugere uma sequência de etapas a serem cumpridas para a realização de uma hermenêutica controlada. São elas : (i) pré-análise, (ii) exploração do material, (iii) tratamento dos resultados, inferência/interpretação.

1.1 Pré-análise

Consiste na fase de organização da pesquisa. Nela são escolhidos os materiais a serem analisados a partir do objetivo e pergunta que norteia a investigação. É nesta fase ainda que ocorre – caso não já se tenha – a formulação de hipóteses, objetivos e elaboração dos indicadores. Estes últimos

podem ser definidos pela elaboração a priori de categorias. Ou seja, a partir da teoria e dos objetivos, são elaborados um conjunto de etiquetas a serem aplicadas a determinados trechos dos textos de forma a identificar o que é procurado. A codificação a posteriori também pode ser feita, mas segue outro caminho: uma amostra dos textos é escolhida para uma "leitura flutuante" (BARDIN, 2009) a partir da qual vai se ter atenção em regularidades e características dos textos a partir das quais serão criadas as categorias ou etiquetas.

Importante destacar que o conjunto de documentos selecionados devem atender 4 condições para poderem ser passíveis de seguir para a análise: exaustividade, representatividade, homogeneidade e pertinência. No primeiro caso, trata-se do fato de que esses documentos devem cobrir todos os elementos do objeto de análises. Tomemos o exemplo da análise de propaganda eleitorais televisivas do candidato x. Para que atenda a regra da exaustividade é necessário que o corpus seja formado por todas as peças veiculadas televisivamente pelo ator em análise. Caso haja algum problema de coleta intransponível isso deve ser posto em claro para garantir o rigor da pesquisa.

O segundo elemento – representatividade – está intimamente relacionado com o primeiro. Isto porque, em alguns casos, pode-se tomar uma amostra do total de comunicações produzidas. Neste caso, é necessário que essa seja representativa do todo a ser investigado de forma a permitir a generalização dos resultados. Quando a pesquisa se dá pelo "n" total de casos, como no exemplo citado acima, a representatividade já está plenamente garantida.

Quanto à homogeneidade, a regra se destina a garantir que todos os documentos que irão compor o corpus[5] de análises sejam homogêneos quanto à sua estrutura e forma de elaboração para, assim, serem passíveis de comparação. Um exemplo neste sentido são entrevistas abertas. Mesmo que as perguntas variem conforme o andamento da entrevistas, todas as entrevistas analisadas devem versar sobre o mesmo tema. No caso de pesquisas que utilizam mais de um meio de comunicação – por exemplo, mensagens orais e escritas – é necessário colocar os documentos na mesma estrutura (i.e. transcrever os áudios) antes de começar a análise. Além disso, neste caso, seria

[5] O corpus é o conjunto de material comunicativo a ser analisado. Ele forma como um "banco de dados" de textos sobre os quais a Análise de Conteúdo será empregada.

recomendável analisar conjuntamente as mensagens originalmente escritas e depois as mensagens transcritas. No caso em que o corpus é composto de apenas um documento – i.e. uma entrevista em profundidade, um livro, um documento histórico – esta regra deve ser desprezada.

Por fim, quando se trata da pertinência refere-se ao fato de que o material a ser analisado seja uma fonte de informação adequada ao objeto de pesquisa. Neste sentido, ao estudar a posição ideológica de partidos a partir de seus documentos impressos, os manifestos, os programas de governos e outros materiais partidários que expressam as propostas da sigla são pertinentes para a pesquisa.

1.2 Exploração do material e categorização

Após a delimitação e coleta do material a ser analisado e sua organização num corpus homogêneo, representativo, exaustivo e pertinente, procede-se o trabalho de codificação ou decomposição do material.

Nesta fase, todo o corpus passa a ser recortado a partir de unidades de análises definidas pelo pesquisador e sobre as quais serão aplicadas as categorias. As unidades de análises podem ser desde parágrafos até palavras. O que importa nesta escolha é novamente a pertinência com o objeto e a pergunta a ser respondida. Cada pesquisador deve definir – na sua pré-análise – qual o ponto a ser observado: se um tema, um personagem, uma representação etc. Assim, isso se expressará numa palavra ou conjunto de palavras – unidade de registro – e seu contexto – unidade de contexto. Estas duas partes compõem a unidade de análise. Por isso, se numa pesquisa de associação livre de palavras dos eleitores com a democracia, por exemplo, a palavra pode ser a unidade de análise (contento registro/ e contexto). Já numa pesquisa sobre temas de políticas públicas em documentos partidários, pode ser necessário considerar uma frase ou parágrafo como recorte mínimo para encontrar as unidades de registro e de contexto.

Isto definido, os textos são divididos nas unidades de análises e estas serão então classificadas nas categorias de pesquisa. A *categorização* é um dos principais fundamentos da AC. É ela que permite a reorganização dos textos em classificações estáveis e exclusivas que servirão de base para a interpretação dos dados. Esta consiste, grosso modo, em classificar as unidades de análise em grupos a partir de critérios definidos pelo pesquisador. Ou seja, ela consiste

numa taxonomia a ser aplicada sobre os textos e podem ter diferentes naturezas. Por exemplo, elas podem indicar classes sociais, gênero, valores, representações, posições ideológicas, quantidades etc. As categorias são, assim, as "variações" de um mesmo aspecto que queremos investigar.

Tomemos o exemplo de uma pesquisa sobre posições ideológicas de partidos políticos a partir de suas propagandas televisivas. As categorias seriam esquerda, direita e centro. Cada uma delas seria aplicada a um conjunto restrito de características expressas textualmente em nossas unidades de análise e que seriam previamente definidas pelo pesquisador[6].

Um aspecto importante é que esta etapa, além de ser uma das mais longas, também exige um trabalho de construção e refinamento das categorias de forma a garantir que estas também sejam pertinentes e homogêneas para os objetivos da pesquisa. Com isso, precisam atender aos critérios de exclusividade, objetividade e fidelidade.

Por exclusividade, entende-se que uma unidade de análise atenda a apenas uma categoria. Se ela atende mais de uma incorre-se num problema de ambiguidade e estas precisam ser revistas. Por objetividade e fidelidade entende-se que um conjunto de categorias deve ser replicável para diferentes corpus textuais e também por diferentes codificadores, de forma a permitir que os resultados sejam objetivos. Ou seja, aqui é necessário que as categorias sejam claras e exatas a ponto de não abrirem precedentes para a subjetividade dos codificadores. E é aqui ainda que reside uma das principais críticas à Análise de Conteúdo.

Assim, diferentes métodos de controle e testes de replicabilidade estatísticos foram desenvolvidos e são aplicados para garantir a pertinência e objetividade da análise. Krippendorff (2004) os classifica em três tipos: estabilidade, a reprodutividade/replicabilidade (reproducibility) e a precisão (accuracy). Trataremos dos dois primeiros já que o terceiro é de rara aplicação[7].

---

[6] Um exemplo interessante neste sentido é o Manifest Research Project (MARPOR), um consórcio de pesquisa que a partir da análise dos temas que aparecem nos manifestos partidários em mais de 50 países nos 5 continentes classificam as siglas num espectro ideológico. O trabalho básico consiste em análise de conteúdo em que os codificadores identificam os temas de cada manifesto, quantificam aqueles que são mais proeminentes e, dependendo das ênfases, posicionam os partidos. Para saber mais acessar https://manifesto-project.wzb.eu/

[7] Para fazer o teste de precisão é necessário comparar a quantidade de acordo intracodificadores com um índice encontrado sobre os mesmo dados em situação ideal. Entretanto, aponta Lima

No primeiro, o que se pretende medir é a estabilidade da classificação de um mesmo codificador. Ou seja, o grau de invariabilidade de um processo de codificação ao longo do tempo. Para tanto são feitos teste-reteste, em que um codificador duplica, num momento posterior, o procedimento de codificação que aplicou a um mesmo conjunto de dados. Não existindo desvios relevantes entre as codificações realizadas em diferentes momentos, conclui-se que os resultados são estáveis. Esta é a forma mais fraca de medir a confiabilidade da codificação e deve ser empregada sempre com outros indicadores da aceitabilidade de uma análise de conteúdo.

O segundo tipo - reprodutividade - mede a confiabilidade intracodificadores, ou seja, quão estável são os resultados daquelas categorias aplicadas por diferentes codificadores. A situação ideal é aplicar as mesmas categorias a um mesmo corpus com diferentes codificadores em separado e medir o grau de concordância. Em alguns casos, a literatura aponta ser suficiente a medição percentual desta concordância por meio da seguinte fórmula geral. Entretanto, como já apontado por Cohen (1960), o acordo entre dois codificadores ou mais pode se dar por mero acaso das condições atribuídas. Krippendorff (2004) admite que este consenso pode abranger até 50% das unidades em análise. Ou seja, o acordo intracodificadores consistente deve então ocorrer para além da que se estima que teria acontecido por mero acaso

E como calcular? existem mais de 20 índices diferentes para se realizar um teste de confiabilidade[8]. Entretanto, três deles se destacam pelo seu uso mais disseminado, o *kappa de Cohen*, o *pi de Scott* e o *alpha de Krippendorff.* O primeiro, desenvolvido por Jacob Cohen (1960), confronta a proporção de acordo observado com o nível de acordo estatisticamente esperado em condições de aleatoriedade a partir de n categorias e x documentos. É um índice considerado conservador pois foi inicialmente pensado para testar a replicabilidade entre dois codificadores. Com múltiplos pesquisadores é sugerido usar a versão modificada - *Kappa de Fleiss*.

(2013, p.12) os padrões comparativos que permitiriam o cálculo deste tipo raramente existem, não sendo possível, na grande maioria dos casos, optar por ela. Assim, a solução mais adequada são os testes de reprodutividade ou replicabilidade.

[8] Para mais sobre as diferenças entre os índices, ver Hayes e Krippendorf (2007) e Feng (2014).

Sampaio e Lycarião (2018) apontam que, dos três citados, o *alpha de Krippendorff*, mostra-se bastante prático e versátil, pois não tem restrição em termos de número de codificadores e de natureza das variáveis, se são ordinais, categóricas ou contínuas (KRIPPENDORFF, 2004). Entretanto, para pesquisas que trabalham com um número alto de categorias pouco presentes, a escolha de outro índice pode ser mais adequada (WOZNIAK, LÜCK & WESSLER, 2015)[9]. O que importa destacar, por fim, é que estes testes de replicabilidade permitem minimizar os efeitos da subjetividade dos resultados, permitindo uma maior validade da análise de conteúdo que é feita pela classificação humana.

1.3 Tratamento dos resultados e inferência

Realizada a categorização do corpus e testes que comprovam a replicabilidade e objetividade das classificações, parte-se para a parte final da análise. Cabe ressaltar neste sentido que, ainda na fase da categorização, é necessário estabelecer quais as métricas que servirão para organizar os resultados numéricos encontrados. Ou seja, a presença/ausência de determinada informação, a frequência total ou ponderada das categorias, variação de frequências ao longo do tempo são apenas alguns exemplos de métricas. E é a partir destas que se constrói a inferência.

Assim, a inferência é resultante do trabalho conjunto de desenho da pesquisa com a elaboração das categorias e indicadores que serão retomados, no tratamento de dados, para voltar ao início do trabalho e responder, a partir dos dados reorganizados, às questões e objetivos da investigação. Como afirma Bardin (2009, p. 41)

> O analista é como um arqueólogo. Trabalha com vestígios: [...] os vestígios são a manifestação de estados, de dados e de fenómenos. Há qualquer coisa para descobrir por e graças a eles. [...] o analista tira partido do tratamento das mensagens que manipula para *inferir* (deduzir de maneira lógica) conhecimentos sobre o emissor da mensagem ou sobre seu meio, por exemplo.

Por fim, importante destacar que esta pode se basear puramente nos indicadores e tratamentos estatísticos – como cálculo de coocorrência e testes

[9] Atualmente, os softwares de análise de conteúdo - como NVivo e MaxQDA - bem como estatísticos - como SPSS e R - possuem ferramentas para o cálculo automáticos desses índices.

multivariados – ou mesmo qualitativa ou mista. Isso traz uma multifuncionalidade ao método e também permite a combinação de diferentes técnicas para estabelecer a inferência.

1.4 Análise textual automatizada: hermenêutica controlada sobre as palavras

É nessa porosidade da AC que o método se conecta com a análise quantitativa automatizada de textos. Bardin (2009) destaca que esse encontro se dá com o surgimento da lexicometria na década de 1970, linha que aponta o léxico - o conjunto de palavras de uma língua - como o lócus da análise.

Ou seja, uma diferença fundamental que já indetificamos é que a AC considera diferentes unidades de análise. Já a análise textual tem como unidade principal a palavra. Neste sentido, o principal pressuposto da análise quantitativa de texto é que os sentidos das trocas comunicativas podem ser encontrados por meio de análises matemáticas sobre as palavras, que passam a ser tratadas estatisticamente por meio da contagem de frequência e testes multivariados, sendo analisada sozinha e em contexto (CÚRCIO, 2006; LEBART E SALEM, 1994). Não precisa ir muito além para perceber que esta abordagem coaduna perfeitamente com a perspectiva da AC, tornando-se um *subset* desta.

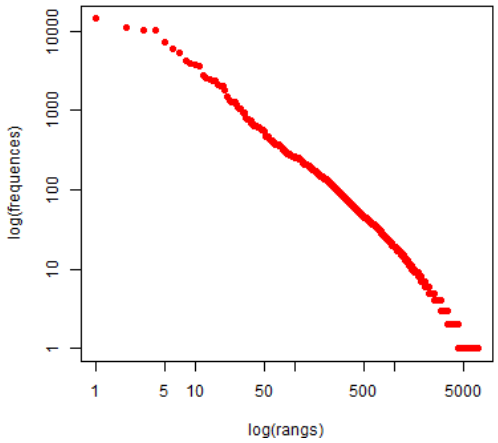
Outro traço distintivo e também fundamental é que, na análise automatizada de textos, as categorias e classificações não ocorrem por meio de categorias a priori estabelecidas pelos pesquisadores, mas são realizadas por softwares que, a partir das palavras dos corpus e parâmetros estabelecidos pelos pesquisadores, estabelece as classificações e repartição dos textos.

Para tanto, as formas materiais destas - a sequência de letras que juntas montam as variações dos léxicos - são o primeiro ponto de organização dos corpus por meio dos processos de *lematização* e *steamming*. Neste, as palavras são reduzidas a sua menor forma completa, retirando suas terminações. Assim, substantivos no plural, vão para o singular. Verbos conjugados em diferentes tempos são colocados no infinitivo. Isso ocorre para que haja um parâmetro inicial de contagem, já que as palavras têm um core: os seus radicais. Depois de reduzidas, elas são agrupadas por radicais - lematização.

É por meio da lematização que é possível caracterizar ainda quais as palavras mais frequentes num dado conjunto de textos. Interessante destacar, neste sentido, que a inferência sobre essas frequências é feita com o suporte de

teorias da linguística. Por exemplo, as formas mais freqüentes na listagem são as chamadas palavras gramaticais, também encontradas como palavras funcionais[10], como os artigos, os dêiticos, as preposições, os pronomes, as conjunções. Entretanto, apesar de trazer indicações estilísticas, estas são menos significantes que os primeiros substantivos listados, pois os mesmos carregam o peso temático (LEBART E SALEM, 1994; BERNARD, 1994). Neste sentido, diferentes linguistas apontam que as palavras relevantes para um determinado léxico nem são as mais frequentes, nem as praticamente ausentes, mas aqueles que estão nos pontos médios de uma distribuição normal.

FIGURA 2: Exemplo de um gráfico de distribuição de Zipf



Fonte: Iramuteq. Elaboração própria

Assim, um segundo elemento na análise textual é entender a importância das classes das palavras. Substantivos, verbos e adjetivos em comparação com preposições e conjunções serão menos frequentes, mas não necessariamente menos importantes. Por isso, outro elemento importante na análise textual automatizada é a definição dos parâmetros de classificação dos textos, levando em conta as classes de palavras relevantes para cada objeto de pesquisa. Isso, deve estar previsto na formulação dos softwares e scripts a serem aplicados.

Outro aspecto relevante é que o desenvolvimento computacional permitiu identificar não apenas as palavras, mas estas e seus contextos, por meio da classificação de conjuntos lexicais estáveis tomando como parâmetros as distâncias lexicais e palavras com que co-ocorrem (LEBART E SALEM, 1994).

[10] Em francês, são conhecidas como *mots-outils, formes fonctionnelles* ou *formes vides*. Nos modelos de língua inglesa, são chamadas de “stop word removal”.

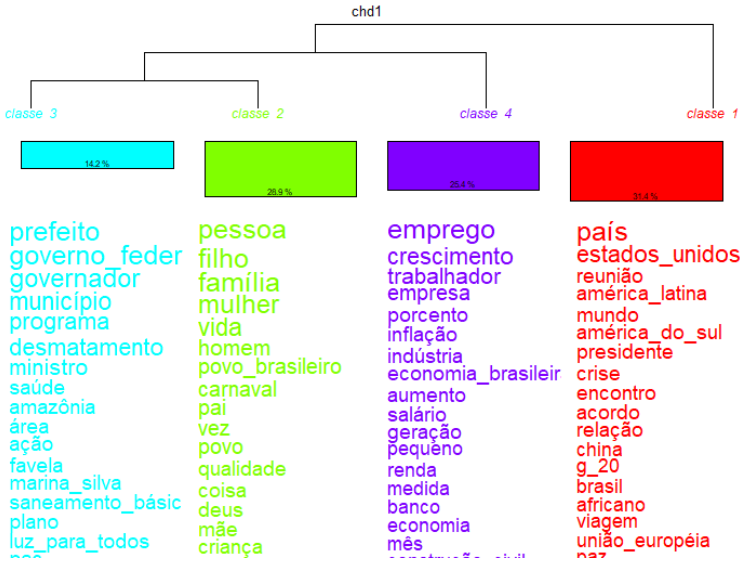
Assim, dois outros processos são desencadeados: a segmentação e a análise estatísticas multivariada.

No primeiro, é definido um espaço mínimo de sentido das palavras - chamado de unidade de contexto (LEBART E SALEM, 1994; CAMARGO E JUSTO, 2016) - que pode ser formado por 40 toques ou 5 palavras ou outro parâmetro a ser definido pelo pesquisador. Assim, depois da separação por palavras, os *corpus* textuais são decompostos por segmentos iguais que servirão para os testes estatísticos multivariado futuros. Em casos de textos curtos - como um tweet ou respostas curtas a uma questionário aberto ou entrevista - é necessário não fazer a decomposição em segmentos, sendo considerado todo o texto como a unidade de contexto das palavras.

Importante ressaltar, neste momento, que grande parte da estatística multivariada aplicada à análise textual é baseada em álgebra matricial. Ou seja, os textos são decompostos em um espaço vetorial e transformados em matrizes de frequência (document-term matrix). Cada palavra se torna em um vetor *v* com distribuição que varia entre 0 e 1 (ausência/presença) num conjunto *z* de documentos/atores (BLEI, NG & JORDAN, 2003). Por *default*, na matriz resultante as linhas são os textos ou segmentos e as colunas são as palavras. Assim, a partir da decomposição das frequências das palavras encontradas pelos segmentos e textos aos quais pertencem é possível proceder testes de correlação de variáveis qualitativas, como x2.

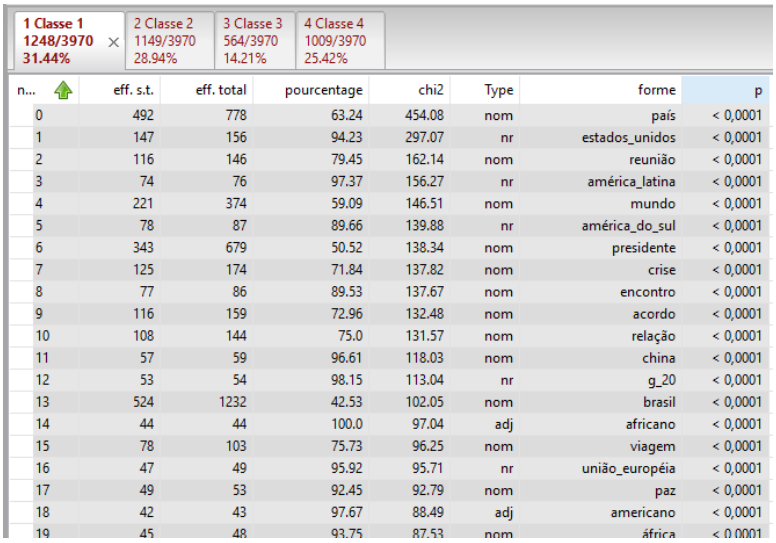
Essa é a base, por exemplo, da Classificação Hierárquica Descendente (CHD) criada por Reinert (1983). Basicamente, a análise é feita em dois movimentos, com sucessivos testes de x2. O primeiro movimento é ordenar quais os segmentos mais semelhantes pelas formas/palavras utilizadas. E depois, dentro desses segmentos semelhantes as formas/palavras mais relevantes entre si. O resultado é apresentado em um dendograma com clusters estáveis (com maior índice de inércia) composto de palavras com alta relevância estatística interna e baixa externa.

FIGURA 3: Exemplo de dendograma da CHD



FONTE: Iramuteq, elaboração própria

FIGURA 4: Perfil da classe 1 com valores de x2 e p-value por palavra



1 Classe 1 1248/3970 31.44% | 2 Classe 2 1149/3970 28.94% | 3 Classe 3 564/3970 14.21% | 4 Classe 4 1009/3970 25.42%

n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
0	492	778	63.24	454.08	nom	país	< 0,0001
1	147	156	94.23	297.07	nr	estados_unidos	< 0,0001
2	116	146	79.45	162.14	nom	reunião	< 0,0001
3	74	76	97.37	156.27	nr	américa_latina	< 0,0001
4	221	374	59.09	146.51	nom	mundo	< 0,0001
5	78	87	89.66	139.88	nr	américa_do_sul	< 0,0001
6	343	679	50.52	138.34	nom	presidente	< 0,0001
7	125	174	71.84	137.82	nom	crise	< 0,0001
8	77	86	89.53	137.67	nom	encontro	< 0,0001
9	116	159	72.96	132.48	nom	acordo	< 0,0001
10	108	144	75.0	131.57	nom	relação	< 0,0001
11	57	59	96.61	118.03	nom	china	< 0,0001
12	53	54	98.15	113.04	nr	g_20	< 0,0001
13	524	1232	42.53	102.05	nom	brasil	< 0,0001
14	44	44	100.0	97.04	adj	africano	< 0,0001
15	78	103	75.73	96.25	nom	viagem	< 0,0001
16	47	49	95.92	95.71	nr	união_européia	< 0,0001
17	49	53	92.45	92.79	nom	paz	< 0,0001
18	42	43	97.67	88.49	adj	americano	< 0,0001
19	45	48	93.75	87.53	nom	áfrica	< 0,0001

FONTE: Iramuteq, elaboração própria

Ou seja, por meio da CHD, é possível identificar, por exemplo, o conjunto de palavras que juntas mostram um vocabulário estável em torno de temas e políticas. Outra técnica baseada na decomposição de textos no plano vetorial é a LDA - Latent Dirichlet Allocation (BLEI, NG E JORDAN, 2003). Ela é um dos mais comuns algorítimos para identificação de tópicos guiada por dois princípios fundamentais:

- Todo documento é um conjunto de tópicos/temas;
- Todo conjunto de temas é composto de um conjunto de palavras;

Blei, Ng e Jordan (2003) demonstram que todo tópico é, de fato, uma variável latente multinomial, ou seja, representa a probabilidade de uma distribuição específica de um conjunto de palavras. E isso é calculado a partir de distribuições multinomiais e testes de correlação entre as palavras que compõem os conjuntos.

Como é possível perceber, nas duas análises, os princípios matemáticos são os mesmos. A diferença é que a LDA primeira permite identificar os diferentes tópicos por documento, podendo, por exemplo, demonstrar como num mesmo discurso um presidente pode tratar de diferentes temas da agenda. No caso da CHD, esses resultados são mostrados para a totalidade de documentos analisados (corpus total). É possível, com a ajuda de alguns softwares, identificar num mesmo documento os clusters montados na CHD, mas isso não é dado automaticamente e requer a ação do pesquisador.

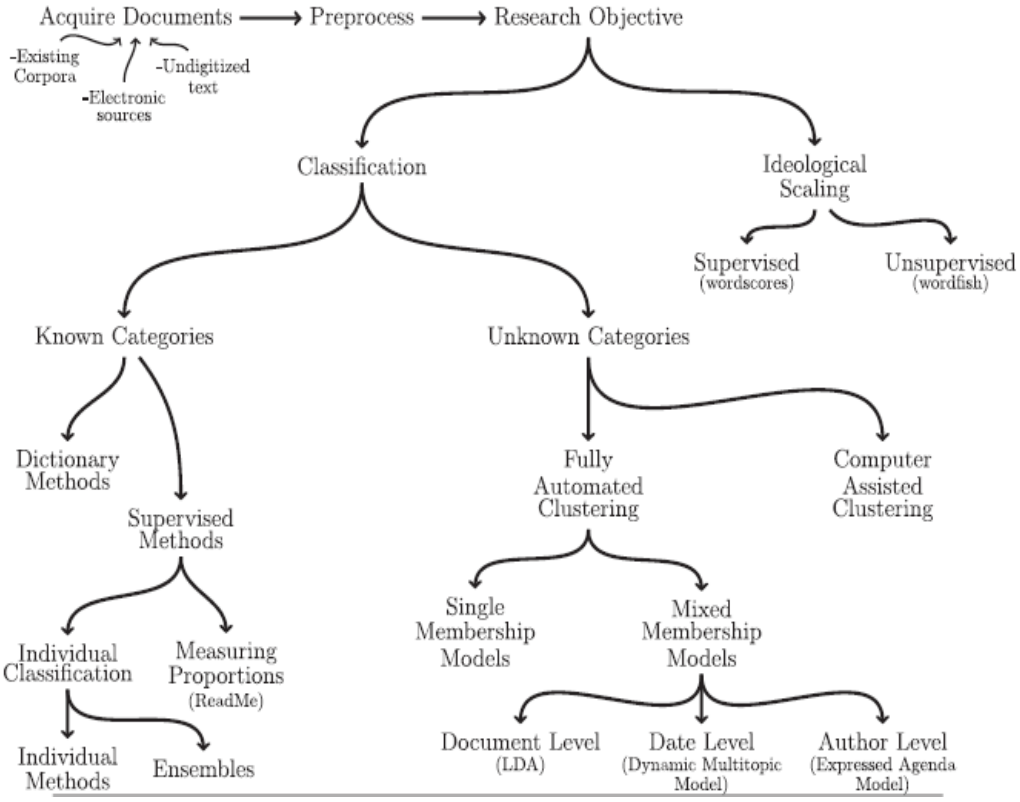
FIGURA 5: Exemplo dos resultados da LDA por documento

1	press,democratic,international,democracy,countries,variables,institutions
2	elections,election,local,members,models,population,levels
3	military,war,conflict,civil,people,leaders,empirical
4	groups,group,interest,problems,foreign,law,latin
5	behavior,cases,society,soviet,values,issue,approach
6	party,parties,model,electoral,vote,voting,effects

FONTE: Nielsen, 2013, p. 22

Por fim, Grimmer e Stewart (2013) apontam que nenhuma análise automatizada prescinde completamente da ação do pesquisador. Isto porque, mesmo que a classificação seja completamente organizada pelo computador, a inferência, ou seja, as interpretações dos dados virão da ação dos cientistas políticos sobre aqueles dados. Assim, eles sugerem uma abordagem combinada - mix methods - onde codificações humanas de uma amostra dos documentos são realizadas, e validadas com os testes de replicabilidade, e servem de parâmetro de controle também das distribuições automatizadas. Outra forma é usar *supervised methods*, ou seja, análises que são organizadas a partir de um *pre set* de palavras organizadas pelos pesquisadores de forma a alimentar os algorítimos na busca e organização das categorias. Por fim, caso seja usada a análise completamente automatizada, importante ter parâmetros de validação dos outputs, como os índices de correlação entre as palavras e critérios de seleção de palavras relevantes. Na figura abaixo, os autores resumem os principais métodos de análise quantitativa de texto e suas abordagens.

FIGURA 6: Métodos de análise quantitativa de textos



Fonte: Grimmer e Stewart, 2013, p.2

Assim, após essa explanação de suas principais características e formulações teórica é possível perceber como a análise de conteúdo e a análise de texto automatizada é uma combinação tanto possível quanto frutífera para diferentes objetos e análises da Ciência Política. As vantagens em termos de rapidez de processamento e do volume de dados a serem tratados são ainda somadas à validação das etapas realizadas a partir de testes de Kappa de Cohen e correlação para garantir a confiabilidade dos dados.

2. AC e AT na produção latinoamericana

Entretanto, apesar da robustez dos métodos e relevância que têm adquirido na Ciência Política internacional, o uso desta metodologia combinada ainda é incipiente na América Latina. Algumas iniciativas têm sido feitas para superar isso, como cursos metodológicos nos encontros da ALACIP voltado especificamente para a Análise Textual. Mas isso ainda não se reflete num campo ou objeto robusto como vemos no caso estadunidense, por exemplo.

Assim, para ilustrar essa lacuna, apresentaremos nesta parte final do trabalho, alguns dados bibliométricos da produção da Ciência Política

latinoamerticana baseada especificamente na análise dos programas dos congressos da Associação Latino-Americana de Ciência Política (ALACIP).

O corpus foi composto de 6 programas gerais dos 9 congressos realizados por serem o material disponível no sítio eletrônicos da associação. A análise foi de conteúdo combinada com textual a partir do software MaxQDA sobre os títulos dos trabalhos que compuseram as mesas e eixos temáticos dos congressos. Ao todo, foram analisados mais de 7.200 títulos de trabalhos e mesas como descritos na tabela abaixo:

Tabela 1: Total de trabalhos por congresso da ALACIP analisado

Congresso	ano	local/país	total de paginas	total de trabalhos
3º congresso	2006	Campinas/Brasil	36	296
4º congresso	2008	San Jose/Costa Rica	78	490
5º congresso	2010	Buenos Aires/Argentina	144	1230
6º congresso	2012	Equador/Quito	127	1851
8º congresso	2015	Peru/Lima	71	1379
9º congresso	2017	Uruguai/Montevideo	223	1971
TOTAL			**679**	**7217**

Fonte: elaboração própria

A escolha dos programas completos se deve ao fato de que eram documentos homogêneos, passíveis assim de serem comparados. Além disso, os eventos da ALACIP se tornaram com o passar do tempo no principal encontro da região, tornando-se vitrine para trabalhos em diferentes níveis – graduação e pós-graduação – e diferentes instituições – universidades, centros de pesquisas, projetos e governos – além de serem resultado da seleção de pares nos eixos temáticos e mesas. Assim, garante a validade destes dados como um parâmetro possível de nos permitir, grosso modo, uma visão geral da produção da Ciência Política latino-americana.

Assim, após a coleta destes programas no site da ALACIP, fizemos o upload no software e submetemos a uma classificação automatizada dos textos a partir de três descritores fundamentais: a palavra “análise” em espanhol, português e inglês, por serem as três línguas que podem ser utilizadas para submeter trabalhos para o evento. O retorno dos resultados segue nas tabelas 2 e 3.

Tabela 2: Total de aparições do termo "Análise" (nas três línguas) por ano

Congresso	Total de trabalhos	Total de aparições	%
3º congresso	296	20	6,76
4º congresso	490	22	4,49
5º congresso	1230	85	6,91
6º congresso	1851	124	6,70
8º congresso	1379	142	10,30
9º congresso	1971	206	10,45
TOTAL	**7217**	**599**	**8,30**

Fonte: elaboração própria

Tabela 3: Aparição dos descritores por língua e por ano

Congresso	Total de aparições	Espanhol	%	Português	%	Inglês	%
3º congresso	20	3	15,00	17	85,00	0	0,00
4º congresso	22	17	77,27	4	18,18	1	0,05
5º congresso	85	69	81,18	16	18,82	0	0,00
6º congresso	124	78	62,90	39	31,45	7	0,06
8º congresso	142	98	69,01	36	25,35	8	0,06
9º congresso	206	102	49,51	95	46,12	9	0,04
TOTAL	**599**	**367**	**61,27**	**207**	**34,56**	**25**	**0,04**

Fonte: elaboração própria

Aqui já é possível perceber que há um incremento com o passar dos anos dos números de trabalhos que possuem o termo "análise" em suas diferentes formas e línguas nos títulos. Entretanto, uma ressalva é importante: sabemos que o termo "análise" pode ser utilizado para descrever abordagens metodológicas – como análise institucional, análise comparada – ou também para descrever o processo de reflexão científica de forma ampla[11].

Assim, para qualificar esses resultados, submetemos cada conjunto de títulos selecionados a uma análise lexicométrica básica com expressão visual – nuvem de palavras – que irá nos ajudar a caracterizar quais os termos mais frequentes e se entre eles a questão da análise de conteúdo e de textos. Os resultados seguem demonstrados abaixo.

[11] Além disso, cabe ressaltar que os títulos podem ser insuficientes para identificar os trabalhos que fazem análise de conteúdo e de textos em suas pesquisas já que bem sempre a metodologia está expressa nos títulos. Entretanto, nem todos os programas traziam os resumos dos trabalhos, o que impedia uma escala maior de comparação. Assim, mantivemos a análise dos títulos, tomando também como pressupostos que se um aspecto é extremamente relevante irá ser destacado no título.

FIGURA 7: Nuvens de palavras do *corpus* dos descritores – ALACIP 4

Fonte: Elaboração própria com software MaxQDA

FIGURA 8: Nuvens de palavras do *corpus* dos descritores – ALACIP 5

Fonte: Elaboração própria com software MaxQDA

FIGURA 9: Nuvens de palavras do *corpus* dos descritores – ALACIP 6

Fonte: Elaboração própria com software MaxQDA

FIGURA 10: Nuvens de palavras do *corpus* dos descritores – ALACIP 8

Fonte: Elaboração própria com software MaxQDA

FIGURA 11: Nuvens de palavras do *corpus* dos descritores – ALACIP 9

Fonte: Elaboração própria com software MaxQDA

O que a análise textual específica dos corpus formado pelos títulos com os descritores de cada congresso nos mostra é que as análises comparada, de políticas públicas, institucional e estudos de casos são as mais proeminentes no conjunto de trabalhos. Os termos “Análise de Conteúdo”, “Análisis de Contenido” e “Content Analysis” aparecem em títulos apenas nos congressos 5, 8 e 9, sendo que numa frequência muito baixa – 1 ou 2 – o que nem as permite aparecer nas nuvens montadas (com frequência mínima de 3 aparições). Uma outra pista que cabe investigação é que aparece, com mais frequência, a palavra “discurso” juntamente com análise, o que aponta ora o método ora o objeto (discursos presidenciais ou legislativos).

Apesar de ser breve, o que esses dados apontam é que este tipo de campo profícuo ainda aparece pouco nos trabalhos analisados ou têm pouca ênfase em seus títulos restando uma investigação mais profunda nos trabalhos, passo que esperamos fazer num trabalho vindouro.

Considerações finais

Este trabalho teve como objetivo discutir as pontes entre análise de conteúdo e análise automatizada de textos, buscando ainda apresentar um breve panorama dos trabalhos que utilizam essas técnicas de forma combinada a partir da análise dos programas dos congressos da ALACIP disponibilizados na página da associação.

Uma conclusão importante é que as técnicas de análise textual automatizada são um *subset* do método da Análise de Conteúdo, permitindo a superação de dificuldades no que se refere à replicabilidade, objetividade e ampliação das análises de AC a partir do uso de computadores, o que aponta para uma combinação metodológica promissora.

Apesar de breve, o trabalho bibliométrico sobre os programas da ALACIP demonstram a ausência do uso destes tipos de métodos, o que torna urgente a ampliação do uso dessa metodologia, bem como o refinamento de suas técnicas a partir dos casos latino-americanos. Entendemos, por fim, que este trabalho é um primeiro movimento no sentido de refletir sobre o uso combinado das metodologias e esperamos, com isso, contribuir para novas discussões e aplicações metodológicas.

REFERÊNCIAS BIBLIOGRÁFICAS:

BARDIN, L. Análise de Conteúdo. Edições 70, Lisboa, 2009

BERNARD, M. **Introduction aux études littéraires assistées par ordinateur**. 1ª edition. Presses Universitaires de France. Paris, 1999.

BLEI, NG & JORDAN, 2003 Latent Dirichlet Allocation. In**: Journal of Machine Learning Research**. v. 3. 2003. p. 993-1022

CAMARGO, B. V; JUSTO, A.M. **Tutorial para uso do software de análise textual Iramuteq.** 2016. Disponível em http://www.iramuteq.org/documentation

CÚRCIO, V.R. Estudos estatísticos de textos literários. In: **Revista Texto Digital**. v. 2, n. 2. 2006.

GRIMMER, J. STEWART, B.M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. In: Political Analysis, 2013. p. 1–31.

KRIPPENDORF, K. Content Analysis: as introduction to its methodology. SAGE Publications, 2004.

LEBART, L; SALEM, A; BARRY, L. Exploring Textual Data. Springer, London, 1998

REINERT, A. Une methode de classification descendente hiérarchique: applicacióna l'analyse lexicale par contexte. In: **Les cahiers d'analyse des donné**, tomme 8, n 02, 1983. p. 187-198

SAMPAIO, R; LYCARIÃO, D. Eu quero acreditar! Da importância, formas de uso e limites dos testes de confiabilidade na análise de conteúdo. In: **Sociologia e Política**. v. 26, n. 66, jun. 2018. p. 31-47.